ELSEVIER

European Journal of Operational Research 119 (1999) 479–494

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/orms

The general behavior of pull production systems: The allocation problems

Nureddin Kırkavak [a,*], Cemal Dinçer [b]

[a] *Department of Industrial Engineering, Eastern Mediterranean University, Gazi Mağusa – TRNC, Mersin 10, Turkey*
[b] *Department of Industrial Engineering, Bilkent University, Ankara 06533, Turkey*

Abstract

The design of tandem production systems has been well studied in the literature with the primary focus being on how to improve their efficiency. Considering the large costs associated, a slight improvement in efficiency can lead to very significant savings over its life. Division of work and allocation of buffer capacities between workstations are two critical design problems that have attracted the attention of many researchers. In this study, first an understanding into how the system works is to be provided. Except for the integration of two allocation problems, the basic model utilized here is essentially the same as the previous studies. Theoretical results that characterize the dynamics of these systems may also provide some heuristic support in the analysis of large-scale pull production systems.

Keywords: Pull production; Production/inventory systems; Performance evaluation; Resource allocation; Throughput maximization; Markov processes; Simulation

1. Introduction

In the last decade, there have been numerous attempts for modelling production systems as queuing systems for the purpose of understanding their behavior. So far, the models in the literature usually involved single-product systems with single or multiple stages for tractability purposes. Cases with multiple products, although closer to reality, proved to be quite difficult to tackle analytically. A production system is usually viewed as an arrangement of production stages in a particular configuration, where each stage consists of a single workstation or several workstations in parallel. These workstations may consist of workers, machines and work-in-process materials.

Performance evaluation in general is concerned with finding out how well the system is functioning provided that certain policies and parameters are set. Typical performance measures for the evaluation of production systems are throughput, average inventory levels, utilizations and customer service levels among others. In obtaining these measures, when analytical techniques become insufficient often numerical techniques such as simulation or approximations could be used.

* Corresponding author. Tel.: +90-392-366-6588; fax: +90-392-365-4029; e-mail: kavak@gantt.ie.emu.edu.tr

0377-2217/99/$ – see front matter
PII: S0377-2217(99)00148-4

An important part of production research literature appeared in the area of production lines. During the last 30 years, performance evaluation models have been developed for many different types of production lines using exact and approximate approaches. The design of tandem production systems has been well studied in the production research literature with the primary focus being on how to improve their efficiency. Considering the large costs associated with these systems, a slight improvement in efficiency can lead to very significant savings over the life of the production system. Division of work among the workstations and allocation of buffer storage capacity between workstations are two critical design factors that have attracted the attention of many researchers and system designers. For a survey of the research in this area, see Ref. [24].

In this study, we analyze the performance of periodic pull production systems for theoretical results that characterize the dynamics of these systems. First, the previous results on the allocation problems will be summarized in Section 2. Then, the system we considered will be described in Section 3 together with an understanding into how these systems work. In Section 4, the two allocation problems and their integration for the objective of throughput maximization will be introduced. Then, the empirical results we obtained through a series of numerical experiments will be discussed in Section 5. Finally, in Section 6 an allocation methodology will be proposed in order to provide some heuristic support for the analysis of large-scale pull production systems.

2. Review of previous results

One significant aspect of production line design is the so-called *line balancing problem*, i.e. allocating the total work content as evenly as possible to workstations and maximizing the utilization through minimizing idle times as well. The solution of line balancing problem specifies a system configuration capable of producing a specified amount of finished product with minimum resource requirements. The operation times can be either deterministic or stochastic. However, line balancing techniques are based on the assumption of deterministic operation times. In practice, a perfect balance of workload may be impossible even with deterministic operation times, since, in most cases, equal allocation of total work content to workstations may be prevented by precedence and technological constraints, and continuous indivisibility of operations. In production systems with stochastic operation times, the balance of workload is attained through allocating the total work content evenly to the workstations based on the means of operation times. However, the balance of stochastic operation times may be impossible due to different variability of operation times at different workstations.

It is intuitively plausible that the variation in the operation times would decrease the mean production (throughput) rate of the system. This can happen in two ways: due to blocking and/or starvation. When there is considerable variability in the operation times at some respective workstations, a perfectly balanced production line may not be optimal. Previous work on *optimal allocation of workload* to production lines has found that, under certain assumptions, the mean throughput rate of a finite buffer production line is maximized by deliberately unbalancing the workload of the line in an appropriate way. In particular, the optimal allocation of work follows a 'bowl phenomenon' whereby the center workstations are given preferential treatment (less workload) over the other workstations towards the beginning and the ending workstations (see Refs. [10,11]). The analogous result of Stecke and Morin [30] is that the mean throughput rate of an infinite buffer production line is maximized by balancing the workload assigned to workstations. In other words, as buffer capacities increases, the degree of unbalance in the optimal workload decreases, until in the limit, a balanced allocation is optimal.

Hillier and Boling [11] report that the improvement in mean throughput rate due to unbalancing grows up to 1.37% for a six workstation serial production line. On the other hand, Magazine and Silver [18] developed an approximation that suggests the improvement from unbalancing is no larger than 1.65% for exponential operation times, regardless of the number of workstations in

the system. One of the main insights emanating from these studies is that balanced systems give acceptable performance and further improvements in mean throughput rate can be made by unbalancing. However, the gains obtained from unbalancing are relatively small – in the order of 1%. The works of El-Rayah [7] and So [28] indicate that the bowl phenomenon is robust. That is, as long as the balance of workload is changed in the direction indicated by the bowl phenomenon, the mean throughput rate function is almost flat near the maximum. On the other hand, if the production line is unbalanced in a different direction, the mean throughput rate decreases quite rapidly.

Muth and Alkaff [20] examine three stage serial production systems in a more general analytical setting in order to give the mean throughput rate as a function of several system parameters, subject to certain constraints. Rao [23] considers the generalization where the coefficient of variation of operation times are different for different workstations. The results found by Rao [23] indicate that unbalancing a serial production system can lead to substantial improvements in mean throughput rate when the variability of the stages differ from one to another. Optimum unbalancing could possibly be achieved by carrying out alternately the following two steps:

1. workload from interior stages should be transferred to the exterior ones (bowl phenomenon),
2. workload from more variable stages should be transferred to less variable ones (variability imbalance).

Step 1 is more important when the differences in the coefficient of variation of the stages are generally less than 0.5 while Step 2 predominates when they exceed 0.5. Then, Wolisz [34] shows that the idea of assigning less workload to more variable workstations is inappropriate for a coefficient of variation greater than one.

For lines longer than three stages and for non-exponential distributions, analytic approaches are quite limited, and some studies used simulation to study the workload allocation problem under more general conditions. Payne et al. [21] simulated production lines with different patterns of processing time variances and observed that a great deterioration in the performance occurs either when processing time variances are increased, or when buffer capacities are highly restricted. In a similar problem, Yamazaki et al. [36] investigated the optimal ordering of workstations that maximizes the mean throughput rate of the system. Based on some theoretical and extensive empirical results, they propose two rules for ordering workstations. The first rule recommends arranging the two worst workstations (apart from each other as far as possible) as the first and the last workstations. A worst workstation refers to the one either with the slowest production rate or with the most variable operation time. The second rule arranges the remaining workstations according to the bowl phenomenon.

