Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 176 (2007) 1066–1083
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

Performance evaluation of a multi-echelon production, transportation and distribution system: A matrix analytical approach

Fong-Fan Wang [a], Chao-Ton Su [b,*]

[a] *Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu, Taiwan*
[b] *Department of Industrial Engineering and Engineering Management, National Tsing Hua University, 101, Kuang Fu Road, Sec. 2, Hsinchu 300, Taiwan*

Received 19 December 2003; accepted 7 September 2005
Available online 18 November 2005

Abstract

This study proposes an originative method to evaluate complex supply chains. A tentative multi-echelon production, transportation and distribution system with stochastic factors built-in is employed as a test bed for the proposed method. The supply subsystem formulated in this study is a two-stage production facility with constant probability of feedback and stochastic breakdowns. The transportation subsystem is a service facility with one server. The distribution subsystem under study is a single central warehouse with M retailers. All the participants of the supply chain use base-stock policies and single-server settings. We investigated both the make-to-order (MTO) and make-to-stock (MTS) policies for different base-stock levels, as adopted at different sites. Applying quasi-birth-and-death (QBD) processes as decomposed building blocks and then using the existing matrix analytical computing approach for the performance evaluation of a tandem queue constitutes the main procedure of this study. We also discuss the possibilities of extending the current model to account for other inventory control policies as well as for multi-server case. Numerical study shows our proposed analytical model is robust for practical use.

Keywords: Performance evaluation; Queueing; Multi-echelon; QBD

* Corresponding author. Tel.: +886 3 574 2936; fax: +886 3 572 2204.
E-mail address: ctsu@mx.nthu.edu.tw (C.-T. Su).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.09.015

1. Introduction

The operation of a production/supply system has become increasingly more complex in recent years. This has resulted in an abundance of recent supply chain management (SCM) studies. In this work, we focus on problems of integrated stochastic supply chain (SC) systems, especially those of tandem-like models. An integrated SC model includes not only a production phase but also several other phases such as distribution and transportation. Usually the operational goal of a SC is to maintain a quick responding SC with minimum operating cost. However, with the sophisticated nature of today's SC, most stochastic multi-echelon SC models can only focus on an individual phase. Some models focus on the interaction between warehouse and retailer operations. Others focus on the study of the production phase. This motivated us to propose a system design and analysis framework, which provides integrated solution approach for multi-echelon supply networks. Our objective is twofold. The first is to develop a flexible modeling approach, which can 'capture' realistic activities inside each stage, such as parallel servers, machine unavailability, etc. We used QBD process to achieve this goal. The second is to extend the applicability of Lee and Zipkin (1992) to include into our model not only a tandem-processing network but also other SC subsystems. Unlike Lee and Zipkin (1992), which assumed single server and exponential distribution at each processing stage, our model relaxes these assumptions and hence allows for more modeling flexibility. Our model is also a variant of classical tandem supply network. Different from previous related works (see literature review below), our work links production/inventory subsystem and distribution/inventory subsystem. And transportation process is considered as well. All three subsystems have limited capacities (actually we use single-server settings in the main part of this study). Through QBD transformation, the original complex topology of an integrated stochastic SC becomes tandem-like and hence tractable. We also used QBD process to model machine unavailability, which makes our model more real than other integrated stochastic SC models such as Cohen and Lee (1988).

In this paper, we assumed that there is an infinite input buffer at each stage along the SC, and that each stage uses the base stock control policy. A policy of this kind demands that each stage starts operation at its own target inventory level at its output buffer. Under such scheme, the output buffer at each stage is set to be finite, while the input buffer at each stage does not have to be set so. However the infinite assumption at the input buffer of each stage releases the difficult analysis of possible blocking effect when units at the upstream stage cannot find any vacancy at the input buffer of downstream stage. Also, unit transfer is assumed and the supply discipline is assumed to be first-come-first-served (FCFS). For practical reason, there is usually a natural quantity unit for both demand and supply (e.g. truckload or 1000 tons/unit load), and in terms of that unit it makes sense to set order quantity to be equal to unity. First we formulated the respective stages as either $M/M/1$ or phase-type queuing model. For the latter type, we then used the quasi-birth-and-death (QBD) model of the Markov process to derive respective sojourn times. Finally, the method of Lee and Zipkin (1992) was applied and then the system-wide performance measures were computed approximately with respect to the base stock levels at all sites. A simulation model was also developed to facilitate the verification study of the accuracy of the proposed approach.

This paper is organized as follows. Section 2 reviews related literature. Section 3 introduces the theory. Section 4 employs a tentative SC model as a test bed for our method and presents the consequent numerical results. Section 5 is a discussion and sensitivity analysis. Section 6 discusses possible extension of the proposed model. Finally, in Section 7 we draw our conclusions.

2. Literature review

Our survey of relevant literature indicates that integrated stochastic models for a production-distribution system are still rare. Research on such models can be found at Cohen and Lee (1988), Pyke and Cohen

(1993, 1994), etc. However they neglected the mutual relationship between the different subsystems. They did not consider factors of uncertainties from upstream stages such as material unavailability, which influences the behavior of the downstream stage. Specifically, Houtum et al. (1996) pointed out that an integrated model for analyzing a multi-stage, multi-product SC problem, which is theoretically sound and numerically tractable would be recognized as a breakthrough in SCM study.

As for the studies of separate subsystems of a SC, there is already a large body of multi-echelon production/inventory and distribution/inventory models in the literature. Axsäter (1993a,b, 2000) and Svoronos and Zipkin (1988) developed different inventory control theories for distribution/inventory systems. Svoronos and Zipkin (1991) first proposed the matrix-analytic approach to solve multi-echelon distribution/inventory problems. The authors assumed unlimited capacity with stochastic lead-times. In particular, the lead-times are unaffected by demand. The major result of Svoronos and Zipkin (1991) is that the transit-time variances play an important role in system performance. Later Lee and Zipkin (1992) used similar approach to discuss the model of a tandem queue with planned inventories. The model therein assumed finite capacity. It used the model of Svoronos and Zipkin (1991) as an approximation. In order to make the approximation reasonable, it set the parameters of the lead-time to correspond with the average lead-time in a queuing system. Hence the lead-times depend on the demand. The simulation results showed the accuracy of the approximation model. Using a different recursive approach, Zipkin (1995) concluded an important concept, namely that a tandem queue with feedback issue built-in can be treated as capacity loss. Other issues of stochastic manufacturing in a tandem queue, which are often investigated by researchers, involve the buffer size allocation, machine breakdowns and blocking effect. Enginarlar et al. (2002) used simulation and regression to set up rules-of-thumb to decide adequate buffer capacity, which guaranteed high production efficiency of a tandem production with unreliable machines. Abboud (2001) used the discrete-time Markov model to study the machine breakdown issue of a one-stage production/inventory model. Mohebbi (2003), Kalpakam and Sapna (1997), Mahmut and Perry (1995) used respective Markov models to formulate supply unavailability as two 'on' and 'off' states, and to study the embedded stochastic process to derive performance measures of interest. As for the study of blocking effect embedded in a tandem-processing queue, Lee and Zipkin (1992) gave review of related literature. Recent example is the work of Gurgur (2002). They assumed two-buffer designs at each stage and the input buffer is finite. These assumptions cause difficulty of analysis associated with blocking. They used the decomposition method to separate the whole SC into several two-node subsystems to facilitate the analysis and then used the iterative approach to integrate all the subsystems to obtain the final system-wide performance measures. They assumed that transfer is in batches, and they used (r, q) inventory control policies at each stage. The current study assumes intermediate stocks at the output buffer of each stage. However the input buffer is assumed infinite. More research, which assumes that only the output buffer at the final stage is positive while the others are zeroes and most involving multiple final products can be found in the literature as reviewed by Lee and Zipkin (1992).

The flexibility of using QBD modeling approach to 'capture' complexity of inner stage processing can be found, for example, in Neuts (1994, pp. 274–286) where arriving customer order is served by multiple parallel machines. The random unavailability of machines is attended by multiple repairmen. As for using phase type distribution as stochastic inventory control models, Zipkin (1988) first proposed the idea and the main results achieved therein were that the marginal distribution of lead-time demand has a discrete phase-type distribution with the same number of phases as the lead-time distribution. Duri et al. (2000) extended Lee and Zipkin (1992) to allow for phase type distribution at each processing stage for more involved supply networks.

