Performance Evaluation Models for Single-Item Periodic Pull Production Systems

Author(s): Nureddin Kirkavak and Cemal Dinçer

Source: *The Journal of the Operational Research Society*, Vol. 47, No. 2 (Feb., 1996), pp. 239-250
Published by: Palgrave Macmillan Journals on behalf of the Operational Research Society
Stable URL: https://www.jstor.org/stable/2584345

Journal of the Operational Research Society (1996) 47, 239–250 0160-5682/96 $12.00

Performance Evaluation Models for Single-item Periodic Pull Production Systems

NUREDDIN KIRKAVAK and CEMAL DINÇER
Bilkent University, Turkey

A number of pull production systems reported in the literature are found to be equivalent to a tandem-queue so that existing accurate tandem-queue approximation methods can be used to evaluate such systems. In this study, we consider developing an exact performance evaluation model for a non-tandem-queue equivalent pull production system using discrete-time Markov processes. It is a periodically controlled serial production system in which a single-item is processed at each stage with an exponential processing time in order to satisfy the Poisson finished product demand. The selected performance measures are throughput, inventory levels, machine utilizations and service level of the system. For large systems, which are difficult to evaluate exactly because of large state-spaces involved, we also propose a computationally feasible approximate decomposition technique together with some numerical experimentations.

Key words: approximate decomposition, Markov processes, performance evaluation, pull production

INTRODUCTION

In the 1970s, the Just-In-Time (JIT) philosophy was introduced into the production literature and has produced an alternative production control system (Kanban System) as an offspring. Golhar and Stamm[1] offer a comprehensive review and provide a framework for classifying the related JIT literature. The first successful example of development and implementation of the JIT concept as a material management system has been reported by Sugimori *et al.*[2] in the Toyota Motor Company describing their production system. At Toyota, the system is actually operated by means of kanbans. The kanban material management system is well described by Sugimori *et al.*[2] and Kimura and Terada[3]. It acts as the nervous system of the JIT production system whose functions are to direct in-process materials just-in-time to the workstations and to pass information as to what and how much to produce. In such systems, the kanbans pull in-process materials from one workstation to another to meet the demand at each workstation at the right time.

In practice, there are many alternative forms of pull production systems that differ in some design or operating characteristics[4]. However, the pull system is commonly distinguished from the conventional push method of production control by the existence of finite buffers for in-process materials and the production triggering process that depends on the inventory level of the succeeding buffer stocks. The well-known pull systems are kanban-controlled production lines.

The simplest form of pull production control system is called a *base stock* system. There exists a single inventory buffer between each workstation. The maximum inventory level permitted in this intermediate buffer is called the base stock level. Each time the downstream workstation (the one closer to final demand) requires in-process material, it withdraws one unit from the intermediate buffer. Production of one unit is then triggered at the upstream workstation since the inventory level falls below the base stock level. Production stops (workstation is blocked) when the inventory level of the buffer reaches the base stock level. Note that the downstream workstation pulls the required in-process materials, which are processed at the upstream workstation.

LITERATURE REVIEW

Many of the kanban systems described in the production literature are equivalent to a tandem queue[5]. A tandem queue is a set of finite queues in series. Note that for two particular queueing

Correspondence: N. Kirkavak, Department of Industrial Engineering, Eastern Mediterranean University, Gazi Mağusa/TRNC, Mersin 10, Turkey

systems to be equivalent to each other, they must have the same joint queue length distribution[5–7]. This is simply because most of the key performance measures are computed using joint queue length distributions. Berkley[5] showed when and how tandem queues can be used to obtain the performance measures of a two-card kanban-controlled pull production line. Two-card kanban systems are designed for batch manufacturing environments where materials in-process are handled periodically. In a two-card system, production kanbans serve as work orders to replace containers withdrawn and withdrawal kanbans serve as material requisitions for the periodic material handling operation (see Figure 1).

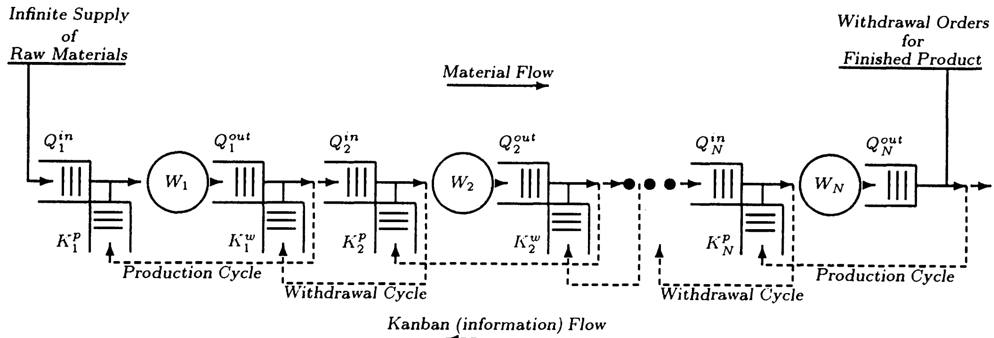

FIG. 1. *Tandem arrangement of workstations in a two-card kanban-controlled pull production line.*

There has been a significant accumulation in the literature of tandem queueing models of production lines over the last 30 years. Various design and operating aspects of these systems have been studied. The exact analysis mostly focused on the special structure of the underlying Markov chains and solved the associated Chapman–Kolmogorov balance equations for the steady-state probabilities[8–10]. As the state-space of the system under study increases, the use of exact methods becomes computationally infeasible because of the magnitude of computational effort and the computer space requirements. The only remaining viable approach for the analysis of large-scale systems appears to be the use of approximation techniques. In an approximate analysis, the system is decomposed into smaller (one or two-node) subsystems that are analysed in isolation and are then related to each other in an iterative manner to obtain the performance measures of the whole system[11–13].

In the 1980s, to represent more general distributions in queueing systems, Altiok[14] introduced phase-type distributions into the production literature. This provided an alternative approach to modelling several issues of production systems and also to approximating general distributions to be used in analytical models as well as in simulation. Altiok and Stidham[15] used a two-stage phase-type distribution, which exactly represents the process completion time distribution of jobs in a system of exponential servers subject to exponential failures and repairs. The advantage of this formulation is that it also provides the approximate representations of tandem queues with general processing times with or without breakdowns. The resultant systems of queues in tandem with phase-type service time distributions and with finite queue capacities were studied through decomposition approximations by Altiok[11].