All of the above studies have assumed that the production system has a serial structure. Baker et al. [4] investigated the behavior of assembly systems in which two or more parts are produced at component lines and put together at an assembly workstation at the end. Their basic finding is that the assembly workstation in a balanced system is intrinsically a bottleneck. Villeda et al. [33] studied an assembly system in which three serial lines (each one composed of three workstations) merge at one assembly workstation which is operating as a pull system. They consider normal processing times with several coefficients of variation and report that mean throughput rate is maximized by assigning decreasing amounts of work closer to assembly workstation at which the mean processing time is fixed.

The effect of bowl phenomenon has been extensively studied in conventional type push production systems, however, studies exploring its effects and validity on pull production systems are rare. The simulation studies made so far show conflicting results. In the simulation experiments performed by Meral [19], the bowl phenomenon is not confirmed for idealized just-in-time production systems. She found that balancing strategies are always superior to the unbalancing strategies based on bowl phenomenon. On the contrary, Villeda et al. [33] analyzed a just-in-time production system by investigating several unbalancing methods and they claim that the only method giving a consistent improvement in the mean throughput rate is the 'high-medium-low'

(decreasing) allocation. They also report that the mean throughput rate with unbalanced workstations are always superior to the perfectly balanced configurations. On the other hand, Sarker and Harris [25] claim that they observed the effect of bowl phenomenon on a just-in-time production system. Recently, Gstettner and Kuhn [9] have classified and studied different pull production systems and show that the buffer capacity (kanban) distribution has significant effect on the performance of the system. Also, they report that different pull policies show similar performance if the buffer capacity distribution is adapted according to the applied control mechanism.

Whatever the case, looking from a labor relations point of view, there may be difficulties in assigning significantly different workloads to different workstations. This raises the question as to whether there might be other ways of achieving this improvement in mean throughput rate by giving preferential treatment to the critical workstations without significantly unbalancing the workloads. One way of doing this is to provide such critical workstations with more buffer storage capacity than the other workstations. As surveyed by Sarker [24] various researchers have considered the general question of optimal allocation of buffer storage capacity in a variety of contexts. In the analogy to workload allocation problem there is a critical difference that the buffer allocation decision variables are discrete (integer) variables whereas the workload allocation decision variables are formulated as continuous variables in the previous studies.

Most of the research on buffer allocation has focused on analytical models of small systems simplified with restrictive assumptions [10,20]. For larger systems, analytical approximations or simulation models have been utilized [3,6]. Conway et al. [6] examined serial production systems via simulation. They find that buffers between workstations increase the production capacity of the system but the returns are reduced sharply with increasing inventory holding costs. They also note that the positioning as well as the capacity of the buffers are important. El-Rayah [8] utilized a computer simulation model to investigate the effect of unequal allocation of buffer capacity on the efficiency with an experiment limited to small production lines. He observed that the lines in which the center workstations are assigned larger buffer storage capacity than the ending workstations (inverted bowl phenomenon) are better (with respect to mean throughput rate) than the other unbalanced configurations. But, according to their experiment the inverted bowl configuration yielded more or less a similar mean throughput rate to that of a balanced line depending upon the total buffer storage capacity.

Hillier and So [12] studied the effect of the variability of processing times on the optimal allocation of buffer storage capacity between workstations. They conclude that either the center workstations or the workstations with high variability should be given more buffer capacity. Consequently, an inverted bowl phenomenon prevails regarding the optimal allocation of buffer storage capacity. In another study, Hillier and So [13] utilized an exact analytical model to conduct a detailed study of how the length of machine up and down times and interstage buffer storage capacity can effect the mean throughput rate of production lines with more than three stages. They developed a simple heuristic to estimate the amount of buffer storage capacity required to compensate for the decrease in mean throughput rate due to machine breakdowns. Sheskin [26] offers some guidelines for the allocation of buffer storage capacity in serial production lines subject to random failure and repair. In the case that all machines have the same reliability, he recommends maximizing the mean throughput rate by allocating the buffers capacities as nearly as possible equal in size. When the machines are different with respect to their reliability, he proposes to allocate more buffer capacity to less reliable machines. This intuitive result is also supported by Soyster et al. [29].

Jafari and Shanthikumar [15] propose a heuristic solution to determine the optimal allocation of a given total buffer capacity among workstations of a serial production line. Their approximate solution is based on a dynamic programming model with an approximate procedure to compute the mean throughput rate of the line. Smith and Daskalaki [27] have developed a design method-

ology for buffer capacity allocation within assembly lines to approximately solve the optimal buffer allocation problem by maximizing mean throughput rate while minimizing holding and storage costs. Baker et al. [3] have examined the effect of buffers on the efficiency of systems in which two serial lines merge at an assembly workstation. They conclude that small buffers are sufficient to regain most of the lost production capacity and buffer capacity should be allocated equally among the workstations.

So far, we review the researchers that proposed rules for allocating buffers to maximize the mean throughput rate in serial production lines operating with push control strategy. In contrast, Andijani and Clark [1] investigate the optimal allocation of buffers (kanbans) in a pull system by considering both the mean throughput rate and the WIP inventories in the maximized objective function. Recently, Askin et al. [2] utilized a continuous time, steady-state Markov model in determining the optimal number of kanbans to use for each part type at each workstation in a just-in-time production system. Their objective was to minimize the sum of inventory holding and backorder costs. Results indicate a need for increased safety stocks for systems where many part types are produced in the same workstation.

Tayur [31,32] developed some theoretical results – *reversibility* and *dominance* – that characterize the dynamics of kanban-controlled manufacturing systems. His study also provides some insights into the behavior of those systems and greatly reduces the simulation efforts required in an investigation. In a serial periodic pull production system with an infinite supply of raw material to the first stage and subject to stochastic demand for finished product at the last stage:

- Increasing the number of identical stages in series, with keeping all other system parameters the same, decreases the mean throughput rate of the system.
- Increasing the demand arrival rate of finished product, with keeping all other system parameters the same, increases the mean throughput rate of the system.
- Increasing the length of the transfer/review period, with keeping all other system parameters the same, decreases the mean throughput rate of the system.
- Increasing the total work content to be allocated to the stages of the system, with keeping all other system parameters the same, decreases the mean throughput rate of the system.
- Increasing the total number of kanbans to be allocated to the stages of the system, with keeping all other system parameters the same, increases the mean throughput rate of the system.
- Increasing the maximum level of allowed backorders, with keeping all other system parameters the same, increases the mean throughput rate of the system.

The characterization of the optimal allocation of scarce resources in a production system requires further investigation with alternate models and techniques through which the results may fit real-life better [14]. One direction is to try non-exponential processing times with different variations or another direction is to broaden the allocation problem by combining the decisions on buffer storage capacity allocation with workload allocation.