3. Matrix analytical approach to evaluate a complex supply chain

The inventory control scheme of our proposed approach is the base stock policy. In practical production/inventory control policies, in contrast to centralized (and push-type) control scheme like MRP, there

are other local (and pull-type) control policies like (r, q), KANBAN and their variants except for base-stock policy. Though we used base-stock policy in this paper, the extension of the existing model to other control policies is possible (this will be investigated in Section 6). Below we illustrate why we select this policy instead of other pull-type control schemes.

The base-stock policy makes sense when economies of scale in the SC are negligible relative to other factors. For example, when each individual unit is very valuable, and hence holding and backorder costs dominate any fixed order (set-up) costs. Likewise, for a slow-moving product (one with a low demand rate where Poisson distribution is adequate to model the arrival process), the economics of the system dynamics clearly rule out batch size (Zipkin, 2000). When the above conditions no longer exist, for example, economies of scale do matter; other control schemes such as (r, q) policy may be more adequate than base-stock policy. In this paper, we assumed processing conditions are like those mentioned above so as to use base-stock policy accordingly. On the other hand, since KANBAN is more restrictive when possible blocking may occur due to no immediately available KANBAN cards at hand when demand arrives, we select base-stock policy as our major control scheme to quickly verify the applicability of the proposed model in the first place. Also, base-stock policy is not uncommon in practical production/inventory control situation. Finally, it is known base-stock policy can be treated as the building block of (r, q) policy and therefore we begin our study from the base-stock policy.

Next, we discuss how the base stock control policy works. This policy is also called $(S - 1, S)$ policy. Where S represents base stocking level. This policy means that whenever demand reaches one unit, the inventory is immediately replenished. Under our proposed model, each stage along the SC has its own input queue (N_j) and output buffer (I_j) physically or imaginarily, where semi-finished or finished products are kept. Assume infinite N_j and finite I_j. Aggregate customer demands at the retailers trigger the delivery from the central warehouse (CW). This demand information propagates to the production facility initiating a production order at each stage. For a specific production, transportation or distribution stage j, a material flow comes from the output buffer of the immediate upper-stage $j - 1$. If the inventory at the buffer is available, one item is immediately deducted from the output buffer of $j - 1$ and sent to the input queue of j. If the inventory at the buffer is not available, one item is backordered and recorded at $j - 1$. When there is one part/product finished at stage j and there is recorded backorder, then the item will be sent immediately to the input queue of the next stage. Otherwise it will just stay at that stage as a base stock item. Under base-stock control, the stage adopting "make to stock" (MTS) policy will maintain its own stock level and reduce customer-waiting times downstream as compared to a pure "make to order" (MTO) policy. Next, the models developed by Svoronos and Zipkin (1991) and Lee and Zipkin (1992) are briefly discussed. In Section 3.2, our proposed approach is presented.

3.1. The approximation approach to evaluate tandem queues with planned inventories

Consider stage j and its immediate predecessor stage i. The following notations were used by Svoronos and Zipkin (1991): L_j: total lead-time of customer order at stage j; D_i: the waiting time at stage i; T_j: processing time at stage j, including waiting and service times at stage j; K_j: the outstanding customer order at stage j; B_i: the backorder level of stage j recorded at stage i; S_j: the stock level at stage j.

Then, the authors defined the following: $L_j = D_i + T_j$ and assume T_j and L_j had continuous phase-type distributions (CPH) as follows: $T_j \sim \mathrm{CPH}(\alpha_j, A_j)$ and

$$L_j \sim \mathrm{CPH}(\boldsymbol{\psi}_j, G_j^*). \tag{1}$$

Let I denote an identity matrix and $\mathbf{1}$ a column vector of ones whose dimension is chosen to fit the content of the context. Then, they indicated that K_j has the same distribution as the lead-time demand. This property combined with Neuts (1994) Theorem 2.2.8 implies that K_j has a discrete phase type distribution (DPH): $K_j \sim \mathrm{DPH}(\pi_j, P_j)$ where

$$\boldsymbol{P}_j = \lambda\left(\lambda\boldsymbol{I} - \boldsymbol{G}_j^*\right)^{-1}, \tag{2}$$

$$\boldsymbol{\pi}_j = \boldsymbol{\psi}_j\boldsymbol{P}_j. \tag{3}$$

Since $B_i = [K_i - S_i]^+$, where $[x]^+ = \max\{x, 0\}$ is a shifted phase-type distribution, it follows that $B_i \sim \mathrm{DPH}(\pi_i P_i^{S_i}, P_i)$ (Neuts, 1994, p. 47). According to Svoronos and Zipkin (1991), B_i has the same distribution as the waiting-time demand. Again this property combined with Neuts (1994) Theorem 2.2.8 implies that $D_i \sim \mathrm{CPH}(\boldsymbol{\psi}_i P_i^{S_i}, G_i^*)$. From the definition of L_j, (1) is the convolution of two phase-type distributions: D_i and T_j. According to Neuts (1981) Theorem 2.2.2, since $L_j = D_i * T_j \sim \mathrm{CPH}(\boldsymbol{\psi}_j, G_j^*)$, where * represents convolution operation, then

$$\psi_j = \left[\psi_i P_i^{S_i}, (1 - \psi_i P_i^{S_i}\mathbf{1})\alpha_j\right], \tag{4}$$

$$\boldsymbol{G}_j^* = \begin{bmatrix} G_i^* & -G_i^*\mathbf{1}\alpha_j \\ 0 & A_j \end{bmatrix}. \tag{5}$$

As Lee and Zipkin (1992) assumed each processing stage to be exponential with one single server, then (5) can be expressed as (6) (see the following) after some recursive algebraic operations starting from stage 1:

$$\boldsymbol{G}_j^* = \begin{bmatrix} -v_1 & v_1 & & & \\ & -v_2 & v_2 & 0 & \\ & & \ddots & \ddots & \\ & 0 & & -v_{j-1} & v_{j-1} \\ & & & & -v_j \end{bmatrix}, \tag{6}$$

where $v_k, k \leqslant j$ represents the inverse of the sojourn time of customer order at stage k. In our approximation approach, we relax the assumptions of exponential and single server. The inverse of sojourn time: v_k is obtained through QBD modeling. Under this approach, the processing activity at each stage can be modeled as complex as possible theoretically. This approach largely enhances the flexibility of the model.

Since there is no waiting time before the first stage, the distribution of L_1 is the same as T_1, which is already known. Starting at $\boldsymbol{\psi}_1 = [1]$, Lee and Zipkin (1992) recursively solved (4) by using (2) and (6) and let $\alpha_j = 1$. From the property of DPH, they finally derived

$$\Pr\{K_j > S_j\} = \boldsymbol{\pi}_j\boldsymbol{P}_j^{\boldsymbol{S}_j}\mathbf{1},$$

and

$$E[B_j] = \boldsymbol{\pi}_j\boldsymbol{P}_j^{\boldsymbol{S}_j}(\boldsymbol{I} - \boldsymbol{P}_j)^{-1}\mathbf{1}, \tag{7}$$

where $\boldsymbol{\pi}_j$ is obtained from (3). Alternatively we find it is simpler to derive (7) as follows

$$E[B_j] = \sum_{K_j > S_j}(K_j - S_j)\Pr\{K_j\} = \sum_{y_j \geqslant S_j}\Pr\{K_j > y_j\} = \sum_{y_j \geqslant S_j}\boldsymbol{\pi}_j\boldsymbol{P}_j^{y_j}\mathbf{1} = \boldsymbol{\pi}_j\boldsymbol{P}_j^{\boldsymbol{S}_j}(\boldsymbol{I} - \boldsymbol{P}_j)^{-1}\mathbf{1}.$$

Here we use the tail probability to derive the second equality. Since $S_j = I_j + K_j - B_j$, where I_j represents on hand inventory at stage j, Lee and Zipkin (1992) gave

$$E[I_j] = S_j - E[K_j] + E[B_j] = S_j - \boldsymbol{\pi}_j(\boldsymbol{I} - \boldsymbol{P}_j)^{-1}\mathbf{1} + E[B_j]. \tag{8}$$

Notice that the first moment of DPH was used to derive the last equality of (8). For $j < J$, this quantity together with $E[N_{j+1}]$, gives the total intermediate inventory between stages j and $j + 1$. When all the S_j equal to zero, the initial probability vector of the Markov chain is $(1, 0, \ldots, 0)$, so that the sojourn time in the queue, if it is a pure tandem one involving no feedback or breakdown issues, is the sum of indepen-

dent J random variables with mean $\frac{1}{v_j}$, where $\frac{1}{v_j}$ is the respective sojourn time at stage j. Consequently the approximation can be verified to be exact. However, in our test model, which we will discuss shortly, the queue involves feedback and breakdown. Under this situation, the respective sojourn time, except for the first stage in the tandem queue is still that of a $M/M/1$ queue. However we have to modify the sojourn time at the first stage to improve the accuracy of the approximation model as discussed in Section 5. For now, we will only focus on the build-up of our approximation model as described below.