There have been a number of attempts at developing analytical models that provide insights into how pull production systems perform. Wang and Wang[16] developed a Markov model for determining the number of kanbans required in a serial JIT system, in which assembly-type operations were allowed. By evaluating Markov chains for an alternative number of production kanbans, they found a solution that minimizes total inventory holding and shortage costs. Recently, Meral[17] developed an analytical model in order to investigate the workload allocation problem on ideal JIT production systems (one kanban at each stage). She also proposed a decomposition approach to handle longer production lines.

The periodic pull system formulated by Kim[18] is a single-item, multi-stage production line utilizing a two-card kanban control system with a fixed withdrawal cycle time. Sarker and

Parija[19] developed a mathematical model to find an optimal batch size for a JIT production system operating under a fixed-quantity, periodic delivery policy. The system they considered procures raw materials from suppliers, processes them and finally delivers the finished products demanded by outside buyers at fixed interval points in time. Deleersnyder *et al.*[20] formulated a discrete time stochastic model in order to demonstrate the key features of JIT production philosophy. The dimensionality problem associated with Markov chains restricts the applicability of this type of model to lines having a relatively small number of workstations (typically not more than three). Recently, Berkley[21] introduced a decomposition approximation using embedded Markov chains for kanban-controlled pull production lines with periodic material handling and Erlang processing times. So and Pinault[22] estimated the amount of buffer stocks needed at each station in order to meet a predetermined level of performance by utilizing an approximation in which the whole system was decomposed into individual $M/M/1$ queues with bulk service. Mitra and Mitrani[23] described an alternative decomposition for a single-card kanban system, which is equivalent to So and Pinault's model. The finished products were assumed to be immediately withdrawn from the system. In another study by Mitra and Mitrani[24], an exogenous demand process was introduced so that their first study turned out to be a special case corresponding to heavy demand arrivals. Analysing the sample path descriptions, Mitra and Mitrani[24] also showed that systems under consideration became equivalent to a tandem queue when the input material queues are eliminated.

Buzacott[25] developed a linked queueing network model to describe the behaviour of a kanban-controlled production system. He pointed out that kanban-controlled systems can be shown to be particular cases of a more general inventory level triggered approach to production control. On the other hand, Badinelli[26] presented a descriptive model for steady-state performance of a serial inventory system in which each facility follows a continuous-review pull policy under stochastic demand. In this model, each downstream facility orders a fixed amount, Q, from the upstream facility whenever the inventory position at the intermediate buffer reaches a reorder point, R.

DESCRIPTION OF THE SYSTEM

In the context of operational design, the periodic review and periodic material handling issues are the widely encountered characteristics in practice for pull production systems[18]. In such periodic pull systems, the transfer of work-in-process (WIP) inventories between stages and the release of collected kanbans as production orders to workstations are initiated at the beginning of the periods. In this study we investigate the steady-state behaviour of a non-tandem-queue (NTQ) equivalent pull production system. To this end it is formulated as a discrete-time Markov process. Note that, a discrete-time model can satisfactorily approximate the continuous model by sufficiently squeezing the time periods.

This basic system consists of N stages in tandem (see Figure 2). At each stage there is only one workstation processing a single-item, so that the term 'stages' and 'workstations' could be used interchangeably. W_j $(1 \leqslant j \leqslant N)$ represents workstations. At any workstation W_j, there are two stocks Q_j^{in} and Q_j^{out} respectively for storing incoming and outgoing WIP inventory items at workstation W_j. W_1 is responsible for the first operation of the item, converting raw material RM (or, alternatively, denoted by component C_0 stored in stock Q_1^{in}) into component C_1 (stored in stock Q_1^{out} until the end of the period then instantaneously transferred to stock Q_2^{in}). W_j $(2 \leqslant j \leqslant N-1)$ converts component C_{j-1} (from stock Q_j^{in}) into component C_j (stored in Q_j^{out} until the end of the period then instantaneously transferred to stock Q_{j+1}^{in}). W_N performs the final operation of the item, converting component C_{N-1} (from stock Q_N^{in}) into finished product FP (which could alternatively be denoted by C_N and stored in Q_N^{out} until the end of the period then instantaneously transferred to Q_{FP} or, alternatively, Q_{N+1}^{in}). The maximum number of items allowed in stocks Q_j^{out} and Q_{j+1}^{in} is K_j; that is, the maximum capacity of buffer space allocated for component C_j between workstations W_j and W_{j+1}. Note that I_j^{in} $(0 \leqslant I_j^{in} \leqslant K_{j-1})$ and I_j^{out} $(0 \leqslant I_j^{\text{out}} \leqslant K_j)$ denote the level of WIP inventories at stocks Q_j^{in} and Q_j^{out}, respectively. Consider the total number of component C_j items between workstations W_j and W_{j+1}, then the inequality for the current level of WIP inventories at stocks Q_j^{out} and Q_{j+1}^{in}; $I_j^{\text{out}} + I_{j+1}^{in} \leqslant K_j$ holds.

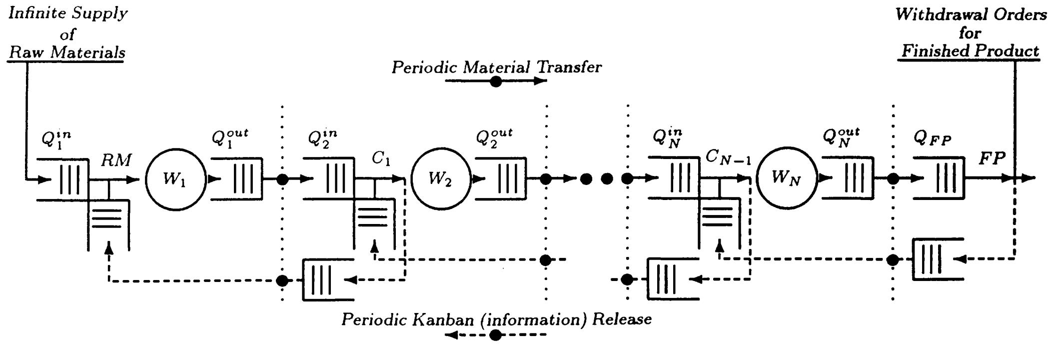

FIG. 2. *Kanban-controlled periodic pull production line.*

For simplification, the rate of supply of RM is assumed to be infinite. Since a kanban-controlled pull production system typically operates with small lot sizes, it is assumed that one kanban corresponds to one item of inventory in this formulation. The analysis can be easily extended to cover the systems operating with lot sizes greater than one at a cost of the dimensionality problem in evaluating transition matrices. In these periodic pull systems, the production is only initiated just for the replenishment of items removed from the buffer stocks during the material handling and inventory review period (transfer/review cycle time) of T time units. Workstation W_j produces components C_j in order to maintain the inventory level of stock Q_{j+1}^{in} at K_j.

