ELSEVIER

Int. J. Production Economics 55 (1998) 71–85

international journal of production economics

A study on production lines with discrete batch WIP inventory transfer

Erdal Erel*

Faculty of Business, Administration Bilkent University, 06533 Ankara, Turkey

Received 20 August 1997; accepted 19 February 1998

Abstract

In this paper we consider asynchronous serial lines in which WIP inventory is stored and transported in containers between stations. Containers may be scarce resources; they may cause a reduction in the capacity utilization of the line. The WIP inventory transfer considered involves determining the number of containers to allocate to each buffer location, the interstation distances and the container size. Some conclusions are drawn that may be utilized as guidelines for the practitioners designing such lines. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Serial production line; WIP inventory; WIP inventory transfer

1. Introduction

For the last four decades, several researchers have examined the effect of interstation buffer capacity on the efficiency of production lines. The ratio of flowtime to total processing time of a product in a modern plant could be in excess of 100 to 1 [1]; thus, the investment cost of WIP inventory and the other cost factors related with storage and transportation would reach substantial levels. However, reducing WIP inventory does not necessarily improve performance measures and the "zero inventory" goal is obviously unattainable [1]. Although several aspects of production lines have been analysed, the problem of WIP inventory transfer design still remains an open question. Furthermore, there is very little in the technical literature to guide practitioners on the role and amount of buffers [1].

The distinctive characteristic of the production line considered in this study is that WIP inventory is stored and transported in tote bins, cars, or containers located in the buffer locations between stations. We will call these storage and transport devices as containers hereafter. A container moves to downstream station when it is filled and it moves back to upstream station when it gets empty. Items are taken out of the containers one at a time as they are processed. Similarly, items are filled into the containers one at a time after being processed. In other words, in addition to transporting items between stations, containers are also used to store items while they are waiting to be processed or to

* Tel.: +90 312 2664164; fax: +90 312 2664958; e-mail: erel@bilkent.edu.tr.

PII S0925-5273(98)00042-5

be transported. More than one container can be assigned to buffer locations and there may be some floating containers which are allowed to move to any location if a need for a container arises.

Capacity utilization of a production line is seldom 100% and the loss arises due to station blockages and starvations. A station gets blocked if a completed item cannot be disposed to a container due to the lack of an available container. Thus, the station stays idle until an empty container returns from the downstream station. A station is starved if there are no available items to process. The occurrence of these two events are attributable to variable station processing times and buffer location capacities. The capacity of a buffer location is defined as the rate at which units can pass through the location. Note that a low capacity of a buffer location may cause starvation and blockage of stations. In other words, containers may be scarce resources; an inappropriate container size or insufficient number of containers may cause a reduction in the capacity utilization of the line. Note also that a station may be simultaneously starved and blocked.

The system described above may seem to be inefficient if one considers the containers being held while the items are produced. One can also argue that the containers can be emptied entirely into the input buffer space of the downstream station and return to the upstream station immediately. Similarly, the blockage of upstream station can be prevented by providing an output buffer space into which items can be delivered or piled up if no container is available. The arguments above are, in fact, true for most of the systems, but the characteristics of some items and processes prohibit the piling up of items in the buffer spaces due to some financial, practical and technological reasons as discussed below.

1. Emptying containers into buffer spaces and loading them from buffer spaces may take some time of considerable size. [2] also point out that even in lines with closely located stations, if the WIP inventory is stored in shunt banks, there may be time lost in storing and retrieving them.
2. Emptying containers into the input buffer space of downstream station requires a storage area which may result in a cost of significant size especially with items of big volumes and/or crowded production environments.
3. Average processing time of the items may be longer if they are picked up from the input buffer space into which items are piled up versus from a container in which items are arranged properly.
4. Variability of the processing time may be larger due to the difficulty of picking up items that were piled up.
5. Some device and/or fixtures may be required to keep the items in the input and output buffer spaces and to load and unload them into containers. The cost of designing and operating these additional devices can be of significant size.
6. It may be inconvenient or technologically impractical to empty containers entirely; items can be scratched, damaged, broken, lost, etc.

Production of easily broken and fragile items such as glassware, processing and canning of perishable foods that decay rapidly, production of items with very small and awkward components such as pens that can create an unmanageable mess if an accident occurs during the unloading of the container are some of the examples the system described above is applicable. As stated by [2], in some systems with WIP inventory transfer performed by Automatic Guided Vehicles (AGV), items remain on the AGVs for the tasks at each station. Note that a portion of the line may have the characteristics of the system above; in other portions, items can be piled up in buffer spaces or the workers in adjacent stations which are closely located can reach the buffer location simultaneously. In summary, although the system above is atypical and uncommon, it is implemented and observed in various production environments due to financial, practical and technological factors discussed.

Decision variables of the WIP inventory transfer design problem are the number of containers allocated to each buffer location, distances between stations and container sizes. In this paper, we analyze the effects of these variables on average throughput, defined as the long-run average output per unit time, and WIP inventory. An expression of expected capacity of a buffer location is developed and compared with the average throughput of the line. The relationship between average throughput and WIP inventory is explored. The effects of bottleneck station(s) on average throughput and WIP inventory in the system are also examined; different designs are tested to reduce the detrimental effects of such stations.

2. Literature review

The work to determine production line efficiency and the effect of interstation buffer capacity on it is of significant size. Exact expressions and numerical methods have been developed to determine throughput for lines with a limited length and/or certain processing time distribution functions [3–5]. Several approximate expressions and simulation models have been proposed to derive approximate expressions for longer lines [6–9]. The effect of unbalancing the line in terms of means and variances of the processing times and the interstation buffer capacity have also been examined by several researchers [1,5,10–12]. However, the existing tools to predict the throughput of unbalanced lines are still very limited [13].

The model examined in this paper differs significantly from the studies reviewed above in the sense that containers provide two services: They transport items between stations and hold the items while they are being processed. Similar models in which items are kept in the same containers until the processing of the load is completed are developed by some researchers. For example, [14] examined the effect of discrete batch WIP inventory transfer with zero transport time. [15] studied the problem of material flow control with AGVs; items are kept in the AGVs while operations are performed. [16] addressed the problem of producing equal-size unit loads in multiple stages with material-handling considerations. In another study, [17] considered the selection of machining rates and the number of units to transport at a time between stations to create an overall production plan that minimizes production costs which consist of machining, transportation and overhead. [18] considered a unit load based system with the primary objective of minimizing total cycle time and a secondary objective of minimizing the number of loads moved between stations.