3. Description of the system

The basic production system considered in this paper consists of N stages in tandem (see Fig. 1). 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 the workstation of stage j. 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. 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} till 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} till the end of the period then instantaneously transferred to stock Q_{j+1}^{in}). Finally, W_N performs the final operation of the item, converting component C_{N-1} (from stock Q_N^{in}) into finished product FP (could be alternatively denoted by C_N and stored in Q_N^{out} till

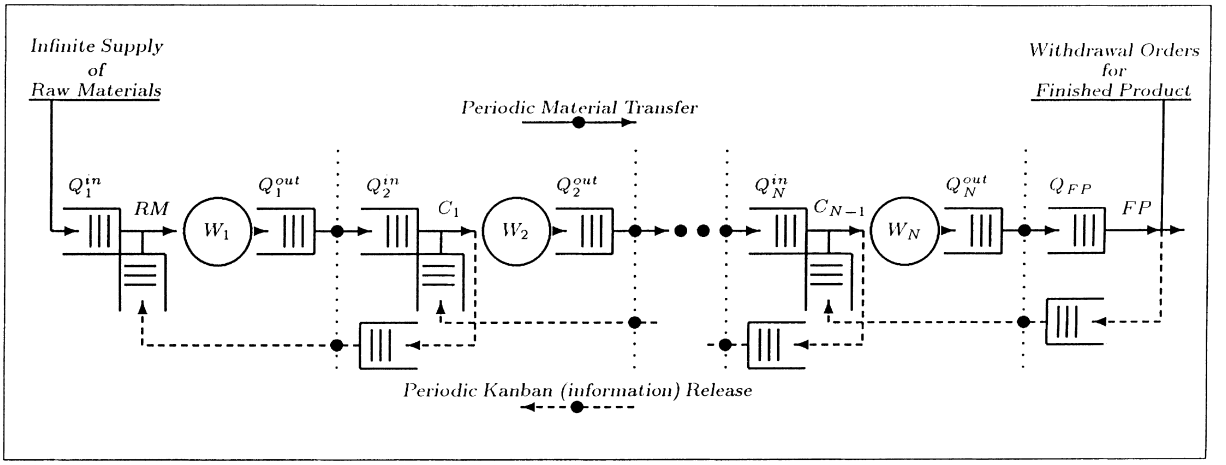

Fig. 1. Tandem arrangement of workstations (W_j: $j = 1, 2, \ldots, N$) in a kanban-controlled periodic pull production line. Each workstation has both an input material queue (Q_j^{in}: $j = 1, 2, \ldots, N$) and an output material queue (Q_j^{out}: $j = 1, 2, \ldots, N$). There are K_j^p production kanbans at workstation W_j and K_j^w withdrawal kanbans circulating between workstations W_j and W_{j-1}.

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 which is the maximum capacity of buffer space allocated for component C_j at workstation W_j. Note that, I_j^{in} $(0 \leqslant I_j^{\text{in}} \leqslant K_{j-1})$ 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} $(1 \leqslant j \leqslant N)$, respectively. Consider the total number of component C_j items between workstations W_j and W_{j+1}, then the inequality for the level of WIP inventories at stocks Q_j^{out} and Q_{j+1}^{in}: $I_j^{\text{out}} + I_{j+1}^{\text{in}} \leqslant K_j$ holds for all stages. However, at the finished product stock Q_{FP} (or alternatively Q_{N+1}^{in}) backordering is allowed up to a maximum allowable amount of B_{FP}. The inventory level at finished product stock is I_{FP} (or alternatively I_{N+1}^{in}, $-B_{\text{FP}} \leqslant I_{N+1}^{\text{in}} \leqslant K_N$).

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 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 of T time units (transfer/review cycle time). That is workstation W_j produces components C_j in order to maintain the inventory level of stock Q_{j+1}^{in} at K_j. Without loss of generality, the production system is assumed to have the same transfer/review cycle times among all stages.

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 material handling function. Then, in the context of 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 time convention used in this study is *beginning of period* in evaluating any state parameter of the system. But, $I_j^{\text{out}}(k)$ denotes the inventory level at 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 the production 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 $(1/\lambda)$ time units. Although backordering is al-

lowed, an arriving finished product demand finding an amount of B_{FP} backordered FP items (that means, I_{N+1}^{in} or alternatively I_{FP} is equal to $-B_{FP}$) is lost. The processing times are assumed to be exponentially distributed. The mean processing time at workstation W_j is $(1/\mu_j)$ time units. For simplification, the workstations are assumed to be reliable. As a result, there are $N+1$ stochastic processes involved in the formulation of the system.

The long-term behavior of the system. In this formulation, the limiting distribution of the states of the system $\vec{\pi}$, of size $|\mathscr{E}|$, could be found (if it exists) by solving the stationary equations of the Markov chain under consideration with the boundary condition imposed:

$$\vec{\pi}M = \vec{\pi} \quad \text{and} \quad \vec{\pi}\,\vec{e}^{\mathrm{T}} = 1,$$

where $\vec{e}$ is a row vector with all elements equal to one, $\vec{\pi}$ the unique solution of the above transition and the boundary equations. A discussion on a variety of methods to compute the stationary probabilities of large Markov chains can be found in [5,22].

Mean throughput rate. Considering the long-term behavior 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

$$\mathrm{MTR} = \mathrm{MTR}_N = \mathrm{MTR}_{N-1} = \cdots = \mathrm{MTR}_2 = \mathrm{MTR}_1.$$

A single-item multi-stage stochastic periodic pull production system is considered in this study to investigate the impacts of system parameters on the mean throughput rate of the system. All descriptive and modelling details of this production system can be found in [17].

4. Statement of the problem

After the brief discussion about the system parameters and the mean throughput rate of the system, it appears that we must progress to the integration of all system parameters simultaneously in the setting of a scarce resource allocation problem. That is, given a set of parameters, the problem is to determine the best choice of these parameters in order to optimize the performance of the system.

Other than the integration of two allocation problems, the basic model utilized here is essentially the same as the previous studies in the literature. The system consists of N production stages corresponding to N workstations in series. Suppose that the set of all production operations required to transform a raw material into a finished product (which is also called the total work content) requires a total of TWC time units. That is, the sum of the mean processing times at all stages, $\sum_{j=1}^{N} 1/\mu_j$, is TWC. On the other hand, the total number of kanbans available for buffer storage in the system (excluding the input buffer stock of the first stage), $\sum_{j=1}^{N} K_j$, is TNK which corresponds to the maximum number of in-process materials and finished product allowed in the system at any instant.

The primary measure of performance of the system is assumed to be the mean throughput rate $\mathrm{MTR}(\vec{\mathscr{W}}, \vec{\mathscr{K}})$, where $\vec{\mathscr{W}} = (1/\mu_1, 1/\mu_2, \ldots, 1/\mu_N)$ represents the allocation of workload to workstations and $\vec{\mathscr{K}} = (K_1, K_2, \ldots, K_N)$ represents the allocation of kanbans between workstations.