At the end of period k, first the components collected at outgoing stocks ($I_j^{\text{out}}(k)$ units of component C_j) are transferred to incoming stocks Q_{j+1}^{in} in the context of the material handling function. Then, in the context of the production/inventory control function, the total number of kanbans released as production orders to start production of components C_j at workstation W_j for the period $k+1$ becomes $K_j - I_{j+1}^{\text{in}}(k+1)$. Note that the convention used in this study is the 'beginning of period' in evaluating any state parameter of the system with the exception of $I_j^{\text{out}}(k)$, which denotes the inventory level of stock Q_j^{out} at the end of the period k, since all output buffers are empty at the beginning of any period.

The two sources of uncertainty considered in this system are the demand and processing time variability. The demand for the finished product FP arrives with exponentially distributed inter-arrival times to the buffer stock Q_{FP}. The mean inter-arrival time of the demand is $(1/\lambda)$. For simplification, backorders are not considered in this formulation, so an arriving finished product demand finding zero FP items at Q_{FP} (that means, I_{N+1}^{in} or alternatively I_{FP} is zero) is lost. The processing times are assumed to be exponentially distributed. The mean processing time at workstation W_j is $(1/\mu_j)$. For simplification, the workstations are assumed to be reliable. As a result, there are $N+1$ stochastic processes involved in the system.

EXACT MODEL

Considering the Poisson demand arrival process for finished product FP, $\{N_D(t), t \geqslant 0\}$, and the satisfied demand during period k, $D_s(k)$ $(0 \leqslant D_s(k) \leqslant I_{N+1}^{\text{in}}(k)$, because of no backorders), the probability distribution is:

$$P[D_s(k) = d_s^0 \mid I_{N+1}^{\text{in}}(\text{k})] = \begin{cases} \dfrac{(\lambda T)^{d_s{}^0}}{d_s^0\,!}\,\mathrm{e}^{-\lambda T} & 0 \leqslant d_s^0 < I_{N+1}^{\text{in}}(k) \\ 1 - \displaystyle\sum_{l=0}^{d_s{}^0-1} \frac{(\lambda T)^l}{l!}\,\mathrm{e}^{-\lambda T} & d_s^0 = I_{N+1}^{\text{in}}(k). \end{cases} \qquad (1)$$

Considering the production/inventory control system, the production orders to be released for period k are determined at the beginning of period k. After the periodic transfer of WIP inventory at the end of the period $k-1$, a production order (the number of production kanbans collected within the period $k-1$) is released at workstation W_j for producing component C_j in period k. The sum of all undelivered production orders (remaining production kanbans to be processed) at workstation W_j at the beginning of period k becomes $K_j - I_{j+1}^{\text{in}}(k)$. This targeted amount of production could be achieved if there is a sufficient amount of component C_{j-1} at workstation W_j. That is, if $K_j - I_{j+1}^{\text{in}}(k) \leqslant I_j^{\text{in}}(k) + W_j^{\text{on}}(k)$ where $W_j^{\text{on}}(k)$ is one if workstation W_j is busy processing component C_{j-1} at the beginning of period k, and zero if the workstation W_j is idle at the beginning of period k. The target production is then adjusted according to the availability of component C_{j-1} at the beginning of period k as:

$$O_j(k) = \min\{K_j - I_{j+1}^{\text{in}}(k),\ I_j^{\text{in}}(k) + W_j^{\text{on}}(k)\}, \qquad 1 \leqslant j \leqslant N. \tag{2}$$

On the other hand, the actual amount of production during period k at workstation W_j is referred to as $P_j(k)$ $(0 \leqslant P_j(k) \leqslant O_j(k))$. Considering the exponential production process of component C_j at workstation W_j, the probability distribution of producing $P_j(k)$ units of component C_j during period k is:

$$P[P_j(k) = p_j^0 \mid O_j(k)] = \begin{cases} \dfrac{(\mu_j T)^{p_j^0}}{p_j^0\,!}\, e^{-\mu_j T} & 0 \leqslant p_j^0 < O_j(k) \\[2ex] 1 - \displaystyle\sum_{l=0}^{p_j^0 - 1} \dfrac{(\mu_j T)^l}{l!}\, e^{-\mu_j T} & p_j^0 = O_j(k). \end{cases} \tag{3}$$

The state of workstation W_j at the beginning of period k can be described by a pair of system parameters, $(I_j^{\text{in}}(k),\ W_j^{\text{on}}(k))$, where $0 \leqslant I_j^{\text{in}}(k) \leqslant K_{j-1}$, $W_j^{\text{on}}(k) \in \{0, 1\}$ and moreover, $I_j^{\text{in}}(k) + W_j^{\text{on}}(k) \leqslant K_{j-1}$. Then, the state of the whole system at the beginning of period k can be satisfactorily described by $2N$ parameters:

$$\mathcal{S}(k) = [W_1^{\text{on}}(k), I_2^{\text{in}}(k), W_2^{\text{on}}(k), I_3^{\text{in}}(k), W_3^{\text{on}}(k), \ldots, I_N^{\text{in}}(k), W_N^{\text{on}}(k), I_{N+1}^{\text{in}}(k)] \tag{4}$$

The one-step transition equations, determining the system state $\mathcal{S}(k)$ are as follows.