[19] have considered a manufacturing system with stations which have input and output buffers. Material handling devices which operate independently and asynchronously move units between stations in nonzero travel times. An empty device is dispatched to the oldest move request located in the system. They have examined the amount of WIP inventory due to processing (units waiting in input buffers plus those that are being processed) and due to material handling (units in output buffers plus those that are being transported) with a simulation study. They have concluded that WIP due to processing usually far exceeds the one due to material handling. They have also concluded that WIP inventory can be decreased significantly by reducing the variance of processing times, whereas the effect of reducing the variance of transportation times is much smaller. The system examined differs from the system in this article mainly in three respects: Stations have input and output buffers, material flow between stations is specified with a routing matrix, and a centralized dispatching rule is utilized for the empty material handling devices.

Transportation of WIP inventory in nonzero time between stations has also been considered by [20]. They have considered a serial production line with deterministic processing times and unreliable stations which are linked to each other with a conveyor travelling at a constant speed. They have defined an equivalent line with zero transportation time by decreasing the buffer sizes by an amount defined as the ratio of the transportation time and the largest processing time of the line. A simulation study conducted on a two-station line with exponentially distributed up and repair times results small errors in the above approximation. However, for larger lines, the error involved gets unacceptably high values.

The system examined in this article has some common features with flow lines in which units are transported between stations by a closed-loop material handling system. In these systems whenever a unit leaves the system, it is replaced by a new unit; thus, the total WIP inventory is limited with the number of pallets carrying the units. The behaviour of the system is closely affected by the number of pallets or job carriers in the system; throughput is maximized when the number of job carriers equals roughly to half the total number of buffer spaces [2].

In comparison with the studies revived above, the system examined in this article has the following two unique characteristics: Stations do not have input and output buffer spaces and containers provide the service of storing the items while they are being processed.

3. Effect of positive transportation time

Transportation time may be one of the major components of flowtime if the stations are far apart from each other with slow-moving containers; the relative weight of transportation time gets large especially with small number of containers at buffer locations. In some production environments, it is possible to observe adjacent stations of the line to be on different floors or even be in different buildings. A sufficiently large transportation time would cause the containers to act as bottlenecks to have a decreasing effect on throughput. The effect of positive transportation time on average WIP inventory is expected to be as follows: If the containers act as bottlenecks, then a decrease in the average WIP inventory is expected along with the decrease in throughput; however, an increase in the average WIP inventory is expected to occur if the increase in transportation time is not sufficient to convert the containers into bottlenecks (i.e., due to large number of containers).

One can view the transportation of WIP inventory in nonzero time as another station in the line; consequently, all the known results about serial production lines would be applicable. In fact, any line with closely located stations can be transformed into a line with no buffer by adding fictitious stations with zero processing times [21]. In the special case of having a single container with unit size at each buffer location, we can also take a similar approach. However, several difficulties arise if two or more larger-sized containers are utilized at the buffer locations. Due to the nonzero transportation time and containers waiting to be filled and emptied, the characteristics of these processing time distributions would be quite difficult to determine since they are functions of the production rates of adjacent stations, container sizes, and the transportation times. One of the objectives of this study is to examine the effect of varying number of containers, container size and transportation time on the average throughput and WIP inventory of the line; it would be impossible to examine the effects of these factors. Furthermore, the known results are quite limited especially with the unspecified processing time characteristics of the fictitious stations; the known results are either applicable to unrealistically short lines or have restrictive assumptions such as identical and specific processing time distributions, buffers having the same capacity. Thus, in this paper we will not follow the approach of considering the transportation of WIP inventory as fictitious stations.

The system performance measures, the average throughput and WIP inventory, with variable processing times are examined with a simulation model. Data is collected during the production of 100 000 units after a warm-up period. The mean processing time is taken as 1.0, and four processing time distributions are used: Uniform $U(0.8; 1.2)$, truncated Normal $N(1; 0.1667)$, Uniform $U(0.5; 1.5)$, and Exponential $E(1)$. The coefficients of variation of these distributions are 0.115, 0.167, 0.289, and 1.0, respectively, and the level of variability represented by the first three distributions can be encountered in practical manual operations; the last distribution is included to explore how bad things could be for wildly changing processing times [1,22,23]. The mean processing time is 1.0; thus, the maximum achievable throughput is also 1.0. With this scaling, throughput can be interpreted as efficiency [22]. We assume that the first station has unlimited input available and the finished goods are collected at the end storage with infinite capacity. No station breakdowns occur and no nonconforming unit is produced. The notation used is as follows:

C container size in number of units,
M number of stations in the line,

T net line length in time units (in terms of transportation time),

t_j transportation time of the containers in buffer location j, $j = 1, \dots, M-1$,

μ_i expected processing time of station i, $i = 1, \dots, M$,

σ_i standard deviation of the processing time of station i, $i = 1, \dots, M$,

n_j number of containers in buffer location j, $j = 1, \dots, M-1$,

N total number of containers utilized along the line,

P average throughput,

W average WIP inventory.

Note that $T = \sum_{j=1}^{M-1} t_j$ and $N = \sum_{j=1}^{M-1} n_j$.

Fig. 1 depicts the effect of the number of stations on average throughput for $U(0.5; 1.5)$, $C = 5$ and $n_j = 2$, $j = 1, \dots, M-1$, with various container transportation times. Stations are located equi-distantly and it is assumed that the time necessary for a filled container to reach the next station is the same as the one of the empty container to return to the previous station. We observe that P is inversely proportional to M; since a longer line facilitates more stations and buffer locations being interfering with each other and consequently decreasing P. In addition, P seems to decrease to a nonzero limit as M is increased; this conclusion has been previously drawn by several researchers [1,21]. We also observe that increasing container transportation time has a decreasing effect on P as expected. However, the decrease in P occurs mainly in the first four or five stations for all the transportation times tested and similar results are obtained for other processing time distributions, container sizes and number of containers assigned to the locations. This observation supports the earlier findings [14], and similarly we will draw conclusions about the behaviour of lines with positive transportation times by analysing a 6-station line. Accordingly, hereafter a line is assumed to have six stations unless stated otherwise.