3.2. The proposed approach

Now we discuss how to use the QBD process combined with the approach of Lee and Zipkin (1992) to derive the performance measures of more complex SCs. First we discuss how to break down the original queuing problem into many smaller queues along the chain. Then we approximate the matrix computation approach developed by Lee and Zipkin (1992) and plug in all the decomposed sub-queues sojourn time information as the matrix parameter to derive the final performance measures that are of interest to us.

Our sub-queues include two types: the $M/M/1$ queue and the phase type queue. For the $M/M/1$ queue, the derivation of sojourn time is well known by simply applying the well known Little's formula. For the phase type queue, our model demands that each job arriving at stage j may have to go through several physical processing phases before it finishes the processing work and releases the occupied resource to the next arriving job waiting in the queue. Under this stochastic process, the infinitesimal generator matrix will have a tri-diagonal block form. Markov chain with this form is a QBD process. Applying the theory of QBD, we can derive the expected sojourn time at this processing stage. As for the distribution subsystem, we can also treat it as an $M/PH/1$ queue and apply the above QBD process derivation procedure. Alternatively, we can accumulate all the retailers as a single stocking site and treat it as an $M/M/1$ queue, which will later be shown to be equal to the $M/PH/1$ queue under some specific conditions. And we then calculate each individual retailer separately and finally we obtain aggregate performance measures for retailer site. In the following, we use the steady-state probability derivation procedure as illustrated in Feldman and Valdez-Flores (1995, see Appendix 1) to derive the sojourn times in an unreliable production stage and a distribution stage, respectively.

First, we derive the sojourn time for an unreliable processing stage. Assume 0 and 1 phases represent the breakdown and operating states, respectively. And, assume all stochastic processes are Markovian with parameters $\lambda, \mu, \zeta, \gamma$, representing mean arrival, processing, and the up and down rates, respectively. We can then formulate the phase type generator as

$$\boldsymbol{G} = \left[\begin{array}{cc|c} -\gamma & \gamma & 0 \\ \zeta & -(\mu+\zeta) & \mu \\ \hline 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{G}_* & \boldsymbol{G}_\Delta \\ \hline \boldsymbol{0} & 0 \end{array}\right].$$

(Note that the bold character form represents vector or matrix.) Assume the initial probability in the phase stage as $\boldsymbol{\alpha}_* = (0, 1)$, apply (A.5), and after some matrix algebraic operations, we get

$$\boldsymbol{R} = \begin{bmatrix} \frac{\lambda}{\lambda+\gamma}\left(1+\frac{\zeta}{\mu}\right) & \frac{\lambda}{\mu} \\ \frac{\lambda}{\lambda+\gamma}\cdot\frac{\zeta}{\mu} & \frac{\lambda}{\mu} \end{bmatrix},$$

which is consistent with Buzacott and Shanthikumar (1993, p. 122). Applying the expectation formula ($L = \sum_{n=1}^{\infty} n \cdot p^n$) and (A.4), we can easily obtain the expected number of orders in the system

$$L = (1-\rho)\boldsymbol{\alpha}_* \boldsymbol{R}(\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{1}, \tag{9}$$

where the traffic intensity rate is

$$\rho = \lambda E[T] = -\lambda \boldsymbol{\alpha}_* \boldsymbol{G}_*^{-1} \mathbf{1}. \tag{10}$$

The last equality of (10) is from the CPH distribution. Then the sojourn time in the processing stage can be obtained by applying the Little's formula $W_s = \frac{L}{\lambda}$.

Since every distribution can be approximated as closely as desired by phase type distribution (Svoronos and Zipkin, 1991), it seems that we can formulate any stage in the SC as a QBD process in a very flexible way. For now, we will now apply the same approach to a distribution subsystem and show that the end result is the same as treating all the retailers as a single stocking unit under some conditions. Basically the random process of a distribution system can be modeled as a Hyper-exponential process. Recall a Hyper-exponential distribution as shown in Fig. 1. We can treat the start node as the input queue to each retailer route. α_i is the probability of which route the transportation will take.

In the long run, under normal conditions, α_i can be approximated as $\frac{\lambda_i}{\sum_i \lambda_i}$, where λ_i represents the mean order rate for retailer i, and the denominator is just the average aggregate demand rate. Node 0 can be thought of as the location of the collective single stock-place. For ease of derivation, assume that the expected delivery rates for all routes are identical, that is $\mu_1 = \mu_2 = \cdots = \mu_m = \mu$. Also assume that there are m retailers, and that all customer demands are identical, that is $\lambda_1 = \lambda_2 = \cdots \lambda_m = \lambda$ and thus $\alpha_1 = \alpha_2 = \cdots \alpha_m = \frac{1}{m}$. Then we have the following phase-type representation:

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m) = \alpha_*, \tag{11}$$

$$\boldsymbol{G} = \left[\begin{array}{ccc|c} -\mu & & & \mu \\ & \ddots & & \vdots \\ & & -\mu & \mu \\ \hline \multicolumn{3}{c|}{\mathbf{0}} & 0 \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{G}_* & \boldsymbol{G}_\Delta \\ \hline \mathbf{0} & 0 \end{array}\right]. \tag{12}$$

Applying (A.5), we get

$$\boldsymbol{R} = \begin{bmatrix} \frac{(\lambda + m\mu)\lambda}{m\mu(\lambda+\mu)} & & \frac{\lambda^2}{m\mu(\lambda+\mu)} \\ & \ddots & \\ \frac{\lambda^2}{m\mu(\lambda+\mu)} & & \frac{(\lambda + m\mu)\lambda}{m\mu(\lambda+\mu)} \end{bmatrix}.$$

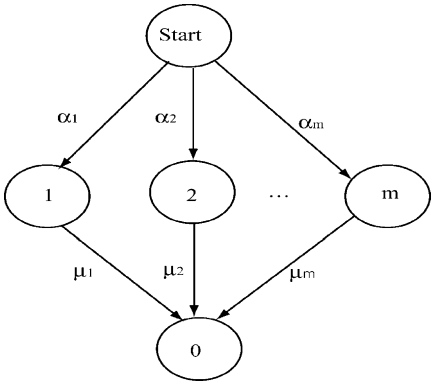


Fig. 1. A transition diagram of a Hyper-exponential distribution.

Applying (10), after some algebraic operations, we get $\rho = \frac{\lambda}{\mu}$ and using (9), we derive

$$L = (1-\rho)\boldsymbol{\alpha}_* \boldsymbol{R}(\boldsymbol{I}-\boldsymbol{R})^{-2}\mathbf{1} = \left(1-\frac{\lambda}{\mu}\right)\left[\frac{1}{m},\ldots,\frac{1}{m}\right]\boldsymbol{R}(\boldsymbol{I}-\boldsymbol{R})^{-2}\mathbf{1} = \frac{\lambda}{\mu-\lambda}. \tag{13}$$

We used the symbolic math toolbox of MATLAB to derive the last equality of (13) by plugging in any number of m greater than or equal to 1, otherwise it becomes too laborious to derive manually. Actually we found that by using the Pollaczek–Khintchine formula, the result is the same as the above. The square of coefficient of variation of the service time of the above $M/H_m/1$ queue is $C_s^2 = \frac{\mathrm{Var}[T]}{E^2[T]} = \frac{2\boldsymbol{\alpha}_* \boldsymbol{G}_*^{-2}\mathbf{1} - (-\boldsymbol{\alpha}_* \boldsymbol{G}_*^{-1}\mathbf{1})^2}{(-\boldsymbol{\alpha}_* \boldsymbol{G}_*^{-1}\mathbf{1})^2}$, which is equal to unity by plugging in (11) and (12) and after some algebraic operations. Notice here that we use the first and the second moments of CPH. So $W_q = \frac{1}{2}(1+C_s^2)\tilde{W}_q = \frac{\lambda}{\mu(\mu-\lambda)}$ where $\tilde{W}_q$ is the waiting time in queue of an $M/M/1$ queue with arrival rate λ and service rate μ.