Workstation status

$$W_1^{\text{on}}(k) = \begin{cases} 1 & \text{if } I_2^{\text{in}}(k-1) < K_1 \\ 0 & \text{if } I_2^{\text{in}}(k-1) = K_1 \end{cases} \tag{5}$$

$$W_j^{\text{on}}(k) = \begin{cases} 1 & \text{if } \begin{array}{c} W_j^{\text{on}}(k-1) = 1 \text{ and } P_j(k-1) = 0 \\ \text{or} \\ W_j^{\text{on}}(k-1) = 0 \text{ and } O_j(k-1) > 0 \text{ and } P_j(k-1) = 0 \\ \text{or} \\ 0 \leqslant P_j(k-1) < O_j(k-1) \end{array} \\ 0 & \text{if } \begin{array}{c} W_j^{\text{on}}(k-1) = 0 \text{ and } O_j(k-1) = 0 \\ \text{or} \\ P_j(k-1) = O_j(k-1) \end{array} \end{cases} \quad 2 \leqslant j \leqslant N. \tag{6}$$

Inventory status

$$I_j^{\text{in}}(k) = I_j^{\text{in}}(k-1) + W_j^{\text{on}}(k-1) + P_{j-1}(k-1) - (P_j(k-1) + W_j^{\text{on}}(k)), \qquad 2 \leqslant j \leqslant N, \tag{7}$$

$$I_{N+1}^{\text{in}}(k) = I_{N+1}^{\text{in}}(k-1) + P_N(k-1) - D_s(k-1). \tag{8}$$

All alternative transitions from $\mathcal{S}(k-1)$ to $\mathcal{S}(k)$ can be found by enumerating all possible values of $N+1$ stochastic processes. The entries of the resulting one-step transition probability matrix M are as follows:

$$m[\mathcal{S}(k-1), \mathcal{S}(k)] = \sum_{\boldsymbol{P}(k-1) \in \mathcal{R}} \xi(\boldsymbol{P}(k-1)) P[D_s(k-1) = d_s^0 \mid I_{N+1}^{\text{in}}(k-1)] \prod_{j=1}^{N} P[P_j(k-1) = p_j^0 \mid O_j(k-1)] \tag{9}$$

where

$$\mathscr{R} = \{\boldsymbol{P}(k-1) = [P_1(k-1), \ldots, P_N(k-1), D_s(k-1)]: \\ 0 \leqslant P_j(k-1) \leqslant O_j(k-1),\ 1 \leqslant j \leqslant N,\ 0 \leqslant D_s(k-1) \leqslant I_{N+1}^{\text{in}}(k-1)\} \quad (10)$$

$$\xi(\boldsymbol{P}(k-1)) = \begin{cases} 1 & \text{if } \boldsymbol{P}(k-1) \text{ causes a transition from } \mathscr{S}(k-1) \text{ to } \mathscr{S}(k) \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

In this formulation, the limiting distribution of the states of the system $\boldsymbol{\pi}$ could be found (if it exists) by solving the stationary equations of the Markov chain under consideration with the following boundary condition imposed:

$$\boldsymbol{\pi M} = \boldsymbol{\pi} \qquad \text{and} \qquad \boldsymbol{\pi e}^T = 1 \quad (12)$$

where $\boldsymbol{e}$ is a row vector with all elements equal to one, and $\boldsymbol{\pi}$ is the unique solution of the above equations. A discussion on the variety of methods to compute the stationary probabilities of large Markov chains can be found in Philippe *et al.*[27] and Baruh and Altiok[28].

Some of the key performance measures

Average inventory levels. The above formulation results in N buffer stocks under consideration Q_j^{in}, $2 \leqslant j \leqslant N+1$. The mean inventory level at Q_j^{in} during the period is:

$$\text{MI}_j = \begin{cases} \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} P[I_j^{\text{in}} = i_j^0,\ W_j^{\text{on}} = w_j^0,\ I_{j+1}^{\text{in}} = i_{j+1}^0] \\ \quad \times \left[(i_j^0 - O_j) + \displaystyle\sum_{p_j^0=1}^{O_j} (O_j + 1 - p_j^0) \frac{\text{MTTP}_j(P_j^0) - \text{MTTP}_j(p_j^0 - 1)}{T} \right] & j = 2, \ldots, N. \\ \displaystyle\sum_{i_j^0=0}^{K_{j-1}} P[I_j^{\text{in}} = i_j^0] \left[\sum_{d_s^0=1}^{i_j^0} (i_j^0 + 1 - d_s^0) \frac{\text{MTTD}_s(d_s^0) - \text{MTTD}_s(d_s^0 - 1)}{T} \right] & j = N+1 \end{cases} \quad (13)$$

where

$$\text{MTTP}_j(p_j^0) = \begin{cases} 0 & p_j^0 = 0 \\ \displaystyle\int_0^T t \frac{\mu_j^{(p_j^0)} t^{(p_j^0 - 1)}}{(p_j^0 - 1)!} \mathrm{e}^{-\mu_j t}\, \mathrm{d}t + \int_T^\infty T \frac{\mu_j^{(p_j^0)} t^{(p_j^0 - 1)}}{(p_j^0 - 1)!} \mathrm{e}^{-\mu_j t}\, \mathrm{d}t & 1 \leqslant p_j^0 \leqslant O_j \end{cases} \quad (14)$$

$$j = 2, \ldots, N.$$

$$\text{MTTD}_s(d_s^0) = \begin{cases} 0 & d_s^0 = 0 \\ \displaystyle\int_0^T t \frac{\lambda^{(d_s^0)} t^{(d_s^0 - 1)}}{(d_s^0 - 1)!} \mathrm{e}^{-\lambda t}\, \mathrm{d}t + \int_T^\infty T \frac{\lambda^{(d_s^0)} t^{(d_s^0 - 1)}}{(d_s^0 - 1)!} \mathrm{e}^{-\lambda t}\, \mathrm{d}t & 1 \leqslant d_s^0 \leqslant i_j^0. \end{cases} \quad (15)$$

Average throughput rate. Considering the long-term behaviour of the system, the throughput rates of the workstations are equal to each other because of the conservation of material flow in the system. The mean throughput rate of workstation W_j is denoted by MTR_j and defined as the expected number of component C_j items produced per unit time. The mean throughput rate of the system is:

$$\text{MTR} = \text{MTR}_N = \text{MTR}_{N-1} = \ldots = \text{MTR}_2 = \text{MTR}_1 \quad (16)$$

where

$$\text{MTR}_j = \begin{cases} \displaystyle\sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} \sum_{p_j^0=0}^{O_j} \left(\frac{p_j^0}{T}\right) P[W_j^{\text{on}} = w_j^0,\ I_{j+1}^{\text{in}} = i_{j+1}^0] P[P_j = p_j^0 \mid O_j] & j = 1 \\ \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_j^0=0}^{1} \sum_{i_{j+1}^0=0}^{K_j} \sum_{p_j^0=0}^{O_j} \left(\frac{p_j^0}{T}\right) P[I_j^{\text{in}} = i_j^0,\ W_j^{\text{on}} = w_j^0,\ I_{j+1}^{\text{in}} = i_{j+1}^0] P[P_j = p_j^0 \mid O_j] & 2 \leqslant j \leqslant N. \end{cases} \quad (17)$$

This is an important performance measure since the other performance measures (workstation utilization and service level) could be computed from the mean throughput rate of the system.