The average WIP inventory levels of the lines represented in Fig. 1 are approximately linearly proportional to line length. This observation is an expected one, since all the stations and buffer locations are identical. The effect of transportation time on average WIP inventory is analysed in detail in the next section.

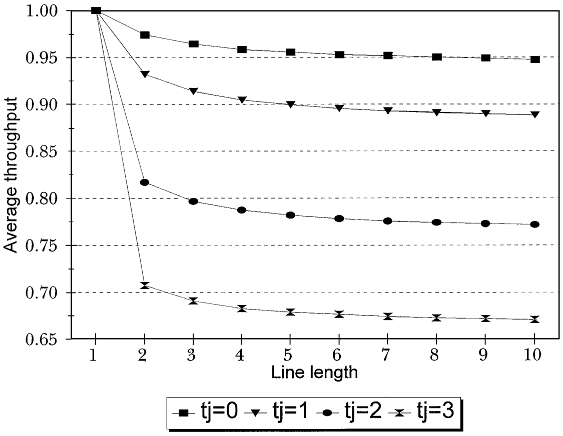

Fig. 1. Effect of number of stations on average throughput.

4. Effect of buffer capacity on throughput and WIP inventory

We first develop an expression for the expected capacity of a buffer location and examine the relationship with the average throughput of the line. We also develop an expression for the average WIP inventory based on the expected capacity of a buffer location and compare it with the ones obtained from the simulation model.

An upper bound on the average throughput of a line, $P_{\max}$, can be expressed as follows:

$$P_{\max} = \min_{i=1,\dots,M,\ j=1,\dots,M-1} \{S_i; L_j\}, \tag{1}$$

where S_i is the expected isolated production rate of station i and L_j is the expected capacity of buffer location j which is defined as the expected rate at which units pass through location j if it were in a system with only the adjacent stations. L_j is measured in number of units transported per unit time. Note that $S_i = 1/\mu_i$ for $i = 1, \dots, M$, and it is the expected production rate that station i would operate at if it were not in a system with the other stations and buffer locations. $P_{\max}$ is specified by a line bottleneck which can be either a station or a buffer location. L_j is specified by μ_j, μ_{j+1}, n_j,

t_j and C; it can be expressed as follows:

$$L_j = \frac{n_j C}{2\left[t_j + (C-1)\frac{\mu_j + \mu_{j+1}}{2}\right]} \quad \text{for} \quad j = 1, \dots, M-1. \tag{2}$$

Eq. (2) is derived from the fact that a container in location j requires, on the average,

$$2\left[t_j + (C-1)\frac{\mu_j + \mu_{j+1}}{2}\right]$$

time units to make a full trip, since no waiting is necessary to load the first unit from the upstream station or after unloading the Cth unit to the downstream station. Note that the expected capacity of a location can exceed the expected production rates of the adjacent stations.

For example, consider a three-station line with $t_j = 2.5$ for $j = 1,2$. Let $\mu_1 = \mu_3 = 1$, $\mu_2 = 1.22$, $n_1 = 3$, $n_2 = 2$ and $C = 10$. The expected production rates of stations 1, 2, and 3 are 1, 0.8196, and 1, respectively. The expected capacities of buffer locations 1 and 2 are 1.2 and 0.8, respectively. Note that $P_{\max}$ is specified by the second buffer location. In other words, an upper bound on the average throughput of the line is 0.8. Suppose further that the processing times are distributed normally with CV = 0.2, then the average throughput obtained from the simulation model is 0.7870.

The difference between the average throughput of the line, P, and $P_{\max}$ is the result of interference between stations and buffer locations. Thus, increasing line length is expected to increase the difference. It is also expected that more interference will occur when the expected production rates of stations and the capacities of buffer locations are close to each other. We have analyzed the relationship between $P_{\max}$ and P of a line by computing the difference between them for various values of C, N, n_j, and t_j, for $j = 1, \dots, M$. Figs. 2–4 depict the difference values of lines with $U(0.5; 1.5)$ distributed processing times and C = 5, 20, and 200, respectively, for various transportation times. The difference values are quite small especially for large C and transportation time values; they are even smaller for the other less variable distributions of $U(0.8; 1.2)$ and $N(1; 0.1667)$. The difference values are the greatest when the expected buffer location capacities are closest to the expected production rates of the stations; namely when L_j is closest to 1, for $j = 1, \dots, M-1$. For example, in Fig. 2, L_j is closest to 1 for t_j values of 0, 1, and 3.5 for n_j = 1, 2 and 3, respectively, and the difference values are the greatest for these points. When L_j for $j = 1, \dots, M-1$ is close to S_i for $i = 1, \dots, M$, more interference between stations and buffer locations occur resulting in the observed increase between P and $P_{\max}$. When L_j for $j = 1, \dots, M-1$ increases, the effect is similar to partitioning the M-station line

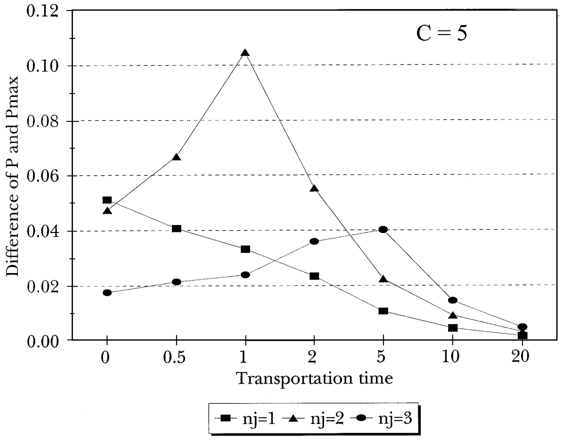

Fig. 2. Effect of expected buffer capacity on the difference of P and $P_{\max}$ for $C = 5$.

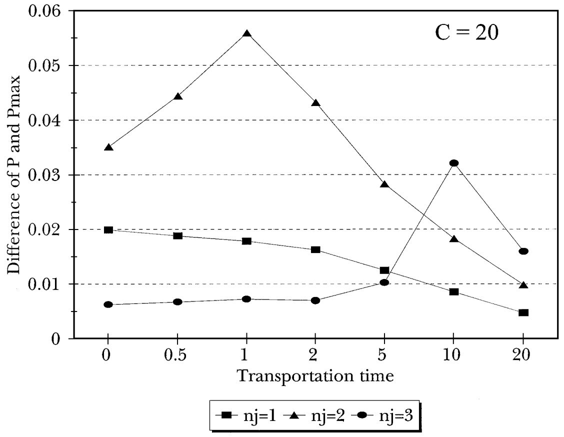

Fig. 3. Effect of expected buffer capacity on the difference of P and $P_{\max}$ for $C = 20$.