We have just shown that if all the initial probability and service rate at each phase of an $M/H_m/1$ queue are identical, then its performance is the same as an $M/M/1$ queue. Though the above result can be easily identified on the probability density function of Hyper-exponential distribution, our purpose here is to illustrate how QBD can handle such distribution structure usually seen in a SC study. Here we use a special case to illustrate the derivation process. However the application is not limited to such special distribution form as assumed above. We have indicated how to derive all the sojourn times inside each stage for a realistic SC. Now we can use Lee and Zipkin (1992) to derive the performance measures of a tandem queue. We test the accuracy of our proposed model by employing it on a tentative multi-echelon production, transportation and distribution system as described below.

4. Implementation

4.1. A production, transportation, and distribution model

A multi-echelon production, transportation and distribution model as shown in Fig. 2 is employed as a test bed for our method. To keep the study manageable, we restrict our attention to a very basic model. The production facility (PF) produces finished goods to downstream retailers. The retailers face a stationary Poisson demand process with mean inter-arrival time of $1/\lambda$. Machining process is as introduced in section three. Successfully finished goods will leave the machine and go to the next stage for final inspection before shipping to a remote CW. After inspection, any imperfect product has to go back to the processing stage for reworking. Assume that the feedback rate is constant with probability δ. For the sake of simplicity we assume that the second (inspection) stage will never fail. Products passing inspection will wait at the shipping area, ready for transportation to CW. Upon arrival at the CW, the product will immediately be transported to the assigned retailer whenever a transporter is available. Again, for the sake of simplicity, we restrict all transporting vehicles between any two sites to one. Assume that all the transportation times are stochastic. Applying the method as described in section three, we formulate this problem as a tandem

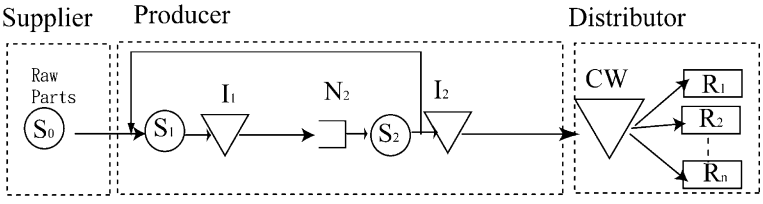


Fig. 2. A production facility with feedback having two machining stages, and one remote central warehouse serving multiple retailers.

queue with five independent stages. The first stage is the production stage with the unreliable machine being formulated as two ''on'' and ''off'' phases. The second stage is the inspection stage. The transportation from PF to CW and the CW itself are formulated as respective $M/M/1$ queuing systems. Finally, the distribution stage is formulated as a phase type, even though it is easier to accumulate all the retailers as one single stocking site, and treat it as an $M/M/1$ alike, as shown in section three.

Please note that we omitted the other N_j and I_j except those of PF in Fig. 2. Specifically, N_3, the input queue of the transit from PF to CW; I_3, the output buffer of the transit from PF to CW; N_4, the input queue of CW; I_4, the output buffer of CW, N_5, the input queue of the transit from CW to retailer; I_5, the output buffer of the transit from PF to CW, which is set to the accumulative retailer inventory level in this design. Further, assume there is an infinite supply at the first stage.

I_3 is always zero, assuming the MTO policy is adopted by this service. At CW, it is reasonable to adopt the MTS policy to lessen the customer order waiting time. And, assume that the warehouse processes its inventory with high efficiency at near zero operation time. This means that each arriving good will be put into stock immediately if there is no backorder recorded. When there is a backorder, the arriving unit will be shipped to the waiting retailer. So, N_4 is always zero as well. Assume that all products satisfying customer demands are all the same. If the customer order arrives, and the stock is out, a situation, which the MTO-type control is sure to encounter, unfilled orders are backlogged and will be satisfied when replenishing goods arrive on a FCFS basis.

4.2. Numerical results

This subsection reports our tests of the approximation of the model illustrated in Section 4.1, and compares its predictions to estimates derived from computer simulation, as illustrated in Appendix 2. Basically we follow the same test approach as reported in Zipkin (1995) with some modifications. The queuing system at PF is just like an open Jackson network. Thus the λ_i are all identical to $\lambda/(1-\delta)$, where δ is the feedback rate. So all ρ_i are equal to $\rho = \lambda/[\mu(1-\delta)]$. To test the taxing condition on the performance of the approximation, we fix $\delta = 0.5$. ρ is determined by λ/μ. Assume that the mean demand rate for each retailer is 0.25 and that there are four retailers. So the combined demand rate is 1. We fix μ to be either 2.5 or 4, and thus ρ is 0.8 or 0.5, respectively. Assume mean failure and repair rate to be 0.25 and 2.5, respectively. And assume that the average transportation time of a back and forth traveling cycle is 1/4. We adopt a similar simulation stopping criteria as reported in Lee and Zipkin (1992) and Zipkin (1995). Each run simulates 30 replications of 10,000 time units. Assume there is a holding cost of 0.5 for work-in-process per unit and per unit time, a holding cost of 1 for the end retailer inventory per unit and per unit time, a backorder cost of 10 for unfilled retailer orders per unit and per unit time. Five key performance measures are measured, TC (the total incurred cost of operating the chain, which is equal to $0.5 \cdot \mathrm{WIP} + E[I] + 10 \cdot E[B]$, see below), SL (average service level measured in no stock-out probability at the retailer site), WIP (the total intermediate inventory, which is defined as all the work-in-process, inventory level at CW, and all the queues of transit, $I1 + N2 + I2 + N3 + I4 + N5$, in this case), $E[I]$ (average retailer inventory, which is omitted for space consideration), $E[B]$ (average retailer backorder). Note that in calculating WIP, $I3$ and $N4$ are always zeroes, as described in Section 4.1. Tables 1 and 2 summarize the results. Note that the parameter setting of Table 1 is the same as in Zipkin (1995).

Also notice that the 'SL' column is not listed in Table 1 since they are all zeros. The column labeled S_j is the initial base stock level at the respective stages. The column labeled 'Sim' represents the simulation estimates; 'App' stands for the approximation, and '%Err' is the percentage error of the approximation compared to the simulation value. It is evident that the approximation is quite accurate for Table 1 with all retailers adopting base stock policies with $S_5 = 0$. Table 2 shows the results when all retailers adopt base stock policies with $S_5 \neq 0$. Also, we adjusted the stock levels for all the other stages according to base stock levels of Table 1. From Table 2, we see that when $S_5 \neq 0$, the accuracy of the matrix approximation

Table 1
Approximation vs. simulation ($S_5 = 0$)

ρ	S_1	S_2	S_3	S_4	TC			WIP			$E[B]$		
					Sim	App	%Err	Sim	App	%Err	Sim	App	%Err
0.5	0	0	0	0	30.605	30.177	−1.40	1.684	1.668	−0.95	2.949	2.93	−0.64
0.5	1	1	0	1	13.237	13.019	−1.65	2.825	2.8908	2.33	1.182	1.1574	−2.08
0.5	3	1	0	1	11.388	10.930	−4.02	4.359	4.597	5.46	0.921	0.863	−6.30
0.5	1	3	0	1	9.029	9.011	−0.20	4.322	4.414	2.13	0.687	0.680	−1.02
0.5	1	1	0	3	8.494	8.425	−0.81	4.28	4.358	1.82	0.635	0.625	−1.57
0.5	1	1	0	5	7.345	7.420	1.02	6.08	6.167	1.43	0.43	0.434	0.93
0.8	0	0	0	0	118.03	125.01	5.91	4.625	4.668	0.93	11.825	12.268	3.75
0.8	1	1	0	1	93.182	97.455	4.59	4.821	4.900	1.64	9.077	9.500	4.66
0.8	3	1	0	1	84.620	83.820	−0.95	5.436	5.507	1.31	8.19	8.107	−1.01
0.8	1	3	0	1	75.606	81.049	7.20	5.126	5.243	2.28	7.304	7.843	7.38
0.8	1	1	0	3	76.335	80.931	6.02	5.154	5.232	1.51	7.376	7.832	6.18
0.8	1	1	0	5	65.185	66.975	2.75	5.828	5.807	−0.36	6.227	6.407	2.89

Table 2
Approximation vs. simulation ($S_5 \neq 0$)