The basic problem is to find the allocation vectors $\vec{\mathscr{W}}$ and $\vec{\mathscr{K}}$ which maximizes $\mathrm{MTR}(\vec{\mathscr{W}}, \vec{\mathscr{K}})$ subject to workload and kanban constraints. In the below formulation of the problem, the parameters N, TWC and TNK are fixed constants, whereas the μ_j are continuous and the K_j are integer decision variables:

$$\begin{aligned}
\text{maximize} \quad & \mathrm{MTR}(\vec{\mathscr{W}}, \vec{\mathscr{K}}) \\
\text{subject to} \quad & \sum_{j=1}^{N} 1/\mu_j = \mathrm{TWC}, \\
& \sum_{j=1}^{N} K_j = \mathrm{TNK}, \\
& 1/\mu_j > 0, \quad K_j > 0 \quad \text{and} \quad K_j \text{ integer} \\
& \quad \text{for } j = 1, 2, \ldots, N.
\end{aligned}$$

The above optimization model can be viewed as a linearly constrained mixed integer non-linear programming problem, where the non-linear function $\mathrm{MTR}(\vec{W},\vec{K})$ cannot be expressed explicitly. Even if the processing and demand inter-arrival times are assumed to be exponential, the limitation imposed by the number of kanbans will cause the output process not to be Poisson. For this reason closed form solutions for the stationary probabilities of the system are not available and numerical methods should be used.

The evaluation of $\mathrm{MTR}(\vec{W},\vec{K})$ for any given $\vec{W}$ and $\vec{K}$ involves formulating the underlying queuing process as a finite state, discrete time Markov chain, and then using an appropriate numerical procedure (such as the Gauss–Seidel method) to solve the resultant system of linear equations to obtain the stationary distribution of the system. Unfortunately, the number of states in the state space of the involved Markov chain, and so the number of equations to be solved, grows very rapidly with N, K_j and B_{FP}. For many of the cases considered in this study, this number is in the thousands. This rapid growth imposes definite limits on the size of the problem that will be computationally tractable.

For the allocation of workload and kanban, there are several empirically observed properties which are first reported by Hillier and Boling [10] in serial production lines. As summarized below, subsequent studies in the literature have supported the validity of these properties as well.

- *Reversibility*: The mean throughput rate of the system is the same if the allocations are reversed, that is

$$\mathrm{MTR}(\vec{W},\vec{K}) = \mathrm{MTR}(\vec{W}',\vec{K}')$$

  for any arbitrary allocation of workload $\vec{W} = (1/\mu_1, 1/\mu_2, \ldots, 1/\mu_N)$, its mirror image is $\vec{W}' = (1/\mu_N, 1/\mu_{N-1}, \ldots, 1/\mu_1)$ and for any arbitrary allocation of kanban (buffer storage capacity) $\vec{K} = (K_1, K_2, \ldots, K_N)$, its mirror image is $\vec{K}' = (K_N, K_{N-1}, \ldots, K_1)$.
- *Symmetry*: The optimal allocation of both workload and kanban (buffer storage capacity) which maximizes the mean throughput rate is symmetric, that is

$$1/\mu_j = 1/\mu_{N+1-j} \quad \text{and} \quad K_j = K_{N+1-j} \quad \text{for } j = 1,2,\ldots,N.$$

- *Monotonicity* (or bowl phenomenon): The workstations receive a decreasing amount of workload or an increasing amount of buffer storage capacity as they get closer to the center of the production line, that is:
  - *in terms of workload allocation*:

$$1/\mu_{j-1} > 1/\mu_j \quad \text{for } 2 \leqslant j \leqslant \left\lceil \frac{N}{2} \right\rceil,$$

$$1/\mu_j < 1/\mu_{j+1} \quad \text{for } \left\lfloor \frac{N}{2} \right\rfloor < j \leqslant N-1$$

  or
  - *in terms of kanban allocation*:

$$K_{j-1} < K_j \quad \text{for } 2 \leqslant j \leqslant \left\lceil \frac{N}{2} \right\rceil,$$

$$K_j > K_{j+1} \quad \text{for } \left\lfloor \frac{N}{2} \right\rfloor < j \leqslant N-1.$$

None of these properties has been proven yet. However, note that the reversibility property immediately implies that if the optimal solution is unique then it must satisfy the symmetry property.

It is empirically shown that the number of serious candidates to be an optimal allocation is generally small. The number of feasible allocations that need to be evaluated can be reduced greatly by using two key theoretical results, reversibility and the concavity of the mean throughput rate function with respect to allocation of both workload and buffer storage capacity [31,32,35,37].

5. Experimental study

These structural results together with the performance of balanced systems (more or less similar to unbalanced systems within 1% or 2% of the optimal) imply that an optimal allocation could be found in some neighborhood of a balanced allocation. Therefore, rather than using an optimum seeking search procedure, an enumeration approach is to be used in this study. An unbalancing measure which shows the degree of imbalance in an arbitrary allocation is to be defined as follows:

- For the allocation of workload:

$$\mathrm{DI}_w = \frac{\max_{1\leqslant j\leqslant N}(1/\mu_j) - \min_{1\leqslant j\leqslant N}(1/\mu_j)}{t^0},$$

where TWC is assumed to be equal to $N \times 10 \times t^0$ ($10 \times t^0$ is the average processing time for each stage) and t^0 is the elemental operation time.

- For the allocation of kanban:

$$\mathrm{DI}_k = \max_{1\leqslant j\leqslant N}(K_j) - \min_{1\leqslant j\leqslant N}(K_j).$$

5.1. Design of experiment

An experiment is designed in order to investigate the optimal allocation of both workload and kanban in multi-stage single-item pull production systems in which the Poisson demand arrives at the last stage with a mean rate of λ. The demand arrivals during the times the finished product buffer is empty are lost (backordering is not allowed, $B_{\mathrm{FP}} = 0$). At each stage of the system, the processing times are exponential with the mean $1/\mu_j$, where $\sum_{j=1}^{N} 1/\mu_j = \mathrm{TWC}$ and the number of kanbans allocated is K_j, where $\sum_{j=1}^{N} K_j = \mathrm{TNK}$. 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 periods. It is also assumed that the raw material supply for the first stage is infinite and the material handling times between stages are zero.