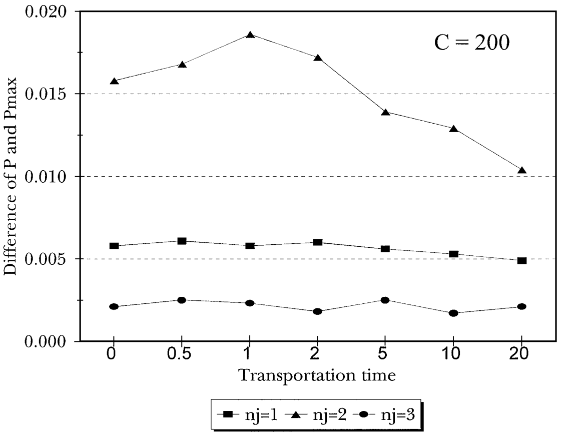

Fig. 4. Effect of expected buffer capacity on the difference of P and $P_{\max}$ for $C = 200$.

into M lines each with one station in length resulting P approaching $P_{\max} = \min\ \{S_i\}$. Similarly, when L_j for $j = 1, \ldots, M-1$ decreases, buffer locations act as bottlenecks resulting P approaching $P_{\max} = \min\ \{L_j\}$.

L_j is a decreasing convex function of C for $t_j < (\mu_j + \mu_{j+1})/2$, an increasing concave function of C for $t_j > (\mu_j + \mu_{j+1})/2$, and $L_j = n_j/(\mu_j + \mu_{j+1})$ for $t_j = (\mu_j + \mu_{j+1})/2$. Note also that $\lim_{C \to \inf}\ L_j = n_j/\mu_j + \mu_{j+1}$. Consequently, for $t_j < (\mu_j + \mu_{j+1})/2$, decreasing C, and for $t_j > (\mu_j + \mu_{j+1})/2$, increasing C would increase L_j.

Assuming the difference between P and $P_{\max}$ is negligible, the average WIP inventory, W, can be approximated with the following expression:

$$W_{\text{est}} = 1 + \sum_{i=2}^{M} P_{\max}\mu_i + \sum_{j=1}^{M-1} \frac{C}{2} n_j^* + \sum_{j=1}^{k-1} (n_j - n_j^*)C, \qquad (3)$$

where k is the order of the line bottleneck and n_j^* is the effective number of containers in location j expressed as follows:

$$n_j^* = \frac{2P_{\max}\left[t_j + (C-1)\dfrac{\mu_j + \mu_{j+1}}{2}\right]}{C}. \qquad (4)$$

The first term in Eq. (3) is associated with the first station; it is never starved. The second term represents the WIP inventory being processed in stations except the first one. The third term is the WIP inventory stored and transported in n_j^* containers for $j = 1, \ldots, M-1$. Note that n_j^* represents the minimum number of containers required to achieve the expected capacity of $P_{\max}$ in location j. Finally, the fourth term represents the WIP inventory stored in containers that are in excess of the effective number of containers. The line bottleneck is either a station or a buffer location; if the jth buffer location is the bottleneck (specifying $P_{\max}$), then k is j, or if the ith station is the bottleneck, then k is $i-1$. Note that the containers assigned in excess of the effective number of containers in location j, $(n_j - n_j^*)$, would, on the average, stay filled if $j < k$, and stay idle if $j > k$. Note also that if the jth buffer location is the bottleneck, then $n_j = n_j^*$. If there are two or more bottlenecks, the first one determines k, and the containers in excess of the effective number of containers in the locations between the bottlenecks are expected to stay filled 50% of the time. Thus, if there are two or more bottlenecks with at least one buffer location between them, W_{est} expression should be increased by $\sum_{j=k_1+1}^{k_2-1} (n_j - n_j^*)C/2$, where k_1 and k_2 denote the orders of the first and the last bottlenecks, respectively.

In the three-station-line example above, the second buffer location constitutes a bottleneck and $n_1^* = 2$. W_{est} is calculated as 32.776 whereas the average WIP inventory obtained from the simulation model is 32.750. Note that the third container in the first buffer location is expected to stay filled with no contribution to the transportation of the WIP inventory.

Fig. 5 depicts the percentage error involved using Eq. (3) for a line with $U(0.5;\ 1.5)$ distributed processing times and $C = 5$ for various transportation times. The percentage error is defined as $100(W_{\text{est}} - W)/W$. As depicted in the figure, the errors are negligible. The same behaviour is observed for other values of C and processing time distributions.

Based on the analyses above and the experiments performed on lines with various C and N values, unequal allocations of the N containers to the locations, $U(0.8;\ 1.2)$ and $N(1;\ 0.1667)$ processing times, we can conclude that $P_{\max}$ can be used to estimate P especially for large C; for relatively small C, the

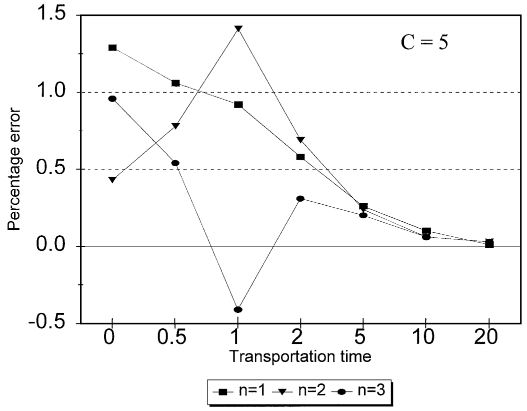

Fig. 5. Percentage error involved using W_{est} for $C = 5$.

approximation is still acceptable if the expected capacities of buffer locations are not close to the expected production rates of stations. However, if the expected capacities of buffer locations and the production rates of stations are very close to each other (analogous to a perfectly balanced serial line) and for relatively small C, $P_{\max}$ may overestimate P by an innegligible amount. On the other hand, considering the lines in practice, $P_{\max}$ can be used to approximate P since bottlenecks are inevitable and containers that can store large number of items are not rare. Similarly, W_{est} can be utilized to estimate W quite accurately.