ρ	S_1	S_2	S_3	S_4	S_5	TC			SL			WIP			$E[B]$		
						Sim	App	%Err	Sim	App	%Err	Sim	App	%Err	Sim	App	%Err
0.5	4	4	0	4	4	9.309	9.512	2.18	0.919	0.994	8.16	10.636	11.079	4.17	0.031	0.029	−6.45
0.5	12	4	0	4	4	13.167	13.497	2.51	0.922	0.994	7.81	18.396	19.074	3.69	0.028	0.027	−3.57
0.5	4	12	0	4	4	13.248	13.484	1.78	0.923	0.994	7.69	18.596	19.068	2.54	0.026	0.026	0
0.5	4	4	0	12	4	13.252	13.482	1.74	0.923	0.994	7.69	18.607	19.067	2.47	0.026	0.026	0
0.5	4	4	0	4	12	16.969	17.198	1.35	0.999	1	0.10	10.624	11.079	4.28	0	0	N/A
0.5	4	4	0	4	20	24.964	25.194	0.92	1	1	0	10.624	11.079	4.28	0	0	N/A
0.8	4	4	0	4	4	29.772	29.923	0.51	0.649	0.702	8.17	8.125	8.331	2.53	2.347	2.335	−0.50
0.8	12	4	0	4	4	23.469	20.958	−10.70	0.742	0.836	12.67	13.379	14.532	8.62	1.416	1.075	−24.09
0.8	4	12	0	4	4	19.066	19.118	0.27	0.816	0.882	8.09	13.947	14.115	1.21	0.930	0.889	−4.44
0.8	4	4	0	12	4	22.175	19.119	−13.78	0.813	0.882	8.49	13.744	14.114	2.69	1.026	0.889	−13.38
0.8	4	4	0	4	12	23.176	21.999	−5.08	0.866	0.885	2.19	8.125	8.331	2.53	0.976	0.888	−9.05
0.8	4	4	0	4	20	25.047	23.902	−4.57	0.944	0.956	1.27	8.125	8.331	2.53	0.407	0.333	−18.08

method is also satisfactory. From Table 2 several useful observations can be made. For example, in the case of $S_5 \neq 0$ with $\rho = 0.5$, an increasing stock level at different stages, except at the last stage, seems to have the same effect of performance influence. The total cost and WIP levels increase and the service levels increase very limitedly while backorder levels decrease slightly. On the other hand, an increasing stock level at the last stage, i.e., retailer inventory level, does increase the service levels and decreases the backorder level, however it does so at the price of higher total cost, which is due to higher retailer inventory levels. This agrees with our intuition.

To conclude, the approximation does seem to work well for all the retailers adopting either MTO or MTS operational strategies with one-for-one replenishment policies. When we incorporate all the stochastic features, including imperfect quality, machine breakdown, random transportation, and random distribution in the system, the degradation of the accuracy is only slight, and is often within the tolerance limits of industrial use. The feedback factor can be treated as capacity loss as concluded in Zipkin (1995).

5. Discussion and sensitivity analysis

For a tandem queue without feedback, every stage behaves just like an independent $M/M/1$ service system. The sojourn time is exact by applying Little's formula $W_s = \frac{1}{\mu-\lambda}$ in each stage, which is not influenced by base-stock setting at each stage. The matrix-algebraic solution of the performance evaluation is approximately correct as reported in Lee and Zipkin (1992). However, the sojourn time varies in the first stage when there is feedback. We compared our findings with the numerical results of Zipkin (1995), which are shown in Tables 3 and 4. Looking at Table 3, which is a two-stage system, it is apparent that the sojourn time (ST) at stage 1 increases when the stock level at stage 1 (S_1) increases for both traffic intensity rates (0.5 and 0.8). Fig. 3 shows this tendency for $\rho = 0.5$. We can see that ST starts from 0.5, when $S_1 = 0$, and then increases when S_1 increases until it finally converges at near 0.7 when S_1 is near 20. After modifying the sojourn time at stage 1, which is obtained by simulation, and applying it to the matrix approximation procedure, we get a closer match between approximation value and simulation value for both performance values of WIP and $E[B]$. Here Sim(1) is the simulation values adopted from Zipkin (1995) for comparison. Sim(2) represents the results from our own simulation model. It shows great agreement when compared

Table 3
A two-stage system

ρ	S_1	ST	WIP						$E[B](S_2 = 0)$					
			Sim (1)	Sim (2)	App	%Err (1)	%Err (2)	%Err (3)	Sim (1)	Sim (2)	App	%Err (1)	%Err (2)	%Err (3)
0.5	0	0.501	N/A	1.005	1	N/A	−0.5	N/A	N/A	2.01	2	N/A	−0.5	N/A
0.5	1	0.548	1.475	1.487	1.477	0.1	−0.7	1.7	1.56	1.57	1.573	0.8	0.2	−3.8
0.5	3	0.626	2.97	2.959	2.963	−0.2	0.1	5.2	1.237	1.213	1.215	−1.8	0.2	−9.1
0.5	5	0.668	4.753	4.748	4.746	−0.2	0	5.9	1.095	1.089	1.082	−1.2	−0.6	−5.9
0.8	0	2	N/A	3.997	4	N/A	0.1	N/A	N/A	7.998	8	N/A	0	N/A
0.8	1	2.089	4.145	4.233	4.2	1.3	−0.8	1.3	7.253	7.421	7.371	1.6	−0.7	−0.7
0.8	3	2.176	4.894	4.895	4.988	1.9	1.9	3.2	6.212	6.253	6.34	2.1	1.4	−2.6
0.8	5	2.299	5.983	6.014	6.121	2.3	1.8	6	5.575	5.622	5.719	2.6	1.7	−4.7

Table 4
A four-stage system

ρ	(S_1, S_2, S_3)	ST	WIP						$E[B]$ $(S_4 = 0)$					
			Sim (1)	Sim (2)	App	%Err (1)	%Err (2)	%Err (3)	Sim (1)	Sim (2)	App	%Err (1)	%Err (2)	%Err (3)
0.5	(0, 0, 0)	0.498	N/A	2.979	3	N/A	0.7	N/A	N/A	3.971	4	N/A	0.7	N/A
0.5	(1, 1, 1)	0.549	4.213	4.229	4.1446	−1.6	−2	−0.6	2.308	2.333	2.2426	−2.8	−3.9	−5.2
0.5	(3, 1, 1)	0.568	5.908	5.92	5.8492	−1	−1.2	0.8	2.036	2.057	1.9852	−2.5	−3.5	−4.1
0.5	(1, 3, 1)	0.575	5.695	5.685	5.6131	−1.4	−1.3	0.4	1.841	1.832	1.763	−4.2	−3.8	−6.6
0.5	(1, 1, 3)	0.591	5.497	5.478	5.328	−3.1	−2.7	−0.9	1.7	1.658	1.51	−11.2	−8.9	−15
0.5	(1, 1, 5)	0.621	7.104	7.093	6.9596	−2.0	−1.9	0.7	1.371	1.327	1.2016	−12.4	−9.5	−15.6
0.8	(0, 0, 0)	2.028	N/A	12.066	12	N/A	−0.6	N/A	N/A	16.126	16.056	N/A	−0.4	N/A
0.8	(1, 1, 1)	1.998	12.16	12.262	12.3033	1.2	0.3	1.2	13.213	13.256	13.2993	0.7	0.3	0.7
0.8	(3, 1, 1)	2.032	13.064	12.901	13.199	1.0	2.3	1.1	12.212	11.957	12.263	0.4	2.6	0.4
0.8	(1, 3, 1)	2.041	12.538	12.723	12.725	1.5	0	1.6	11.735	11.809	11.807	0.6	0	1.2
0.8	(1, 1, 3)	2.033	12.352	12.513	12.4949	1.12	−0.1	1.2	11.501	11.584	11.5609	0.5	−0.2	1.1
0.8	(1, 1, 5)	2.095	12.939	12.951	12.8978	−0.3	−0.4	−0.1	9.924	10.136	10.0878	1.7	−0.5	−1.8

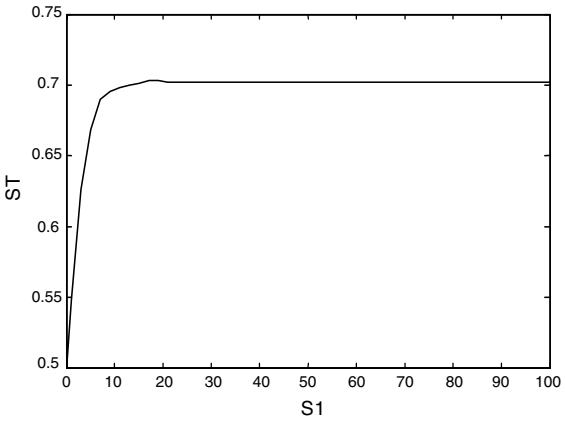


Fig. 3. The sojourn time of stage 1 as a function of S_1 for a two-stage PF with $\rho = 0.5$.

with that of Zipkin (1995). For comparison, in Table 3 we show the relative percentage error between App and Sim(1) as indicated in the %Err(1) column. The %Err(2) is the relative percentage error between App and Sim(2). The %Err(3) is the relative error before adjusting the sojourn time at stage 1, which is reported by Zipkin (1995). It is clear that the sojourn time at stage 1 does influence the accuracy of the theoretical approximation value.