Average workstation utilization. Although the long-term mean throughput rates of the workstations are equal, the utilization of workstations MU_j could be different because the production rates of workstations may differ. The mean utilization of workstation W_j is:

$$\mathrm{MU}_j = \frac{\mathrm{MTR}_j}{\mu_j} = \frac{\mathrm{MTR}}{\mu_j}. \tag{18}$$

Average service level. The formulation of this system considers a loss system in which the demand for finished product FP, arriving at times when Q_{N+1}^{in} is empty, is lost. The mean service level of the system is:

$$\mathrm{MSL} = \frac{\mathrm{MTR}}{\lambda}. \tag{19}$$

APPROXIMATE DECOMPOSITION

The approximation method decomposes the production system into several individual sub-systems: starting with the last stage, each of the stages is approximated by a single-stage model with appropriately revised material supply, production and demand arrival functions. This decomposition procedure is repeated several times in order to approximate adequately the performance measures of the production system as a whole. The goal is to approximate the whole system given in Figure 2 by a sequence of isolated single-stage pull production sub-systems, $\mathscr{X}_j$, $1 \leqslant j \leqslant N$ (see Figure 3). The first and the last sub-systems are atypical since, in the first sub-system, the raw material input is assumed to be infinite and in the last stage the Poisson demand arrivals for the finished product are external to the system.

The state of sub-system $\mathscr{X}_j$ at the beginning of period k can be described by a pair of system parameters, $(W_j^{\mathrm{on}}(k), I_{j+1}^{\mathrm{in}}(k))$, where $0 \leqslant I_{j+1}^{\mathrm{in}}(k) \leqslant K_j$, $W_j^{\mathrm{on}}(k) \in \{0, 1\}$. In our formulation, the state of the isolated single-stage periodic pull production sub-system at the beginning of period k is simply denoted by:

$$\mathscr{S}_{\mathscr{X}_j}(k) = [W_j^{\mathrm{on}}(k), I_{j+1}^{\mathrm{in}}(k)]. \tag{20}$$

The one-step transition equations, determining the state of sub-systems, are the same as equations (5)–(8). All alternative transitions from $\mathscr{S}_{\mathscr{X}_j}(k-1)$ to $\mathscr{S}_{\mathscr{X}_j}(k)$ can be found by enumerating all possible realizations of related random variables; $I_j^{\mathrm{in}}(k-1)$, $P_j(k-1)$, $W_{j+1}^{\mathrm{on}}(k-1)$, $I_{j+2}^{\mathrm{in}}(k-1)$

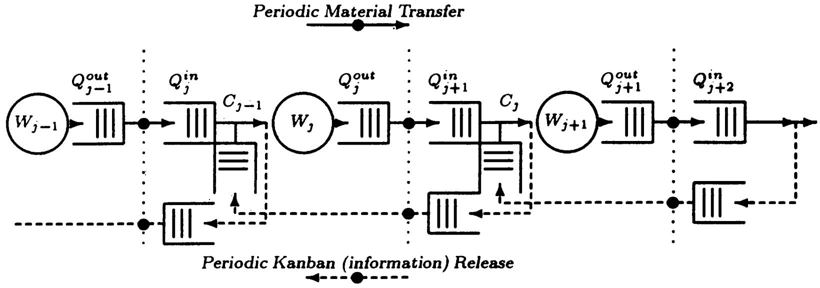

FIG. 3. *An isolated single-stage pull production subsystem* $\mathscr{X}_j$

and $P_{j+1}(k-1)$. The entries of the resulting one-step transition probability matrix $M_{\mathscr{X}_j}$ could be approximately computed as follows:

$$m_{\mathscr{X}_j}[\mathscr{S}_{\mathscr{X}_j}(k-1), \mathscr{S}_{\mathscr{X}_j}(k)] \approx$$

$$\begin{cases} \displaystyle\sum_{w_{j+1}^0=0}^{1} \sum_{i_{j+2}^0=0}^{K_{j+1}} \sum_{p_j^0=0}^{O_j} \sum_{p_{j+1}^0=0}^{O_{j+1}} P[W_{j+1}^{\text{on}} = w_{j+1}^0, I_{j+2}^{\text{in}} = i_{j+2}^0] P[P_j = p_j^0 \mid O_j] \\ \qquad\qquad \times P[P_{j+1} = p_{j+1}^0 \mid O_{j+1}] \xi(\bullet) \qquad j = 1 \\ \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{w_{j+1}^0=0}^{1} \sum_{i_{j+2}^0=0}^{K_{j+1}} \sum_{p_j^0=0}^{O_j} \sum_{p_{j+1}^0=0}^{O_{j+1}} P[I_j^{\text{in}} = i_j^0] P[W_{j+1}^{\text{on}} = w_{j+1}^0, I_{j+2}^{\text{in}} = i_{j+2}^0] P[P_j = p_j^0 \mid O_j] \\ \qquad\qquad \times P[P_{j+1} = p_{j+1}^0 \mid O_{j+1}] \xi(\bullet) \qquad 1 < j < N \\ \displaystyle\sum_{i_j^0=0}^{K_{j-1}} \sum_{p_j^0=0}^{O_j} \sum_{d_s^0=0}^{I_{j+1}^{\text{in}}} P[I_j^{\text{in}} = i_j^0] P[P_j = p_j^0 \mid O_j] P[D_s = d_s^0 \mid I_{j+1}^{\text{in}}] \xi(\bullet) \qquad j = N \end{cases} \tag{21}$$

$$\xi(\bullet) = \begin{cases} 1 & \text{if the realizations of the related random variables cause a transition from } \mathscr{S}_{\mathscr{X}_j}(k-1) \text{ to } \mathscr{S}_{\mathscr{X}_j}(k) \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

In this formulation, the limiting distribution of the states of the sub-system $\pi_{\mathscr{X}_j}$ could be found (if it exists) in the same manner. The aim of the proposed decomposition approach is to represent the whole production system by a sequence of isolated single-stage periodic pull production sub-systems, where the streams of raw material and demand for component C_j to be produced at sub-system $\mathscr{X}_j$ are provided by sub-systems $\mathscr{X}_{j-1}$ and $\mathscr{X}_{j+1}$, respectively (see Figure 4). The parameters of these isolated sub-systems must be coordinated in such a way that the performance characteristics of the resulting sequence are as close as possible to those of the production system as a whole.