5. Determining interstation distances

A problem encountered in practice is to locate the intermediate stations between the first and the last stations in such a way that the average throughput is maximized and the average WIP inventory is minimized. The locations of the first and the last stations are usually imposed by the locations of the raw material and finished goods storage areas. The locations of the other stations have typically much less restrictive constraints. The locations of these intermediate stations may affect P and W along with the number of containers assigned to each buffer location and container capacity.

We have analysed lines with different C, N, and T values and determined the optimal locations for the intermediate stations along with the optimal allocation of the N containers to the buffer locations. In other words, we have identified the designs specified by C, n_j and t_j for $j = 1, \dots, M-1$ that maximizes the minimum expected buffer location capacity, $L_{\min}$, defined as $L_{\min} = \min_{j=1,\dots,M-1}\{L_j\}$. The expected processing times of the stations are taken as 1; thus, P cannot exceed 1. However, the expected capacity of buffer location j, L_j, (as well as $L_{\min}$) can exceed 1. Although P cannot exceed 1 (due to $\mu_i = 1$ for $i = 1, \dots, M$), we study and report $L_{\min}$ values larger than 1 as well, since an expected location capacity of 1 (or values even larger than 1) may have a decreasing effect on throughput. A similar analysis has been made by [13]; they have considered the largest expected processing time or within 95% of the largest expected processing time as a bottleneck.

First, we consider lines with N changing from M to $2M-3$; in other words, at least one buffer location would house a single container. It is assumed that at most two containers are assigned to a buffer location. The above assumption follows from the studies made on the optimal allocation of buffer capacity [1,11]; allocation of buffer capacity should be made as nearly equally as possible. It is also assumed that containers travel at a unit speed and the net total distance between the first and the last station is measured in time units. The net total distance, T, is defined as the distance between the first and the last station minus the distance occupied by the stations. Let the superscripts $'$ and $''$ indicate the single- and double-container locations, respectively; for example, t''_j represents the transportation time in location j which houses two containers. In the design that maximizes $L_{\min}$, the interstation distances of all the single-container locations are equal to each other; the same property holds also for the double-container locations. Note that L_j can be increased by decreasing t_j; however, the amount of decrease in t_j is offset by another buffer location and consequently, $L_{\min}$ is decreased. Thus, $t' = t''_j$ and $t'' = t''_j$ for all j. Let u_j represent $(\mu_j + \mu_{j+1})/2$ for $j = 1, \dots, M-1$. Note that $u = u_j$ for all j. Since the expected capacities of the single- and double-container locations can be equated to each other by varying the distances between the stations, the relation between t' and t'' can be expressed as follows: $t'' = 2t' + u(C-1)$. Note that there are $2M - N - 2$ single-container locations and $N - M + 1$ double-container locations; thus, t' can be expressed as follows:

$$t' = \frac{T - (N - M + 1)u(C-1)}{N}. \tag{5}$$

t' decreases as C is increased, and is equal to zero for

$$C \geqslant \frac{T + u(N - M + 1)}{u(N - M + 1)}.$$

For $t' > 0$, $L_{\min}$ is an increasing and concave function of C for $T > u(M-1)$ and a decreasing and convex function of C for $T < u(M-1)$. For $t' = 0$, $L_{\min}$ is a decreasing and convex function of C. When N is set to a multiple of $M-1$, t' and t'', corresponding to $M-1$ and $2(M-1)$ containers, respectively, equal to $T/(M-1)$. On the other hand, it is clear from the literature that the average WIP inventory is an increasing function of C.

The analysis above indicates that for $T < u(M-1)$, W is a decreasing function of $L_{\min}$. For $T > u(M-1)$, W is an increasing function of $L_{\min}$ if N is a multiple of $(M-1)$; otherwise, W is an increasing function of $L_{\min}$ for $t' > 0$, and a decreasing function of $L_{\min}$ for $t' = 0$. The above analysis provides valuable information to practitioners due to the relationship between $L_{\min}$ and P. Utilizing small-sized containers is the best policy if $T < u(M-1)$; otherwise, determining container size depends on the costs of holding inventory and underutilization of the line capacity.

The analysis can be easily extended to the lines with larger number of containers. For example, if $2M - 1 \leqslant N \leqslant 3M - 4$ (that is, at least one buffer location would house two or less containers), then $t''' = 1.5t'' + ((C-1)/2)u$, where

$$t'' = \frac{2T - (C-1)u(N - 2M + 2)}{N},$$

where t''' is the transportation time of the containers in triple-container locations.

Figs. 6–8 depict the relationship between W and $L_{\min}$ for T values of 4, 6 and 25 time units, respectively. Note that these figures belong to 6-station lines with $\mu_i = 1$ for all i. The curves are constructed by varying C for different number of containers utilized along the line. For example, in Fig. 7 with $N = 6$ containers, it is possible to design a line with

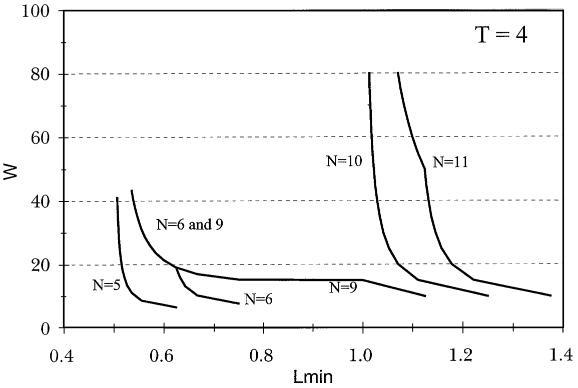

Fig. 6. Relationship between $L_{\min}$ and W for $T = 4$.

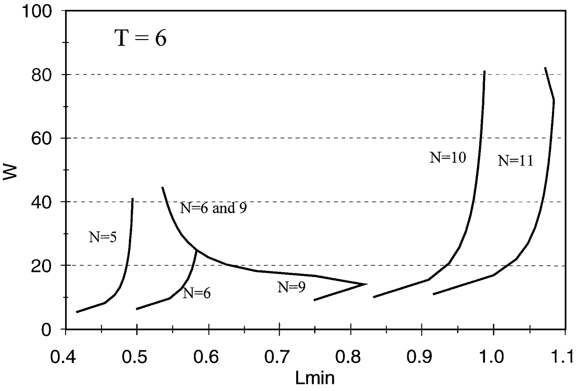

Fig. 7. Relationship between $L_{\min}$ and W for $T = 6$.