In the context of this experiment, 48 two-stage systems, 36 three-stage systems and 20 four-stage systems are evaluated. The framework of the experiment is as follows:

- *Case 1*: Two-stage systems.
  - Mean demand arrival rate is fixed, $\lambda = 1.0$.
  - Total work content is set equal to three different levels, $\mathrm{TWC} = 1.0, 1.50, 2.0$, corresponding to three different levels for the demand load, $\rho = 0.50, 0.75, 1.0$.
  - Total number of kanbans is varied from 2 to 9, $\mathrm{TNK} = 2, 3, 4, 5, 6, 7, 8, 9$.
  - Length of the transfer/review period is set to two different values, $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system.
  - The maximum allowable value for the degree of imbalance is less than or equal to 5, that is $\mathrm{DI}_w \leqslant 5$ and $\mathrm{DI}_k \leqslant 5$.
- *Case 2*: Three-stage systems.
  - Mean demand arrival rate is fixed, $\lambda = 1.0$.
  - Total work content is set equal to three different levels, $\mathrm{TWC} = 1.50, 2.25, 3.0$, corresponding to three different levels for the demand load, $\rho = 0.50, 0.75, 1.0$.
  - Total number of kanbans is varied within two disjoint sets, $\mathrm{TNK} = 3, 4, 5$ and $12, 13, 14$.
  - Length of the transfer/review period is set to two different values, $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system.
  - The maximum allowable value for the degree of imbalance is less than or equal to 5, that is $\mathrm{DI}_w \leqslant 5$ and $\mathrm{DI}_k \leqslant 5$.
- *Case 3*: Four-stage systems.
  - Mean demand arrival rate is fixed, $\lambda = 1.0$.
  - Total work content is set equal to two different levels, $\mathrm{TWC} = 2.0, 4.0$, corresponding to two different levels for the demand load, $\rho = 0.50, 1.0$.
  - Total number of kanbans is varied from 4 to 8, $\mathrm{TNK} = 4, 5, 6, 7, 8$.
  - Length of the transfer/review period is set to two different values, $T = 0.0001, 1.0$, where $T = 0.0001$ approximates the continuous review instantaneous order pull system.
  - The maximum allowable value for the degree of imbalance is less than or equal to 4, that is $\mathrm{DI}_w \leqslant 4$ and $\mathrm{DI}_k \leqslant 4$.

In order to obtain the general behavior of the systems in some neighborhood of balanced allocations, 960 two-stage, 18786 three-stage and 26040 four-stage MTR functions are evaluated by solving the involved one-step transition matrices obtained from discrete-time Markov chain models.

5.2. Empirical results

We will present our findings on the optimal allocation of workload and kanban by focusing on two-, three- and four-stage pull production lines, respectively. In the context of the designed experiment 104 different systems are evaluated in 500

(on the average) different configurations. Because of the huge amount of raw I/O data (input: 462,234 data items and output: 995,334 data items), we will briefly discuss some of the findings as empirical observations, factorial regression models and optimal allocations.

5.2.1. Empirically observed properties

Throughout the experiments, according to optimal allocation results the properties – reversibility, symmetry and monotonicity (or bowl phenomenon) – are not verified. *The periodic pull production system* modeled and analyzed in this study is *not reversible*. The stages closer to the finished product demand require more resources (more production rate and/or more buffer storage capacity) relative to the stages closer to raw material supply. This is because of our infinite assumption of raw material supply to the first stage.

Then, the empirical results show that the optimal allocation is *not symmetric*. The optimal allocation in general follows a pattern of decreasing workload and increasing kanban allocation towards the end of the production line. As a result, the bowl-phenomenon is not observed in these periodic pull production lines. Although we have evaluated all possible allocations within the limitations on DI_w and DI_k, giving preferential treatment to center workstations does not yield better mean throughput rates than we found by giving preferential treatment to the ending stages which are closer to finished product demand.

In the correlation analysis of the MTR and its independent factors (input parameters defining the whole system) this result is also verified. Mean throughput rate of the system is negatively correlated with TWC and positively correlated with TNK as it is intuitively clear. It is observed from Table 1 that, the correlation coefficients of both the amount of workload and the number of kanbans allocated to stages is monotone increasing towards the end of the production line. Thus, the preferential treatment should be focused on the last stages whose allocation variables are the most significantly correlated to MTR. See Table 1 for the correlation coefficients of K_1 and K_2 as -0.0062 and 0.6157, respectively. Although, TNK is positively correlated with MTR, small negative correlation of K_1 is simply because of $K_1 + K_2 = \text{TNK}$. This means that increasing the number of kanbans in the first stage directly decreases the number of kanbans in the second (last) stage. Since, the production capacity lost due to decreasing the number of kanbans in the second stage is significantly greater than the production capacity gained due to increasing the number of kanbans in the first stage, the correlation coefficient of K_1 is

Table 1
Correlation analysis of the factors affecting the mean throughput rate of two, three and four-stage systems

Workload factors	Dependent factor: MTR		Buffer factors	Dependent factor: MTR	
	Continuous approximated by $T = 0.0001$	Periodic with $T = 1.0$		Continuous approximated by $T = 0.0001$	Periodic with $T = 1.0$
TWC_2	−0.6491	−0.3037	TNK_2	0.5295	0.6933
$1/\mu_1$	−0.5202	−0.2540	K_1	−0.0062	0.2188
$1/\mu_2$	−0.6229	−0.2808	K_2	0.6157	0.5793
TWC_3	−0.7416	−0.4279	TNK_3	0.5006	0.6278
$1/\mu_1$	−0.5751	−0.3396	K_1	0.0964	0.1770
$1/\mu_2$	−0.6205	−0.3627	K_2	0.1452	0.2272
$1/\mu_3$	−0.6769	−0.3782	K_3	0.4872	0.5104
TWC_4	−0.8404	−0.6843	TNK_4	0.1802	0.2539
$1/\mu_1$	−0.7466	−0.6169	K_1	−0.1377	−0.1271
$1/\mu_2$	−0.7589	−0.6249	K_2	−0.0910	−0.0359
$1/\mu_3$	−0.7712	−0.6292	K_3	−0.0082	0.0194
$1/\mu_4$	−0.8003	−0.6344	K_4	0.4170	0.3975

turned out to be negative. A similar effect is also observed for four stage systems.

On the other hand, concavity is the only property of mean throughput rate function observed empirically in all cases. It is very difficult to visualize the concavity of MTR function of systems with three or more stages on a three-dimensional graph. See as an example of the mean throughput rate function of a two-stage periodic pull system around the balanced allocation in Fig. 2.

In periodic systems, with decreasing the transfer/review period length T, the mean throughput rate is increased. Thus, the mean throughput rate of a system controlled periodically is always lower