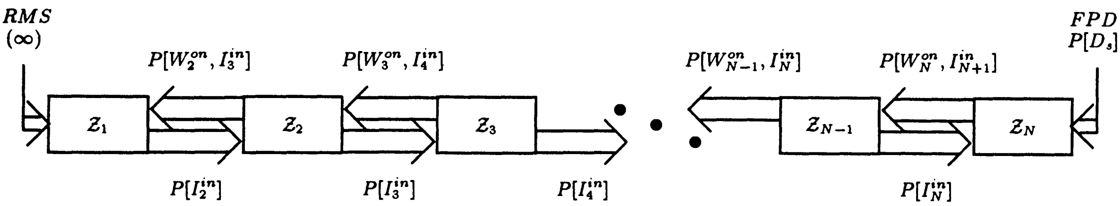

FIG. 4. *Model of the production system constituted from models of isolated single-stage sub-systems.*

While decomposing the whole production system, we start with the last sub-system, $\mathscr{X}_N$, and work our way backwards until we reach the first sub-system, by considering an infinite supply of raw material at all input buffer stocks, Q_j^{in}, in order to initialize the steady-state probabilities of states of all decomposed sub-systems. In this backward initialization pass, the starvation of all sub-systems is ignored and only blocking is considered. Then, two consecutive passes, backward and forward passes, are executed iteratively until a satisfactory level of approximation in evaluating the performance measures of the whole production system is obtained. The level of approximation is determined by the deviation between throughput rates of the subsystems at consecutive iterations. During these iterations, both starvation and blocking of sub-systems are considered. More precisely, the steps summarizing the decomposition approach are as follows.

Step 0. Initialization

Set iteration index, $l \leftarrow 0$.
Set $P^{(l)}[I_j^{\text{in}} = K_{j-1}] = 1$, for $j = 2, \ldots, N+1$.
Set sub-system (stage) index, $j \leftarrow N$.
Set the level of approximation ($\varepsilon \leftarrow 10^{-8}$).
Backward loop. For $j := N$ down to 1:
compute $M_{\mathscr{X}_j}^{(l)}$ and $\pi_{\mathscr{X}_j}^{(l)}$.

Step 1. Iterations

Set $l \leftarrow l + 1$,
Backward loop. For $j := N$ down to 1:
compute $M^{(l)}_{\mathscr{X}_j}$, $\pi^{(l)}_{\mathscr{X}_j}$ and $\mathrm{MTR}^{(l_b)}_{\mathscr{X}_j}$.
Forward loop. For $j := 2$ to N:
compute $M^{(l)}_{\mathscr{X}_j}$, $\pi^{(l)}_{\mathscr{X}_j}$ and $\mathrm{MTR}^{(l_f)}_{\mathscr{X}_j}$.

Step 2. Stopping criteria

If $\max_{2 \leqslant j \leqslant N} |\mathrm{MTR}^{(l_b)}_{\mathscr{X}_j} - \mathrm{MTR}^{(l_f)}_{\mathscr{X}_j}| < \varepsilon$ then
compute the performance measures of the system and stop; otherwise go to Step 1.

We do not have a proof of convergence. However, in the many examples we have examined the method has always converged within a reasonable number of iterations (low 10s), only moderately dependent on the number of stages. As a result, the computational complexity of our approach grows relatively moderately (but more than linearly) with the number of stages in the system.

The key performance measures

Average inventory levels. According to the above formulation of the sub-systems, there are N buffer stocks under consideration, Q^{in}_j, $2 \leqslant j \leqslant N + 1$. The mean inventory level at Q^{in}_j during the period is:

$$\mathrm{AMI}_j \approx \begin{cases} \displaystyle\sum_{i_j{}^0=0}^{K_{j-1}} \sum_{w_j{}^0=0}^{1} \sum_{i_{j+1}{}^0=0}^{K_j} P[I^{\mathrm{in}}_j = i^0_j]P[W^{\mathrm{on}}_j = w^0_j, I^{\mathrm{in}}_{j+1} = i^0_{j+1}] \\ \displaystyle\quad \times \left[(i^0_j - O_j) + \sum_{p_j{}^0=1}^{O_j} (O_j + 1 - p^0_j) \frac{\mathrm{MTTP}_j(p^0_j) - \mathrm{MTTP}_i(p^0_j - 1)}{T}\right] & j = 2, \ldots, N \\ \displaystyle\sum_{i_j{}^0=0}^{K_{j-1}} P[I^{\mathrm{in}}_j = i^0_j]\left[\sum_{d_s{}^0=1}^{i_j{}^0} (i^0_j + 1 - d^0_s) \frac{\mathrm{MTTD}_s(d^0_s) - \mathrm{MTTD}_s(d^0_s - 1)}{T}\right] & j = N + 1. \end{cases} \tag{23}$$

Average throughput rate. The mean throughput rate of sub-system $\mathscr{X}_j$ is denoted by $\mathrm{MTR}_{\mathscr{X}_j}$ and is defined as the expected number of component C_j items produced per unit time. The mean throughput rate of the whole system is:

$$\mathrm{AMTR} = \mathrm{MTR}_{\mathscr{X}_N} \approx \mathrm{MTR}_{\mathscr{X}_{N-1}} \approx \ldots \approx \mathrm{MTR}_{\mathscr{X}_2} \approx \mathrm{MTR}_{\mathscr{X}_1} \tag{24}$$

where

$$\mathrm{MTR}_{\mathscr{X}_j} \approx \begin{cases} \displaystyle\sum_{w_j{}^0=0}^{1} \sum_{i_{j+1}{}^0=0}^{K_j} \sum_{p_j{}^0=0}^{O_j} \left(\frac{p^0_j}{T}\right) P[W^{\mathrm{on}}_j = w^0_j, I^{\mathrm{in}}_{j+1} = i^0_{j+1}]P[P_j = p^0_j \mid O_j] & j = 1 \\ \displaystyle\sum_{i_j{}^0=0}^{K_{j-1}} \sum_{w_j{}^0=0}^{1} \sum_{i_{j+1}{}^0=0}^{K_j} \sum_{p_j{}^0=0}^{O_j} \left(\frac{p^0_j}{T}\right) P[I^{\mathrm{in}}_j = i^0_j]P[W^{\mathrm{on}}_j = w^0_j, I^{\mathrm{in}}_{j+1} = i^0_{j+1}]P[P_j = p^0_j \mid O_j] \\ & 2 \leqslant j \leqslant N. \end{cases} \tag{25}$$