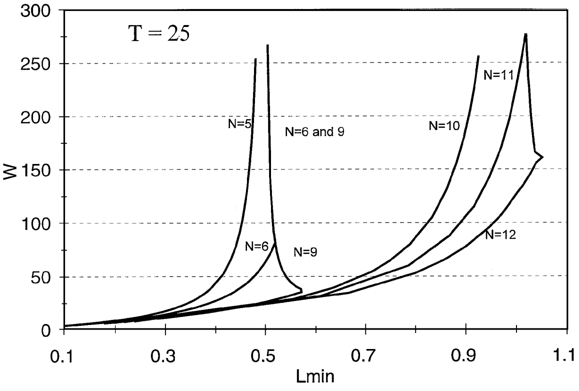

Fig. 8. Relationship between $L_{\min}$ and W for $T = 25$.

$L_{\min}$ equal to 0.5 and W equal to 6.5 or with $N = 5$ containers a line with $L_{\min}$ equal to 0.4934 and W equal to 40.9671; the container sizes associated with these designs are 1 and 15, respectively. Figs. 6 and 7 are obtained by increasing C from 1 to 15, and in Fig. 8, C is increased up to 100. In Fig. 6, W is a decreasing function of $L_{\min}$, since $T = 4 < u(M-1) = 5$. Note that the same $L_{\min}$ may correspond to different W values for different N. In Figs. 7 and 8, W is an increasing function of $L_{\min}$ for $N = 5$ and 10, since 5 and 10 are multiples of $M-1$; for the other values of N, W is an increasing and decreasing function of $L_{\min}$ for $t' > 0$ and $t' = 0$, respectively.

Based on the analysis above, we can draw the following conclusions:

1. If $T \leqslant u(M-1)$, container size should be minimized. For a fixed value of N, increasing C decreases $L_{\min}$ and increases W. With an appropriate choice of C, the same $L_{\min}$ can be achieved with various N values and higher N values correspond to higher W values. In other words, if the line length measured in terms of transportation time is small, then small containers should be used.
2. If $T > u(M-1)$ and N is a multiple of $M-1$, then increasing C increases both $L_{\min}$ and W. Thus, the costs associated with W and P should be considered to determine the container size. In other words, if the line length is large (larger than $u(M-1)$) measured in terms of transportation time and the number of containers is a multiple of $M-1$, then using large containers increases $L_{\min}$, but also the average WIP inventory.
3. If $T > u(M-1)$ and N is not a multiple of $M-1$, increasing C would decrease t' and t'' to zero in lines with $M \leqslant N \leqslant 2M-3$ and $2M-1 \leqslant N \leqslant 3M-4$, respectively. Specifically, t' and t'' are equal to zero for
$$C \geqslant \frac{T + u(N-M+1)}{u(N-M+1)},$$
and
$$C \geqslant \frac{2T + u(N-2M+2)}{u(N-2M+2)},$$
in line with $M \leqslant N \leqslant 2M-3$ and $2M-1 \leqslant N \leqslant 3M-4$, respectively. Increasing C up to the above values increases both $L_{\min}$ and W; however, increasing C further with the transportation times set to zero would decrease $L_{\min}$ and increase W. Thus, $L_{\min}$ is maximized when
$$C = \frac{T + u(N-M+1)}{u(N-M+1)},$$
and
$$C = \frac{2T + u(N-2M+2)}{u(N-2M+2)},$$
in line with $M \leqslant N \leqslant 2M-3$ and $2M-1 \leqslant N \leqslant 3M-4$, respectively. In other words, if the line length is large (larger than $u(M-1)$) measured in terms of transportation time and the number of containers is not a multiple of $M-1$, then increasing container size increases the average WIP inventory; $L_{\min}$ also increases up to a specific C value, but it decreases beyond this specific C value. Thus, the costs associated with W and P, and the specific C value should be considered to determine the container size.

The conclusions above are tested with the simulation model; the resulting curves closely resemble the curves obtained from the above analysis. This is even true for the most variable processing time of $E(1)$. Table 1 depicts the simulation results of the $N = 6$ curve in Fig. 8. The first column depicts the C values. $L_{\min}$ and W obtained from Eq. (3) are depicted in the second and third columns. Note that these values are used to draw the the $N = 6$ curve in Fig. 8. Simulation results for $U(0.8; 1.2)$ and $E(1)$ are in columns 4–7. Note that the values associated with the uniform distribution follows the theroetical results very closely. The exponential distribution results deviate somewhat from the theoretical results, as expected. Note also that as C increases, W values for both distributions are almost equal to the values obtained from Eq. (3).

In summary, we can state that if $T \leqslant u(M-1)$, then minimizing C assures the maximum $L_{\min}$ (and consequently P), and the minimum W values. In other words, utilizing small container sizes is the best policy; note that utilizing large number of

Table 1
Simulation results of P and W for $N = 6$ and $T = 25$

C	Theoretical results		Simulation results			
			$U(0.8; 1.2)$		E(1)	
	$L_{\min}$	W	P	W	P	W
1	0.120	4.600	0.1198	6.452	0.1196	7.331
5	0.333	17.667	0.3274	17.710	0.2877	17.971
10	0.429	33.143	0.4214	33.233	0.3731	33.817
20	0.500	63.500	0.4930	63.681	0.4450	64.779
50	0.510	141.306	0.5058	140.921	0.4736	138.588
100	0.505	266.152	0.5020	265.727	0.4798	262.594

containers increases $L_{\min}$. If $T > u(M-1)$ and N is a multiple of $(M-1)$, then utilizing large container sizes increases both $L_{\min}$ and W; the costs associated with holding inventory and underutilizing the line capacity should be considered to determine C. If $T > u(M-1)$ and N is not a multiple of $(M-1)$, C should not exceed the values given in the third conclusion above; up to those values, the costs associated with holding inventory and underutilizing the line capacity should again be considered.