The WIP and $E[B]$ of stage 2 compared to the $S1$ for a two-stage system with $\rho = 0.5$ are also shown in Figs. 4 and 5. App (adj) means the performance by applying adjusted sojourn time to the matrix solution. App ($'$adj) is the performance by not plugging in adjusted sojourn time. We can observe minor differences between the approximation and the simulation results regarding the base-stock level at stage 1.

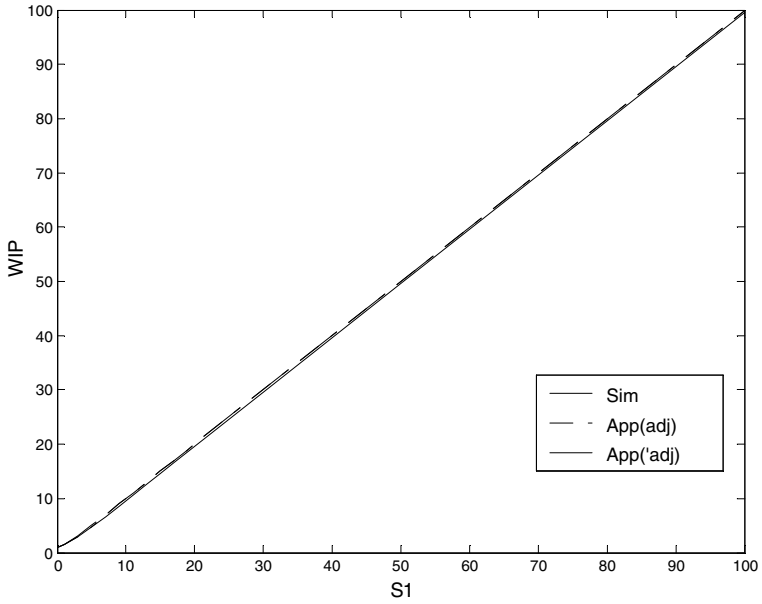


Fig. 4. WIP as a function of S_1 for a two-stage PF with $\rho = 0.5$.

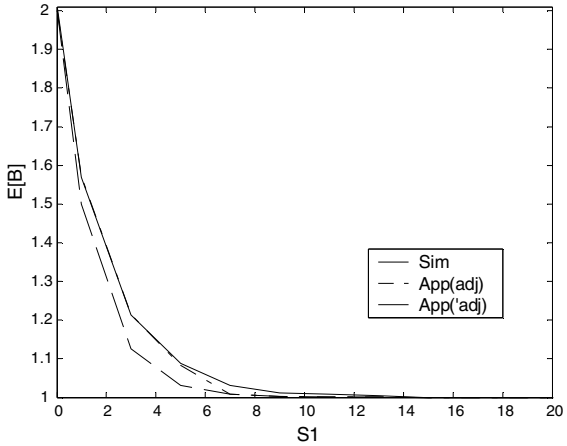


Fig. 5. $E[B]$ as a function of S_1 for a two-stage PF with $\rho = 0.5$.

Table 4 shows the comparison of simulation and approximation of a four-stage system. It seems that the accuracy does not improve as expected for a system composed of more stages.

In conclusion we find that the impact of the above analysis of the accuracy of our matrix approximation method is limited. In the worst case, the absolute error between Sim and App of WIP is only near 0.5. Therefore, there is no adjustment made in the computing code of the implementation section.

As for our tentative SC model we also tested the case when there is only feedback and no machine break down issue incorporated. Basically the difference between App and Sim is also small, when compared to the results of Section 4.2. In addition we investigated when there is only the influence of machine breakdown, and it behaved as expected when compared to the simulation results.

As a final remark, from the open Jackson network, the feedback impact on traffic intensity rate $\rho = \lambda/[\mu(1-\delta)]$ can be explained in a different way, by either increasing the input arriving rate from λ to $\lambda/(1-\delta)$ or by losing capacity from μ to $\mu(1-\delta)$. From our computing experience with the performance of a tandem queue with feedback, both methods achieve the same results. In analyzing the impact of ST of Tables 3 and 4, we used the arrival increase method. However, it ifs better to use the method of capacity loss when there is also a machine breakdown issue, otherwise the outcome will differ largely from the simulation results.

6. Extension

For the derivation of (r,q) policy, it is natural by using the fact that it is built upon base-stock policy. The key performance measures such as the steady-state backorder level can therefore be represented as the equal weighted sum of respective performance measures at different levels of inventory positions (Axsäter, 2000 or Zipkin, 2000). Specifically, after some algebraic operations, we may express the above argument as:

$$E[B] = \sum_{S=r+1}^{r+q} E[B(S)] = \left(\frac{1}{q}\right)\boldsymbol{\pi P}(\boldsymbol{I}-\boldsymbol{P})^{-2}(\boldsymbol{I}-\boldsymbol{P}^{\boldsymbol{q}})\boldsymbol{P}^{\boldsymbol{r}}\boldsymbol{e}. \tag{14}$$

Note when $q = 1$, (14) becomes (7). To illustrate our argument, assume we have a two-echelon SC: a PF directly serves four identical retailers. The PF uses base stock policy to control its inventory while the retailers use (r,q) policies to control their stocks. The demand process at the PF is not Poisson but it is a superposition of several independent renewal processes, which under suitable conditions resembles a Poisson

process (Svoronos and Zipkin, 1988). Assume the PF produces in units of retailer batches and each retailer has its dedicated transporter. Here we follow Svoronos and Zipkin (1988) and assume the arrival process at PF as Poisson processes. We then express the aggregated arrival rate at the PF as $N\lambda_R/q$, where λ_R is the arrival rate for each retailer and N is the number of the retailers. We also approximate the arrival process at the respective transit stage as Poisson process with mean λ_R/q. And finally we use the same modeling approach for CW as shown in 4.1 to model retailer activity, assuming that the retailer processes its inventory with high efficiency at near zero operation time. Alternatively, we can formulate this problem as a 2-stage SC, with respective retailer-stocks representing planned inventories at the second stage. Using (14) and the fact: WIP $= E[I_1] + E$[inventory in transit], we obtain performance measures for different combinations of inventory control parameters at each stage as listed in Table 5.

Here, S is the base stock level at PF and r an q represent reorder point and fixed order quantity at the retailers, respectively. Under the arrival assumptions at respective echelons, ρ_1 and ρ_2 are calculated traffic intensities by changing different level of q and letting λ_R fixed at either 0.5 or 0.8. And μ is fixed at 2 for the server at respective echelons (no feedback concern in this case). Also for simplicity we do not consider breakdown issue. From the table we see acceptable accuracy exists when ρ_1 is low. We also test other cases when ρ_1 is high and q is large by varying λ_R. Unfortunately, the approximation is not satisfactory for $E[B]$ on most of the test cases. Some tests show Erlang distribution may be more appropriate than the proposed Poisson distribution for the arrival process at respective echelons. However such conjecture is related to phase type arrival and needs further analytic efforts and numerical verifications.

As stated in Section 1, we may use QBD process to achieve the goal of more modeling flexibility. For example if we want to model multi-server at each subsystem with each server suffering random breakdowns, we have an $M/PH/m$ queueing system at each subsystem. To model such queueing system by using the QBD approach, we may express the state space as $nx(1)\cdots x(i)\cdots x(m)$ where $n =$ customer number, $x(i) = 0$ (down) or 1 (on), $1 \leqslant i \leqslant m$, and have 2^m states for $n \geqslant m$. And we conjecture that the QBD modeling approach for each subsystem may be treated independently from the linkage of the whole SC. To justify our argument, we employed the same 2-stage example (also no feedback concern in this case) as in Section 5 with some modifications that there are multiple parallel machines at the PF and so are there at the second stage. Specifically we assume 2 servers for each stage. Assume the parameters are the same as in 4.2. We form a QBD process for the decomposed server queue at each stage. Table 6 lists the results for different combinations of base-stock level at each stage.