Average utilization. Although the long-term mean throughput rates of the sub-systems are equal, the utilization of sub-systems $\mathrm{MU}_{\mathscr{X}_j}$ could be different because the production rates of the sub-systems may differ. The mean utilization of sub-system $\mathscr{X}_j$ is:

$$\mathrm{AMU}_j = \mathrm{MU}_{\mathscr{X}_j} = \frac{\mathrm{MTR}_{\mathscr{X}_j}}{\mu_j} \approx \frac{\mathrm{AMTR}}{\mu_j}. \tag{26}$$

Average service level. This is the ratio of finished product demand satisfied from stock to the total demand arrived within a period. The mean service level of the whole system is:

$$\text{AMSL} \approx \frac{\text{AMTR}}{\lambda}. \tag{27}$$

NUMERICAL EXPERIMENTATION

An experiment is designed in order to investigate the general behaviour and the accuracy level of the single-stage approximate decomposition technique. A three-stage system is selected, because it is the smallest system that requires a significant amount of reduction in computation while solving the exact model. In the context of this experiment, 320 different three-stage systems were evaluated using both the exact and the approximate models. The range of system parameters is as follows.

- Mean arrival rate of finished product demand; $\lambda = (0.1, 0.5, 1.0, 2.0, 10.0)$.
- Number of kanbans at each stage; $K = (1, 2, 3, 4)$.
- Mean production rate at each stage; $\mu = \lambda/\rho$,
  where ρ is the traffic intensity or the demand load, $\rho = (0.45, 0.60, 0.75, 0.90)$.
- Length of the transfer/review period; $T = (1, 2, 3, 4)$.

These pull systems consider a single product with a Poisson demand that arrives at the third (last) stage of the system with a mean rate of λ. The demand arrivals during the times the finished product buffer is empty are lost (backordering is not allowed). At each stage of the system, the processing times are exponential with the same mean $1/\mu$ and the number of kanbans allocated are equal to K. The status of the system is reviewed periodically with a period length of T. The production and material withdrawal orders are released at the beginning of the periods. It is assumed that the raw material supply for the first stage is infinite and the material handling times between stages are zero.

The mean throughput rate is selected as a primary measure of performance for this experiment. All comparisons are based on this primary measure. Numerical experience suggests that when the mean throughput rates of the workstations converge to a unique solution during the iteration process, it agrees closely with the exact model. The percentage absolute error between the exact and the approximate mean throughput rates is computed as follows:

$$\%\text{ absolute error} = 100\left|\frac{\text{AMTR} - \text{MTR}}{\text{MTR}}\right| \tag{28}$$

See Table 1 for the percentage absolute errors obtained from the results of the experiment and for the effect of system parameters on the accuracy of the approximate decomposition technique.

The effect of the number of kanbans at each stage is very important. When there is only one kanban at each stage, the average of percentage absolute errors is greater than 20. This is because the starvation and blocking probabilities are very significant and an estimation error in these probabilities causes a large error in the computation of performance measures of the whole system. For the case of an increasing number of kanbans at each stage the average of the percentage absolute errors, although fluctuating within an acceptable range, is decreasing in the limit. Very low and very high demand arrival rates have a relatively modest effect on the accuracy level for the number of kanbans exceeding one. The average of the percentage absolute errors seems to be insensitive to the variation in the traffic intensity. On the other hand, the errors slightly increase with an increase in transfer/review period length, and note that the average of the percentage absolute errors is comparatively small for the number of kanbans exceeding one.

The overall average of the percentage absolute errors between AMTR and MTR is less than 10. Generally speaking, it is accepted that the error level of an approximate decomposition technique should not exceed 3%. Note that, the average of the percentage absolute errors for the systems with $K \geqslant 2$ and $0.5 \leqslant \lambda \leqslant 2.0$ is less than 2.90 (see the summary report in Table 1). As a result, the

TABLE 1. *The average absolute percentage errors between the exact and the approximate mean throughput rates*

With respect to λ	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 2.0$	$\lambda = 10.0$
Overall	3.3511	6.3558	9.0488	10.5789	12.8461
$2 \leqslant K \leqslant 4$	3.1546	**2.1602**	**2.9847**	**3.5394**	6.0173
With respect to K	$K = 1$	$K = 2$	$K = 3$	$K = 4$	
Overall	23.0313	2.8095	6.6299	1.2739	
$0.5 \leqslant \lambda \leqslant 2.0$	25.9610	**2.0096**	**5.3654**	**1.3085**	
With respect to ρ	$\rho = 0.45$	$\rho = 0.60$	$\rho = 0.75$	$\rho = 0.90$	
Overall	8.5393	8.6048	8.1220	8.4785	
$0.5 \leqslant \lambda \leqslant 2.0$	8.9500	8.7396	8.5569	8.3981	
$2 \leqslant K \leqslant 4$	3.4607	3.7103	3.2490	3.8643	
$0.5 \leqslant \lambda \leqslant 2.0$, $2 \leqslant K \leqslant 4$	**2.8039**	**2.8128**	**2.9056**	**3.0558**	
With respect to T	$T = 1$	$T = 2$	$T = 3$	$T = 4$	
Overall	6.7071	7.9913	8.6080	10.4381	
$0.5 \leqslant \lambda \leqslant 2.0$	5.9610	8.5240	9.7510	10.4085	
$2 \leqslant K \leqslant 4$	2.7920	3.0637	3.2482	5.1805	
$0.5 \leqslant \lambda \leqslant 2.0$, $2 \leqslant K \leqslant 4$	**2.1603**	**2.7392**	**3.1931**	**3.4856**	
Summary	Overall	$0.5 \leqslant \lambda \leqslant 2.0$	$2 \leqslant K \leqslant 4$	$0.5 \leqslant \lambda \leqslant 2.0$, $2 \leqslant K \leqslant 4$	
report	8.4361	8.6612	3.5712	**2.8948**	

proposed approximate decomposition technique could be used for the evaluation of NTQ equivalent periodic pull production systems having more than one kanban at each stage and a demand arrival rate not in extreme values relative to other system parameters such as K, ρ and T.