6. Unbalanced lines

Balanced production lines with equal expected processing times at each station is a rough approximation; real production lines are never perfectly balanced [1] and a perfectly notionally balanced line is virtually unattainable [7]. In addition, nonidentical variance and higher moments of processing times can further unbalance lines. Consequently, identification of bottleneck station(s) that limit the average throughput of the line becomes a difficult task; there are conflicting views and confusion in the literature [24]. For example [25] recommend to increase the buffer near the bottleneck which is the station with the largest expected processing time, whereas the recommendation of [21] is to place buffers near the stations with the greatest disruptions. In this section, unbalanced lines only in terms of expected processing times are examined.

A station with a higher expected processing time would both decrease the expected production rate of the station itself and the capacities of the adjacent buffer locations. If two adjacent stations both have higher expected processing times, then the decreasing effect on the expected capacity of the buffer location between these stations would be magnified. In this section, we have examined the effect of 1 and 2 bottleneck stations on P and W. Bottleneck stations are created by increasing the original expected processing times by 20% with no change in the variances of the processing times. Note that the above procedure is not applicable to exponentially distributed processing times.

The following properties about the effect of a bottleneck station on P and W are expected to be observed. If the bottleneck station is the first or the last station in the line, then the decreasing effect on P is expected to be the least due to the fact that the first and the last stations are adjacent to only one buffer location. In other words, these stations lowers the expected capacity of only one buffer location, whereas stations other than the first and the last one affect the output capacities of two buffer locations. The effect of a bottleneck station on W is clear from the literature and intuition: If the bottleneck station is the first station, then W is expected to be smaller than the one of balanced case, since the bottleneck station would decrease P. As the bottleneck station is shifted towards the end of the line, W is expected to increase; in the buffer locations between the first and the bottleneck station, containers assigned in excess of the effective number of containers would stay, on the average, filled.

Tables 2 and 3 depict the P and W values of the line obtained from the simulation model with $U(0.5;\ 1.5)$, $C = 5$, $t_j = 1$ for all j, single and double-container locations, respectively. Processing times of the bottleneck stations are distributed uniformly between 0.7 and 1.7. Note that the expected production rate of the bottleneck station is 0.833, whereas the expected capacity of the location adjacent to the bottleneck station is 0.4629 and 0.9259 for the single and double-container locations, respectively. In other words, a buffer location and a station constitute the line bottleneck in the single and double-container cases, respectively. The boxes in the first column of the tables represent the stations and the darkened one is the bottleneck station. We observe a slight bowl phenomenon for the single-container location design. This is in accordance with the bowl phenomenon reported in the literature that slightly less work should be assigned to the stations in the middle of the line. As expected, average WIP inventory gets smaller as the bottleneck station is shifted towards the beginning of the line for both of the cases. The balanced cases for the single and double container locations are shown in the first rows. Note that the above observations match with the conjectures above.

Table 2
P and W of the line with single-container locations

Bottleneck station	Average throughput	Average WIP inventory
□□□□□□	0.4667	15.854
■□□□□□	0.4549	15.247
□■□□□□	0.4456	15.194
□□■□□□	0.4450	15.609
□□□■□□	0.4446	16.042
□□□□■□	0.4455	16.497
□□□□□■	0.4552	16.449

Table 3
P and W of the line with double-container locations

Bottleneck station	Throughput	Average WIP inventory
□□□□□□	0.8953	30.569
■□□□□□	0.8275	26.955
□■□□□□	0.8238	28.130
□□■□□□	0.8233	29.613
□□□■□□	0.8236	31.085
□□□□■□	0.8232	32.575
□□□□□■	0.8262	33.729

Tables 4 and 5 depict the P and W values of the line with 2 bottleneck stations, $U(0.5;\ 1.5)$, $C = 5$, $t_j = 1$ for all j, single- and double-container locations, respectively. The expected capacities of the locations adjacent to one and two bottleneck stations are 0.4629 and 0.4310, respectively, for the single-container locations. The figures are 0.9259 and 0.8629 for the double-container locations. The designs in Tables 4 and 5 are arranged in the descending order of P values; the designs with adjacent bottleneck stations are listed at the bottom of the tables. Note that the maximum P value belongs to the design with the maximum distance between the bottleneck stations. Examining the tables reveals that no bottleneck station should be adjacent to another bottleneck station; for the single-container locations the mean of the P values of the designs with and without adjacent bottleneck stations are 0.42558 and 0.44135, respectively

Table 4
P and W of the line with 2 bottleneck stations and single-container locations

Bottleneck station	Throughput	Average WIP inventory
■□□□□■	0.4505	15.870
■□□□■□	0.4436	16.067
□■□□□■	0.4435	15.695
■□□■□□	0.4420	15.677
□□■□□■	0.4420	16.053
□□□■□■	0.4401	16.412
■□■□□□	0.4400	15.282
□■□□■□	0.4390	15.906
□□■□■□	0.4365	16.173
□■□■□□	0.4363	15.582
■■□□□□	0.4269	14.547
□■■□□□	0.4248	15.084
□□■■□□	0.4245	15.795
□□□■■□	0.4248	16.518
□□□□■■	0.4269	17.020

Table 5
P and W of the line with 2 bottleneck stations and double-container locations

Bottleneck station	Throughput	Average WIP inventory
■ □ □ □ □ ■	0.8208	30.449
□ ■ □ □ □ ■	0.8176	30.829
■ □ □ □ ■ □	0.8170	30.047
□ □ ■ □ □ ■	0.8151	31.696
■ □ □ ■ □ □	0.8144	29.135
□ ■ □ □ ■ □	0.8138	30.429
□ □ □ ■ □ ■	0.8112	32.645
■ □ ■ □ □ □	0.8107	28.139
□ ■ □ ■ □ □	0.8098	29.597
□ □ ■ □ ■ □	0.8089	31.282
■ ■ □ □ □ □	0.7957	26.794
□ ■ ■ □ □ □	0.7951	28.456
□ □ ■ ■ □ □	0.7946	30.324
□ □ □ ■ ■ □	0.7952	32.211
□ □ □ □ ■ ■	0.7950	33.795

($p < 0.0005$). For the cases with adjacent bottleneck stations, we have also examined the effect of assigning an extra container between them. Assigning an extra container between the bottleneck stations improves throughput significantly; the mean of the P values increases to 0.45624. The same property holds also for the double-container locations; the mean of the P values of the designs with and without adjacent bottleneck stations are 0.79512 and 0.81393, respectively ($p < 0.0005$). Assigning an extra container between the bottleneck stations improves the mean of the P values to 0.82236.