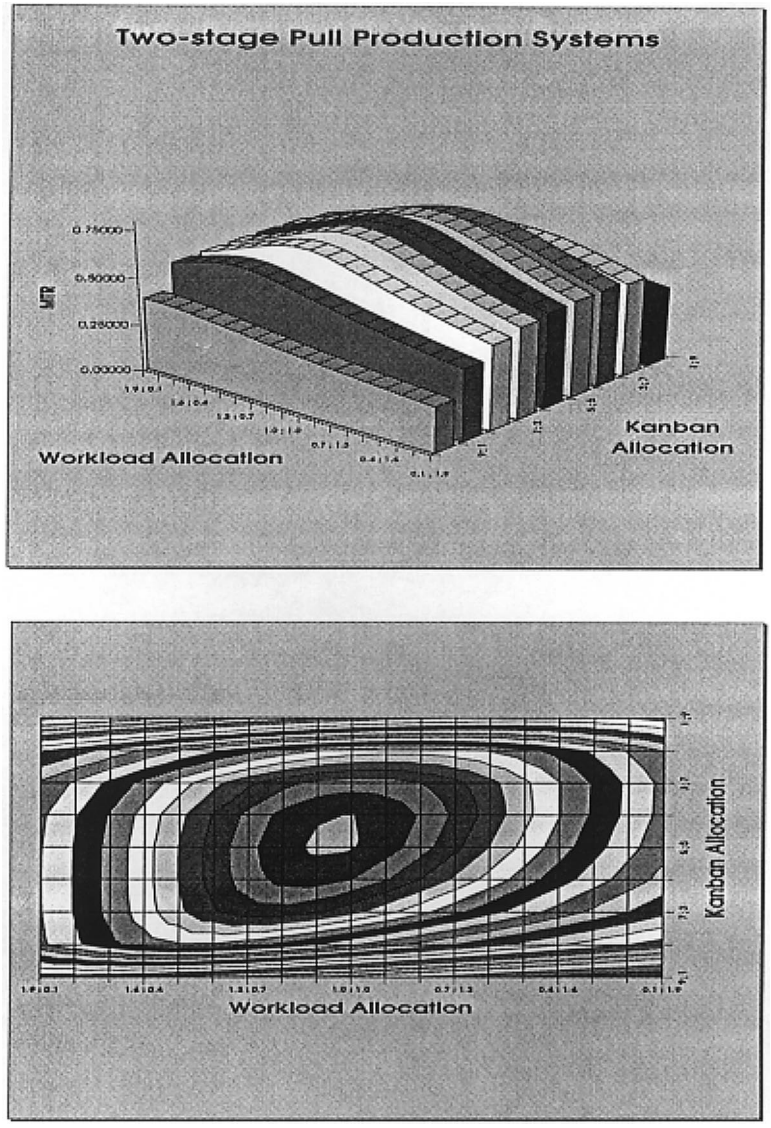

Fig. 2. The mean throughput rate function in a two-stage periodic pull production system. The function is concave with respect to both allocation of workload and kanbans. In the contour plot, the maximum is at the quadrant in which the second stage gets less workload and more number of kanbans. (Fixed parameters of the two-stage system: mean demand arrival rate $\lambda = 1.0$; transfer/review period length $T = 1.0$; total work content TWC = 2.0; total number of kanbans TNK = 10).

than its continuous counterpart. But, on the other hand, the periodic systems carry less inventory than the continuous systems. There is a trade-off between throughput and the inventory depending on the transfer/review period length so that one cannot prefer continuous control, simply that the system could produce more relative to its periodic counterpart, without further analysis of the cost structure.

5.2.2. Factorial regression models

The amount of output data obtained throughout the experiment is very large so that one cannot simply analyze the whole data and point out some rules for the optimal allocation of workload and kanban in pull production systems. In order to summarize the output data some regression models are utilized.

In this regression analysis, there is a single dependent variable (or response) $\mathrm{MTR}(\vec{\mathscr{W}}, \vec{\mathscr{K}})$, that depends on $2 \times N$ independent (or regressor) variables $\vec{\mathscr{W}}$ and $\vec{\mathscr{K}}$. The relationship between these variables is characterized by a mathematical model. The regression model is fit to the output data obtained from the designed experiment. However, the true functional relationship between the response and the regressors is unknown.

Linear factorial regression model:

$$\mathrm{MTR}^{1}_{\mathrm{reg}}(\vec{\mathscr{W}}, \vec{\mathscr{K}}) = a_0 + \sum_{i=1}^{N} a_i 1/\mu_i + \sum_{i=1}^{N} a_{N+i} K_i.$$

Here, we like to determine the linear relationship between the single response variable and the regressor variables. The unknown parameters in the above linear factorial regression model are called regression coefficients and the method of least squares is used to estimate them. Some of the statistical measures showing how well the linear factorial regression model fits the data for two-stage pull systems is summarized in Table 2. The linear factorial regression model fits better to data of continuous pull systems than the data of periodic pull systems. One of the most important measures, R-square, showing the proportion of variability in the data explained or accounted for by the regression model is above 0.8 for continuous pull systems and 0.6 for periodic pull systems. Another measure, mean square error, showing the average error per data point of the regression model is around 0.01. These are quite satisfactory results for linear factorial regression model. The significance of these linear models is that the coefficient estimates point the stage where the preferential treatment (less workload and more kanban) should be focused.

Table 2
The summary of factorial regression models between the independent factors and the mean throughput rate of a two-stage pull system

	Continuous approximated by $T = 0.0001$			Periodic with $T = 1.0$		
	Linear	Quadratic	MTR	MTR	Quadratic	Linear
Mean	0.7616	0.7616	0.7616	0.5831	0.5831	0.5831
St. deviation	0.1322	0.1408	0.1427	0.1813	0.1742	0.1430
Variance	0.0175	0.0198	0.0204	0.0329	0.0304	0.0205
CV	17.3547	18.4893	18.7359	31.0944	29.8799	24.5277
Skewness	0.0290	−0.1113	−0.2307	0.0552	0.3057	−0.2017
Kurtosis	−0.5507	−0.6135	−0.8185	−1.1406	−0.7358	−0.5792
Minimum	0.4500	0.4107	0.4269	0.2830	0.2452	0.2232
Maximum	1.0842	1.0419	0.9907	0.9411	1.0049	0.8900
Corl. coefficient	0.9263	0.9868	1.0000	1.0000	0.9609	0.7888
R-square	0.8580	0.9739	1.0000	1.0000	0.9234	0.6222
SS (error)	1.3850	0.2550	0.0000	0.0000	1.2059	5.9478
MS (error)	0.0029	0.0005	0.0000	0.0000	0.0026	0.0125
F-Value	717.5100	1237.2100	∞	∞	400.4200	195.5900
DF	4	14	480	480	14	4

- *Two-stage systems*: The coefficient estimates of linear factorial regression model has the relation, $a_1 > a_2$ and $a_3 < a_4$. This means: in order to increase mean throughput rate of the system allocate less workload and more kanban to the second stage than the first stage.
- *Three-stage systems*: The coefficient estimates of linear factorial regression model has the relation, $a_1 > a_2 > a_3$ and $a_4 < a_5 < a_6$. This means: in order to increase mean throughput rate of the system a decreasing workload and an increasing kanban allocation should be utilized. The most critical stage that requires preferential treatment is the last stage.
- *Four-stage systems*: The coefficient estimates of linear factorial regression model has the relation, $a_1 > a_2 > a_3 > a_4$ and $a_5 < a_6 < a_7 < a_8$. This means: in order to increase mean throughput rate of the system a decreasing workload and an increasing kanban allocation should be utilized. The most critical stage that requires preferential treatment is the last stage.

Response surface methodology is a collection of mathematical and statistical techniques that are useful for the modelling and analysis of problems in which a response, like mean throughput rate MTR, is influenced by several variables, like workload and kanban allocations $\vec{\mathscr{W}}$ and $\vec{\mathscr{K}}$, and the objective is to optimize the response. If the fitted surface is an adequate approximation of the response function, then analysis of the fitted surface will be approximately equivalent to analysis of the actual system. Since the form of the relationship between the response and the independent variables is unknown, a low-order (second order) polynomial is employed.