CONCLUSIONS

A variety of production systems appearing in the literature has been investigated. There have been a few attempts to develop analytical models for the performance evaluation of kanban-controlled stochastic pull production systems. Most of the existing models address tandem-queue equivalent systems. There are a number of NTQ equivalent pull production systems to be considered in a research study. A periodic review–instantaneous order/periodic transfer system is selected as the basic system to start the research on modelling and analysis of NTQ equivalent pull production systems. This basic system is formulated as a discrete time Markov process. Because of the dimensionality problem inherited in the exact solution technique, it could be exactly evaluated up to three stages in tandem.

An approximate decomposition approach is proposed to handle larger periodic pull production systems that are analytically intractable. The proposed approach generates results that are quite close to the exact solution of the three stage systems. In order to improve the overall accuracy level of the approximation, a further study could be the development and analysis of a two-node decomposition technique. This type of approximation might lower the average errors on performance measures since one of the approximated probabilities utilized in the decomposition technique could be exactly evaluated. On the other hand, the computation requirements of a two-node decomposition increase both in terms of memory and time.

Note that the proposed approximation technique is demonstrated on our basic periodic pull production system, in which the arrival and the production processes are both Markovian. Other research could be based on the interaction of the variation coming from the stochastic processes in the system and the accuracy level of the approximation technique. In this way, several discrete distributions with different levels of variation could be utilized in the formulation. The extensions of the model to cover back-orders and unreliable machines are straightforward. In terms of the configuration of the network, the approximation could be extended to cover periodic pull production systems in the flow shop configuration by formulating the split and merge sub-systems.

REFERENCES

1. D. Y. Golhar and C. L. Stamm (1991) The just-in-time philosopy: a literature review. *Int. J. Prod. Res.* **29**, 657–676.
2. Y. Sugimori, K. Kusunoki, F. Cho and S. Uchikawa (1977) Toyota production system and kanban system: materialization of just-in-time and respect for human system. *Int. J. Prod. Res.* **15**, 553–564.
3. O. Kimura and H. Terada (1981) Design and analysis of pull systems: a method of multi-stage production control. *Int. J. Prod. Res.* **19**, 241–253.
4. B. J. Berkley (1992) A review of the kanban production control research literature. *Prod. Opns Mgmt* **1**, 393–411.
5. B. J. Berkley (1991) Tandem queues and kanban-controlled lines. *Int. J. Prod. Res.* **29**, 2057–2081.
6. A. Brandwajn (1985) Equivalence and decomposition in queuing systems—a unified approach. *Perf. Eval.* **5**, 175–186.
7. R. O. Onvural and H. G. Perros (1986) On equivalencies of blocking mechanisms in queuing networks with blocking. *Opns Res. Lett.* **5**, 292–297.
8. F. S. Hillier and R. W. Boling (1967) Finite queues in series with exponential or erlang service times: a numerical approach. *Opns Res.* **15**, 286–303.
9. G. C. Hunt (1956) Sequential arrays of waiting lines. *Opns Res.* **4**, 674–683.
10. E. J. Muth (1984) Stochastic processes and their network representations associated with a production line queuing model. *Eur. J. Opl Res.* **15**, 63–83.
11. T. Altiok (1989) Approximate analysis of queues in series with phase-type service times and blocking. *Opns Res.* **37**, 601–610.
12. M. B. M. De Koster (1988) Approximate analysis of production systems. *Eur. J. Opl Res.* **37**, 214–226.
13. S. B. Gershwin (1987) An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Opns Res.* **35**, 291–305.
14. T. Altiok (1985) On the phase-type approximation of general distributions. *IIE Trans.* **17**, 110–116.
15. T. Altiok and J. S. Stidham (1983) The allocation of interstage buffer capacities in production lines. *IIE Trans.* **15**, 292–299.
16. H. Wang and H. Wang (1990) Determining the number of kanbans: a step toward non-stock production. *Int. J. Prod. Res.* **28**, 2101–2115.
17. S. Meral (1993) A design methodology for just-in-time production lines. PhD Thesis, Department of Industrial Engineering, Middle East Technical University, Ankara, Turkey.
18. T. Kim (1985) Just-in-time manufacturing system: a periodic pull system. *Int. J. Prod. Res.* **23**, 553–562.
19. B. R. Sarker and G. R. Parija (1994) An optimal batch size for a production system operating under a fixed-quantity, periodic delivery policy. *J. Opl Res. Soc.* **45**, 891–900.
20. J. L. Deleersnyder, T. J. Hodgson, H. Müller(-Malek) and P. J. O'Grady (1989) Kanban type controlled pull systems: an analytic approach. *Mgmt Sci.* **35**, 1079–1091.
21. B. J. Berkley (1992) A decomposition approximation for periodic kanban-controlled flow shops. *Decis. Sci.* **23**, 291–311.
22. K. C. So and S. C. Pinault (1988) Allocating buffer storages in a pull system. *Int. J. Prod. Res.* **26**, 1959–1980.
23. D. Mitra and I. Mitrani (1990) Analysis of a kanban discipline for cell coordination in production lines, I. *Mgmt Sci.* **36**, 1548–1566.
24. D. Mitra and I. Mitrani (1991) Analysis of a kanban discipline for cell coordination in production lines, II: stochastic demands. *Opns Res.* **39**, 807–823.
25. J. A. Buzacott (1989) Queuing models of kanban and MRP controlled production systems. *Eng. Costs and Prod. Economics* **17**, 3–20.
26. R. D. Badinelli (1992) A model for continuous-review pull policies in serial inventory systems. *Opns Res.* **40**, 142–156.
27. B. Philippe, Y. Saad and W. J. Stewart (1992) Numerical methods in Markov chain modeling. *Opns Res.* **40**, 1156–1179.
28. H. Baruh and T. Altiok (1991) Analytical perturbations in Markov chains. *Eur. J. Opl Res.* **51**, 210–222.

Received September 1994; accepted May 1995 after one revision