Clearly the accuracy is degraded when traffic intensity is high, but not significantly serious, as compared to all the previous examples.

In short, it is possible a more general framework to accommodate for versatile control policies may be developed by combining the QBD technique and Lee and Zipkin (1992). Since the basic assumption for the

Table 5
Approximation vs. simulation for the case where retailers use (r, q) policies

ρ_1	ρ_2	S	r	q	WIP			$E[B]$		
					Sim	App	%Err	Sim	App	%Err
0.5	0.125	0	0	2	0.53	0.571	7.74	0.659	0.83	25.95
0.25	0.063	1	0	4	1.001	1.016	1.50	0.091	0.12	31.87
0.167	0.042	3	0	6	2.978	2.974	−0.13	0.041	0.045	9.76
0.125	0.031	5	0	8	4.988	4.986	−0.04	0.031	0.033	6.45
0.8	0.200	0	0	2	0.915	1	9.29	4.287	5.777	34.76
0.4	0.100	1	0	4	1.003	1.044	4.09	0.309	0.48	55.34
0.267	0.067	3	0	6	2.939	2.929	−0.34	0.109	0.129	18.35
0.2	0.050	5	0	8	4.967	4.96	−0.14	0.081	0.088	8.64

Table 6
Approximation vs. simulation for cases with multi-server and breakdowns

ρ	S_1	S_2	WIP			$E[B]$		
			App	Sim	%Err	App	Sim	%Err
0.5	0	0	1.597	1.620	−1.42	3.594	3.224	11.48
0.5	1	0	1.982	1.883	5.26	2.579	2.491	3.53
0.5	3	0	3.371	3.256	3.53	1.968	1.853	6.21
0.5	5	0	5.140	5.078	1.22	1.737	1.676	3.64
0.8	0	0	8.214	7.076	16.08	16.428	13.846	18.65
0.8	1	0	8.322	7.176	15.97	15.536	13.049	19.06
0.8	3	0	8.819	7.355	19.90	14.033	10.981	27.79
0.8	5	0	9.624	8.336	15.45	12.839	10.237	25.42

approximation model of Lee and Zipkin (1992) is that the queueing system at each subsystem is independent. And the QBD approach often faces the problem of largeness, i.e., too many states may make the solution intractable (for example, for the above mentioned $M/PH/m$ queueing system, if we have 20 parallel machines, the states become more than one million). The challenge lies in how large and how sophisticated the QBD modeling approach can allow as well as how accuracy this combining process can provide. All these need further study as well as thorough numerical verifications.

7. Conclusions

We have demonstrated that by using the matrix analytical approach, the evaluation of a complex SC where all the participants, including PF, transporters, CW and retailers use base-stock control policies, performs as expected through simulation verification. The relative errors between App and Sim are all below 10% for retailers adopting MTO policies. When all the retailers adopt the MTS policy, numerical studies also show that the approximation is accurate for medium traffic intensity and acceptable for high traffic intensity. In this present study the results are somehow similar to those of Zipkin (1995) where the base stock level at the end stage is set to zero. The present study shows that the matrix analytical approach is very accurate, not just for the application of tandem processing queue as reported by Lee and Zipkin (1992) and Zipkin (1995) but also for the application of tandem SC where the end stage can be of a distribution system with identical retailers. In the literature on the stochastic production-distribution system, most models are developed and analyzed separately. Unlike our model, these evaluation models are usually difficult to integrate as one single model.

The most significant contribution of this paper is that we proposed an originative and useful system design and analysis tool for evaluating the performance of an integrated stochastic SC. Although a rich body of multi-echelon inventories systems in the literature, which use the same base-stock policies as we used herein, our idea is to provide a viable scheme for solving integrated stochastic supply network in a flexible and realistic way. So we used the simplest inventory control scheme of base-stock as the first step towards more involved inventory control technique. Under the matrix analytical approach, decision makers can easily formulate stochastic and/or factors of uncertainty, which are often encountered in real life, as adequate queuing form and later integrate them together as a single tandem queue. The performance measures are then readily available by simple matrix-manipulated computation.

In this study, we also found that, the Hyper-exponential queue $M/H_m/1$ can be used adequately to model a distribution subsystem of a supply chain. The phase-type structure can then be handled as a usual QBD process. We illustrate how it works by proposing a special structure, under which the distribution subsystem behaves just like an $M/M/1$ queue. However numerical studies show this modeling approach is not lim-

ited to such special form (we omit the numerical details herein). We believe this modeling approach is new in supply chain study. The other finding is that the behavior of the beginning stage of a tandem queue may differ from the other stages when MTS stock policies are applied throughout the chain except for the end stage. This seems to violate the inherent theory of an open Jackson network. However, sensitivity study shows that the matrix analytical approach still approximates well.

In the extension section we test the applicability of the proposed approach for another control scheme as well as for multi-server setting. We employed two 2-echelon problems. Numerical studies of (r, q) policy are satisfactory for low to medium traffic intensities when arrival rates are fixed at either 0.5 or 0.8. In the multi-server case the approximated results are more satisfactory with medium traffic intensities than with heavy traffic intensities. Generally speaking, we see the promising future of the proposed model as a quick and accurate SC evaluation tool not just for base stock inventory control schemes but also for (r, q) policy employed at the retailer site if traffic intensity of the studied queueing system is medium or low. Other control policies of pull type such as KANBAN, etc., may be incorporated into the current model. However it needs more involved analytic techniques. Another advantages of the current study is that the proposed method herein seems more tractable when compared with existing multi-echelon stochastic models in the literature, which often used more involved stochastic process to derive performance measures of interests. Finally, the closed-form solutions of the current model may be used as later SC optimization applications.

Acknowledgements

We would like to thank Professor Paul Zipkin for his helpful comments during the development of this research. We are also grateful to the anonymous referees for their helpful comments.

Appendix 1. A QBD process

Define the phase type distribution of a Markov process as a stochastic process having parameters m, G_*, and α_* if it can be expressed as a first-passage time random variable, that is, $T = \min\{t \geqslant 0: Y_t = \Delta\}$, for a Markov process with state space $E = \{1, \ldots, m, \Delta\}$, generator

$$\boldsymbol{G} = \left[\begin{array}{c|c} \boldsymbol{G}_* & \boldsymbol{G}_\Delta \\ \hline \boldsymbol{0} & 0 \end{array}\right],$$

and the initial probability vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}*|0)$. The generator of an $M/PH/1$ queue can then be represented in the following matrix form:

$$\boldsymbol{Q} = \begin{bmatrix} -\lambda & \lambda\boldsymbol{\alpha}_* & & & & \\ \boldsymbol{G}_\Delta & \boldsymbol{G}_* - \boldsymbol{\Lambda} & \boldsymbol{\Lambda} & & & \\ & \boldsymbol{G}_\Delta\boldsymbol{\alpha}_* & \boldsymbol{G}_* - \boldsymbol{\Lambda} & \boldsymbol{\Lambda} & & \\ & & \boldsymbol{G}_\Delta\boldsymbol{\alpha}_* & \boldsymbol{G}_* - \boldsymbol{\Lambda} & \boldsymbol{\Lambda} & \\ & & & \boldsymbol{G}_\Delta\boldsymbol{\alpha}_* & \boldsymbol{G}_* - \boldsymbol{\Lambda} & \cdots \\ & & & & \vdots & \ddots \end{bmatrix}, \tag{A.1}$$

where $\boldsymbol{\Lambda} = \lambda\boldsymbol{I}$. Then we can find the steady-state probability vector $\boldsymbol{p} = (p_0|p_{11}, p_{12}, \cdots |p_{21}, p_{22}, \cdots |\cdots) = (p_0|\boldsymbol{p}_1|\boldsymbol{p}_2|\cdots)$ as follows.