Quadratic factorial regression model:

$$\begin{aligned}\mathrm{MTR}^2_{\mathrm{reg}}(\vec{\mathscr{W}},\vec{\mathscr{K}}) &= a_0 + \sum_{i=1}^{N} a_i 1/\mu_i + \sum_{i=1}^{N} a_{N+i} K_i \\ &\quad + \sum_{i=1}^{N}\left[\sum_{j=i}^{N} a_{i,j} 1/\mu_i 1/\mu_j + \sum_{j=1}^{N} a_{i,N+j} 1/\mu_i K_j\right] \\ &\quad + \sum_{i=1}^{N}\sum_{j=i}^{N} a_{N+i,N+j} K_i K_j.\end{aligned}$$

The method of least squares is again used to estimate the regression coefficients. The quadratic factorial regression model better fits the data than the linear model in terms of all statistical measures considered. *R*-square is above 0.9 and 0.8 for continuous and periodic pull systems, respectively. Mean square error is reduced to 0.005. But, on the other hand, individual interpretation of regression coefficients with the inclusion of second order terms becomes meaningless. See Table 3 for the increase in number of terms to be utilized in a third order polynomial relative to linear and quadratic models.

5.2.3. Optimal allocations

Throughout this experiment an overall average of 1.35% improvement is obtained in the mean throughput rate over the balanced (as possible as) systems. See Table 4 for the average improvement in MTR of the systems evaluated. Note that, in the design of experiment, there are several cases in which the total number of kanbans cannot be equally allocated to the stages in the system. In such cases, a composite measure of the degree of imbalance in both allocation of workload and kanban is defined as

$$\mathrm{DI} = \left[1 - \left(\frac{N}{\mathrm{TWC}}\sqrt[N]{\prod_{i=1}^{N}\frac{1}{\mu_i}}\right)\right] + \left[1 - \left(\frac{N}{\mathrm{TNK}}\sqrt[N]{\prod_{i=1}^{N} K_i}\right)\right].$$

This aids to find the most closely balanced configuration with maximized mean throughput rate. The level of the average improvement ob-

Table 3
The number of terms utilized in factorial regression models developed for pull production systems

Factorial regression models	Number of regression terms		
	2-stage	3-stage	4-stage
$\mathrm{MTR}^1_{\mathrm{reg}}(\vec{\mathscr{W}},\vec{\mathscr{K}})$	5	7	9
$\mathrm{MTR}^2_{\mathrm{reg}}(\vec{\mathscr{W}},\vec{\mathscr{K}})$	15	28	45
$\mathrm{MTR}^3_{\mathrm{reg}}(\vec{\mathscr{W}},\vec{\mathscr{K}})$	35	84	165
$\mathrm{MTR}^l_{\mathrm{reg}}(\vec{\mathscr{W}},\vec{\mathscr{K}})$	$1+\sum_{j=1}^{l}\binom{2N-1+j}{j}$		

Table 4
Average MTR of optimal and balanced allocations

	Continuous approximated by $T = 0.0001$		Periodic with $T = 1.0$	
	Optimal	Balanced	Optimal	Balanced
2-Stage	0.7817	0.7695	0.6332	0.6282
3-Stage	0.7567	0.7435	0.6023	0.5900
4-Stage	0.6639	0.6410	0.4112	0.3999

tained is similar to the results reported in the literature. The results regarding the optimal allocation of both workload and kanban in pull production systems could be briefly summarized as follows:

- *General rule*: Select kanbans to allocate first. Allocate kanbans in a monotone increasing pattern in which first stage gets less kanban than the last stage of the system. Allocate workload in a monotone decreasing pattern in which first stage gets more workload than the last stage of the system.
- *Exceptions*: If TNK is low, then the effect of one unit of imbalance in the allocation of kanban is high. That is, giving one kanban to any stage results in high preferment to that stage, instead of taking some amount of this effect back, some extra workload could be transferred to that stage. As a result, in such cases an increasing pattern of workload may give the best performance.
- *Continuous vs periodic*: The number of exceptions increases with the number of stages in the system and also with increasing the length of transfer/review period.

Note that kanban allocation variables are discrete. On the other hand, although workload allocation variables were assumed continuous in the formulation, they are made discrete as multiples of elemental task time t_0 in the context of the experiment. This also causes some exceptions in the optimal allocation of workload.

6. Proposed allocation methodology

The allocation methodology we propose utilizes an evaluative modelling approach. The evaluation of mean throughput rate, $\mathrm{MTR}(\vec{\mathscr{W}}, \vec{\mathscr{K}})$, for any given $\vec{\mathscr{W}}$ and $\vec{\mathscr{K}}$ involves formulating the system as a finite state, discrete time Markov process and then using an appropriate technique to solve the resultant system of linear equations to obtain the stationary distribution of the system. The objective of the allocation methodology is to achieve the maximum mean throughput rate of the system with providing the best set of parameters regarding the allocation of total work content and the total number of kanbans among workstations. In this respect, the process through which the best set of allocation decisions generated is semi-generative. See Ref. [16] for more details on the development of this methodology. Our proposed allocation methodology can be outlined as:

1. Allocate the number of kanbans to workstations as equal as possible.
2. Allocate the amount of workload to workstations as equal as possible.
3. If the resulting configuration is a pure balanced allocation, then all stages are identical to each other. In such a system the last stage which produces the finished product becomes the bottleneck because the other stages on top of their buffer stocks utilize the intermediate buffers of stages up to last stage as extra stocks. So, the system should be configured in such a way that all stages should be bottleneck (critical) at the same instant.
4. Either the resulting system has to possess imbalances because of indivisibility of the operations and precedence relations or not, depending on the total number of kanbans to be allocated, giving more preferential treatment to the last stage might improve MTR. That is:
   (a) If TNK is low,
      (i) allocate the kanbans as equal as possible, if balanced allocation is not possible then allocate more kanban to the last stage(s),
      (ii) select a pattern (decreasing, balanced or increasing) for the allocation of workload depending on the effect of imbalance in the allocation of kanban.
   (b) Otherwise, if TNK is sufficient,
      (i) select a monotone increasing pattern for kanban allocation with special emphasis given to the last stage,

(ii) select a monotone decreasing pattern for workload allocation in which the first stage gets more workload than the last stage.

Note that, decreasing the workload and increasing the number of kanbans in a system have similar effect on mean throughput rate. In this respect they are treated as substitute of each other.

7. Conclusion

In the recent years, with parallel to the developments in manufacturing and computer technology, classical production facilities are being replaced by advanced systems and the companies have entered into a new age of global competitiveness. Because of the scarcity of world's natural resources, it becomes necessary to look for ways of improving productivity and reducing costs through a system of waste elimination. One such system is the JIT production system in which the waste is greatly reduced by adapting to changes. Thus, having all processes produce the necessary parts at the necessary time and having on hand only the minimum stock needed to hold the processes together. The pull production system is a way of implementing the JIT principles, with the finished product 'pulled' from the system at the actual demand rate.

The major decisions for pull production systems are concerned with the allocation of workload (operations) to workstations, the determination of the number of kanbans between workstations and the production/transfer batch sizes. An experiment is designed in order to investigate the optimal allocation of both workload and kanban in two-stage, three-stage and four-stage systems. The results do not support the properties – reversibility, symmetry and monotonicity – in pull production systems. Similar to the results reported by Villeda et al. [33], a decreasing workload and an increasing kanban allocation strategy gives always a consistent improvement (1–10% relative to balanced allocation) in the mean throughput rate. That is, the stages closer to demand are intrinsically bottleneck in a balanced system and requires preferential treatment (less workload and more buffer storage capacity) over the other stages.

With the insight gained in this study, developing both exact and approximate performance evaluation models for multi-item multi-stage pull production systems could be an interesting future research. Note that, when there are more than one item in the system, because of some shared resources, set-up times and scheduling priorities the formulation becomes complicated. The use of vacation queues could be helpful in the development of the approximate model.

References

[1] A.A. Andijani, G.M. Clark, Kanban allocation to serial production lines in a stochastic environment, in: A. Şatr (Ed.), Just-in-time Manufacturing Systems: Operational Planning and Control Issues, Elsevier, Amsterdam, 1991, pp. 175–190.

[2] R.G. Askin, G. Mitwasi, J.B. Goldberg, Determining the number of kanbans in multi-item just-in-time systems, IIE Transactions 25 (1) (1993) 89–98.

[3] K.R. Baker, S.G. Powell, D.F. Pyke, Buffered and unbuffered assembly systems with variable processing times, Journal of Manufacturing and Operations Management 3 (1990) 200–223.

[4] K.R. Baker, S.G. Powell, D.F. Pyke, Optimal allocation of work in assembly systems, Management Science 39 (1) (1993) 101–106.

[5] H. Baruh, T. Altok, Analytical perturbations in Markov chains, European Journal of Operational Research 51 (1991) 210–222.

[6] R. Conway, W. Maxwell, J.O. McClain, L.J. Thomas, The role of work-in-process inventory in serial production lines, Operations Research 36 (2) (1988) 229–241.

[7] T.E. El-Rayah, The efficiency of balanced and unbalanced production lines, International Journal of Production Research 17 (1) (1979) 61–75.

[8] T.E. El-Rayah, The effect of inequality of interstage buffer capacities and operation time variability on the efficiency of production line systems, International Journal of Production Research 17 (1) (1979) 77–89.

[9] S. Gstettner, K. Kuhn, Analysis of production control systems kanban and CONWIP, International Journal of Production Research 34 (11) (1996) 3253–3273.

[10] F.S. Hillier, R.W. Boling, The effect of some design factors on the efficiency of production lines with variable operation times, Journal of Industrial Engineering 17 (1966) 651–658.

[11] F.S. Hillier, R.W. Boling, On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times, Management Science 25 (8) (1979) 721–728.

[12] F.S. Hillier, K.C. So, The effect of the coefficient of variation of operation times on the allocation of storage space in production line systems, IIE Transactions 23 (2) (1991) 198–206.
[13] F.S. Hillier, K.C. So, The effect of machine breakdowns and interstage storage on the performance of production lines, International Journal of Production Research 29 (10) (1991) 2043–2055.
[14] F.S. Hillier, K.C. So, R.W. Boling, Notes: Toward characterizing the optimal allocation of storage space in production line systems with variable processing times, Management Science 39 (1) (1993) 126–133.
[15] M.A. Jafari, J.G. Shanthikumar, Determination of optimal buffer storage capacities and optimal allocation in multi stage automatic transfer lines, IIE Transactions 21 (2) (1989) 130–135.
[16] N. Kırkavak, Modelling and analysis of pull production systems, Ph.D. Dissertation, Department of Industrial Engineering, Bilkent University, Ankara, Turkey, 1995.
[17] N. Kırkavak, C. Dinçer, Performance evaluation models for single-item periodic pull production systems, Journal of the Operational Research Society 47 (2) (1996) 239–250.
[18] M.J. Magazine, G.L. Silver, Heuristics for determining output and work allocations in series flow lines, International Journal of Production Research 16 (3) (1978) 169–181.
[19] S. Meral, A design methodology for just-in-time production lines, Ph.D. Dissertation, Department of Industrial Engineering, Middle East Technical University, Ankara, Turkey, 1993.
[20] E.J. Muth, A. Alkaff, The bowl phenomenon revisited, International Journal of Production Research 25 (2) (1987) 161–173.
[21] S. Payne, N. Slack, R. Wild, A note on operating characteristics of 'balanced' and 'unbalanced' production flow lines, International Journal of Production Research 10 (1) (1972) 93–98.
[22] B. Philippe, Y. Saad, W.J. Stewart, Numerical methods in Markov chain modeling, Operations Research 40 (6) (1992) 1156–1179.
[23] N.P. Rao, A generalization of the 'bowl phenomenon' in series production systems, International Journal of Production Research 14 (4) (1976) 437–443.
[24] B.R. Sarker, Some comparative and design aspects of series production systems, IIE Transactions 16 (3) (1984) 229–239.
[25] B.R. Sarker, R.D. Harris, The effect of imbalance in a just-in-time production system: A simulation study, International Journal of Production Research 28 (5) (1988) 879–894.
[26] T.J. Sheskin, Allocation of interstage storage along an automatic production line, AIIE Transactions 8 (1) (1976) 146–152.
[27] J.M. Smith, S. Daskalaki, Buffer space allocation in automated assembly lines, Operations Research 36 (2) (1988) 343–358.
[28] K.C. So, On the efficiency of unbalancing production lines, International Journal of Production Research 27 (4) (1989) 717–729.
[29] A.L. Soyster, J.W. Schmidt, M.W. Rohrer, Allocation of buffer capacities for a class of fixed cycle production lines, AIIE Transactions 11 (2) (1979) 140–146.
[30] K.E. Stecke, T.L. Morin, The optimality of balancing workloads in certain types of flexible manufacturing systems, European Journal of Operational Research 20 (1985) 68–82.
[31] S.R. Tayur, Properties of serial kanban systems, Queueing Systems 12 (1992) 297–318.
[32] S.R. Tayur, Structural properties and a heuristic for kanban-controlled serial lines, Management Science 39 (11) (1993) 1347–1368.
[33] R. Villeda, R. Dudek, M.L. Smith, Increasing the production rate of a just-in-time production system with variable operation times, International Journal of Production Research 26 (11) (1988) 1749–1768.
[34] A. Wolisz, Production rate optimization in a two-stage system with finite intermediate storage, European Journal of Operational Research 18 (1984) 369–376.
[35] G. Yamazaki, T. Kawashima, H. Sakasegawa, Reversibility of tandem blocking queueing systems, Management Science 31 (1985) 78–83.
[36] G. Yamazaki, H. Sakasegawa, J.G. Shanthikumar, On optimal arrangement of stations in a tandem queueing system with blocking, Management Science 38 (1) (1992) 137–153.
[37] D.D. Yao, Some properties of throughput function of closed networks of queues, Operations Research Letters 3 (6) (1985) 313–317.