Combining (A.1) with the equations $\boldsymbol{PQ} = 0$ yields the following system of equations,

$$\begin{aligned}
&-\lambda p_0 + \boldsymbol{p}_1 \boldsymbol{G}_{\Delta} = 0,\\
&\lambda \boldsymbol{\alpha}_* p_0 + \boldsymbol{p}_1(\boldsymbol{G}_* - \boldsymbol{\Lambda}) + \boldsymbol{p}_2 \boldsymbol{G}_{\Delta} \boldsymbol{\alpha}_* = 0,\\
&\boldsymbol{p}_1 \boldsymbol{\Lambda} + \boldsymbol{p}_2(\boldsymbol{G}_* - \boldsymbol{\Lambda}) + \boldsymbol{p}_3 \boldsymbol{G}_{\Delta} \boldsymbol{\alpha}_* = 0,\\
&\boldsymbol{p}_2 \boldsymbol{\Lambda} + \boldsymbol{p}_3(\boldsymbol{G}_* - \boldsymbol{\Lambda}) + \boldsymbol{p}_4 \boldsymbol{G}_{\Delta} \boldsymbol{\alpha}_* = 0.\\
&\vdots
\end{aligned} \tag{A.2}$$

The solution of (A.2) involves the characteristic equation

$$\boldsymbol{\Lambda} + \boldsymbol{R}(\boldsymbol{G}_* - \boldsymbol{\Lambda}) + \boldsymbol{R}^2 \boldsymbol{G}_{\Delta} \boldsymbol{\alpha}_* = 0. \tag{A.3}$$

The matrix-geometric solution of (A.3) is

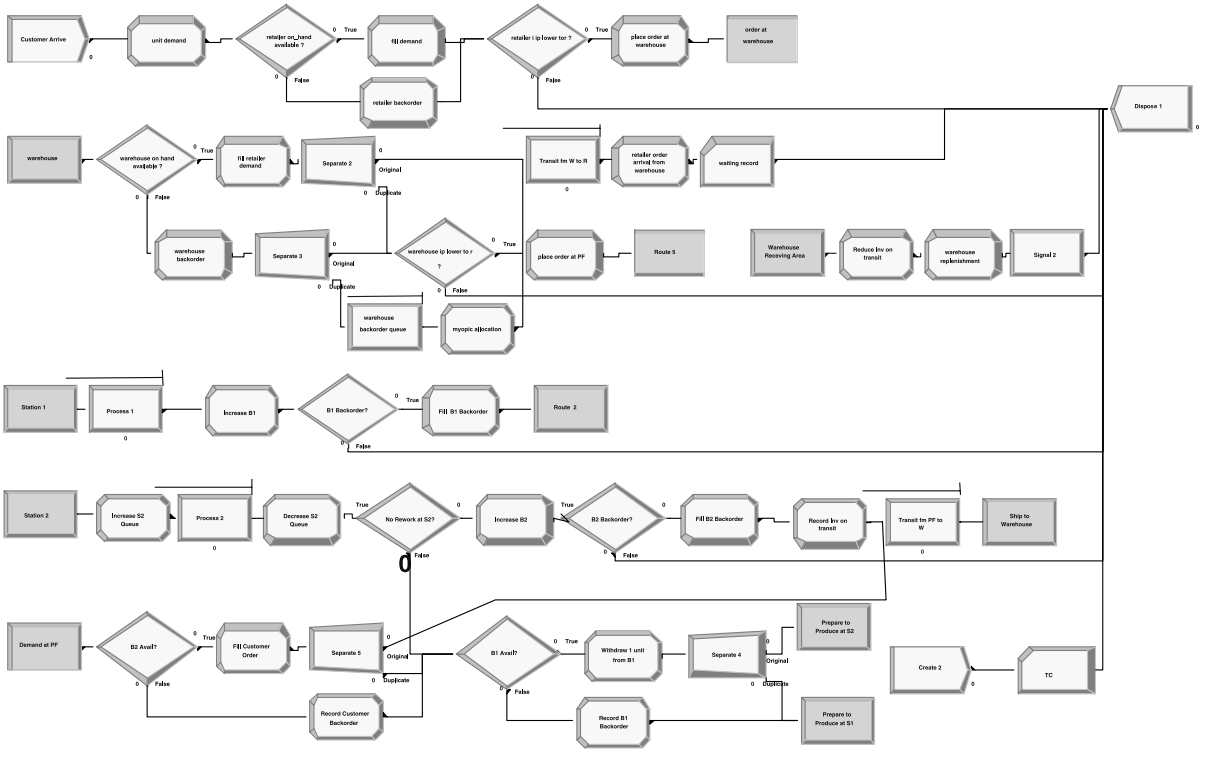
$$\boldsymbol{p}_n = \boldsymbol{c}\boldsymbol{R}^n \quad \text{for} \quad n = 1, 2, \ldots, \tag{A.4}$$

where $\boldsymbol{c}$ is a vector of constants. It can be shown (Neuts, 1981, p. 84.) that

$$\boldsymbol{R} = \lambda(\boldsymbol{\Lambda} - \boldsymbol{\lambda}\boldsymbol{1}\boldsymbol{\alpha}_* - \boldsymbol{G}_*)^{-1}. \tag{A.5}$$

And $\boldsymbol{c} = p_0 \boldsymbol{\alpha}_* = (1 - \rho)\boldsymbol{\alpha}_*$, where $\rho = \frac{\lambda}{\mu} = -\lambda \boldsymbol{\alpha}_* \boldsymbol{G}_*^{-1}\mathbf{1}$, $\mathbf{1}$ is a column vector of ones, whose dimension is chosen to fit the context.

Appendix 2. Simulation Model of Section four, which is implemented using ARENA 5.0



References

Abboud, N.E., 2001. A discrete-time markov production-inventory model with machine breakdowns. Computers & Industrial Engineering 39, 95–107.

Axsäter, S., 1993a. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. Operations Research 41 (4), 777–785.

Axsäter, S., 1993b. Continuous review policies for multi-level inventory systems with stochastic demand. In: Graves, S.C. et al. (Eds.), Logistics of Production and Inventory, Handbooks in OR and MS, vol. 4. Elsevier (North-Holland), Amsterdam, pp. 175–197 (Chapter 4).

Axsäter, S., 2000. Exact analysis of continuous review (r, q) policies in two-echelon inventory systems with compound Poisson demand. Operations Research 48 (5), 686–696.

Buzacott, J.A., Shanthikumar, J.G., 1993. Stochastic Models of Manufacturing Systems. Prentice Hall, New Jersey.

Cohen, M.A., Lee, H.L., 1988. Strategic analysis of integrated production-distribution systems: models and methods. Operations Research 36 (2), 216–228.

Duri, C., Frein, Y., Di Mascolo, M., 2000. Performance evaluation and design of base stock systems. European Journal of Operational Research 127 (2000), 172–188.

Enginarlar, E., Jingshan, L., Meerkov, S.M., Zhang, R.Q., 2002. Buffer capacity for accommodating machine downtime in serial production lines. International Journal of Production Research 40 (3), 601–624.

Feldman, R.M., Valdez-Flores, C., 1995. Applied Probability & Stochastic Processes. PWS Publishing Co., Boston.

Gurgur, G.Z., 2002. Performance analysis and capacity planning of multi-stage, multi-product, decentralized and market-driven manufacturing systems. Ph.D. Dissertation, Graduate School, New Brunswick Rutgers, The State University of New Jersey.

van Houtum, G.J., Inderfurth, K., Zijm, W.H.M., 1996. Materials coordination in stochastic multi-echelon systems. European Journal of Operational Research 95, 1–23.

Kalpakam, S., Sapna, K.P., 1997. A lost sales inventory system with supply uncertainty. Computers and Mathematical Application 33 (3), 81–93.

Lee, Y.-J., Zipkin, P.H., 1992. Tandem queues with planned inventories. Operations Research 40 (5), 936–947.

Mahmut, P., Perry, D., 1995. Analysis of a (Q, R, T) inventory policy with deterministic and random yields when future supply is uncertain. European Journal of Operational Research 84, 431–4434.

Mohebbi, E., 2003. Supply interruptions in a lost-sales inventory system with random lead time. Computers & Operations Research 30, 411–426.

Neuts, M.F., 1994. Matrix-geometric Solutions in Stochastic Models. Dover Publications Inc., New York.

Pyke, D.F., Cohen, M.A., 1993. Performance characteristics of stochastic integrated production-distribution systems. European Journal of Operational Research 68, 23–48.

Pyke, D.F., Cohen, M.A., 1994. Mutiproduct integrated production-distribution systems. European Journal of Operational Research 74, 18–49.

Svoronos, A., Zipkin, P., 1988. Estimating the performance of multi-level inventory systems. Operations Research 36 (1), 57–72.

Svoronos, A., Zipkin, P., 1991. Evaluation of one-for-one replenishment policies for multi-echelon inventory systems. Management Science 37, 68–83.

Zipkin, P., 1988. The use of phase-type distributions in inventory-control models. Naval Research Logistics 35, 247–257.

Zipkin, P., 1995. Processing networks with planned inventories: tandem queues with feedback. European Journal of Operational Research 80, 344–349.

Zipkin, P., 2000. Foundations of Inventory Management. McGraw-Hill, New York.