W is observed to increase as the bottleneck stations are shifted towards the end of the line as stated above as a conjecture. Shifting the single bottleneck station from the first station to the last one increases W by 7.9% and 25.1% for the single and double container locations, respectively. With two bottleneck stations, the figures are 17.0% and 26.1%, respectively.

The above experiments have been conducted with various different settings such as negligible transportation times, three bottleneck stations, other processing time distributions, different expected processing time values for the bottleneck station(s), and similar observations have been made. Based on the analyses and observations above, we can state the following conclusions as guidelines for the design of unbalanced production lines:

1. The effect of bottleneck station(s) on decreasing P is expected to be the least if the station(s) are the first (and the last) stations.
2. The effect of bottleneck stations on decreasing P is expected to magnify significantly if these stations are adjacent to each other.
3. As the bottleneck station(s) are shifted towards the end of the line, W is expected to increase.
4. If two bottleneck stations are adjacent to each other, then increasing the buffer capacity between these stations either by assigning extra containers or by decreasing the distance between them would diminish the decreasing effect on P.

The first two conclusions above are especially valuable to guide practitioners in designing lines. The first conclusion follows from the fact that the first and the last stations are adjacent to only one buffer location; the decreasing effect of the large processing time on the expected buffer output capacity works only on one location. The second conclusion follows directly from the expression of the expected capacity of a buffer location. Practitioners should consider these results in designing lines since the effect on P and W can reach to significant amounts. Bottleneck stations are shifted along the line in various ways; for example, the expected output rates of stations may be determined by the number of workers, and bottleneck stations are shifted by moving workers between stations. Bottleneck stations can also be shifted by assigning different portions of the work to the stations. The third conclusion is clear from the literature and intuitive since a bottleneck at the first station decreases W by restricting input to the rest of the line whereas a bottleneck at the end of the line allows large W since the line accepts any input which then gets blocked at the last station. The last conclusion is also intuitive since the decreasing effect of two adjacent bottleneck stations on the expected output capacity of a buffer location can be increased with an extra container. The last two conclusions are given to complete the analysis on the bottleneck station location.

7. Conclusions

In this paper we have examined some aspects of production lines with discrete batch WIP inventory transfer; the conclusions and the guidelines developed can be employed by practitioners to regain the lost production capacity due to the WIP inventory transfer and to control the level of average WIP inventory. We have developed an expression for the expected capacity of a buffer location in which containers store and transport WIP inventory in nonzero transportation time and examined the relationship with the average throughput of the line. We have also developed an expression for the average WIP inventory; the difference between the two expressions above and simulated values is of negligible size under certain conditions. We have also analysed lines to determine the locations of the intermediate stations given a fixed distance between the first and the last stations. Based on the fixed distance above, we have developed rules to specify container size. Finally, we have examined unbalanced lines in terms of expected processing times and developed some guidelines for the design of such lines.

Although the results above and the ones found earlier [14] answer some of the questions about the design of WIP inventory transfer, there are still several areas remaining to be addressed. Some interesting future research items can be stated as follows: The mode of operation of the line examined is "push"; in other words, stations operate as long as there is an item available to work on. The other modes of operation are "pull" and "CONWIP"; in the pull-mode, operation is started after getting a signal from the adjacent downstream station. As [26] point out, a vacancy at either a station or a buffer signals the need to draw the next item. They also point out that there is no distinction between push and pull modes in a serial line with closely located stations and finite buffers. The "CONWIP" mode of operation is a mixture of pull and push modes; the first station starts operating on an item after getting a signal from the last station. Will the policies developed in this paper stay valid for the other modes of operation? The performance measures considered in this study are average throughput and WIP inventory; performance measures based on average cost or profit per unit time can also be of interest. The variance of throughput should also be studied along with the average throughput; as stated by [21], prediction of the variance is as important as the prediction of the average throughput with the current emphasis on just-in-time production. Finally examining the effect of transporting WIP inventory in containers as considered in this study on assembly systems would be interesting. Simple assembly systems in which two or more serial lines or parallel stations feed the assembly operation have been examined by [22,24,27]; practitioners would benefit from future research addressing more complex and realistic assembly systems in which WIP inventory is transported similarly.

References

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

[2] J.A. Buzacott, J.G. Shanthikumar, Stochastic Models of Manufacturing Systems Prentice-Hall, New Jersey (1993).

[3] F. Hillier, R. Boling, Finite queues in series with exponential or Erlang service time – A numerical approach, Operations Research 15 (1967) 286–303.

[4] E.J. Muth, A. Alkaff, The throughput rate of three- station production lines: A unifying solution, International Journal of Production Research 25 (1987) 1405–1413.

[5] F. 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 (1991) 198–206.

[6] D.E. Blumenfeld, A simple formula for estimating throughput of serial production lines with variable processing times and limited buffer capacity, International Journal of Production Research 28 (1990) 1163–1182.

[7] G.E. Martin, Predictive formulae for unpaced line efficiency, International Journal of Production Research 31 (1993) 1981–1990.

[8] K.R. Baker, S.G. Powell, D.F. Pyke, A predictive model for the throughput of unbuffered three-station serial lines, IIE Transactions 26 (1994) 62–71.

[9] C.M. Liu, S.F. Su, C.L. Lin, Predictive models for performance evaluation of serial production lines, International Journal of Production Research 34 (1996) 1279–1291.

[10] F. Hillier, R. 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] S.G. Powell, Buffer allocation in unbalanced three-station serial lines, International Journal of Production Research 32 (1994) 2201–2217.

[12] R. Pike, G.E. Martin, The bowl phenomenon in unpaced lines, International Journal of Production Research 32 (1994) 483–499.

[13] C.M. Liu, C.L. Lin, Performance evaluation of unbalanced serial production lines, International Journal of Production Research 32 (1994) 2897–2914.

[14] E. Erel, Effect of discrete batch WIP transfer on the efficiency of production lines, International Journal of Production Research 31 (1993) 1827–1838.

[15] P.J. Egbelu, N. Roy, Material flow control in AGV/unit load based production lines, International Journal of Production Research 26 (1988) 81–94.

[16] P.J. Egbelu, Batch production time in a multi-stage system with material-handling consideration, International Journal of Production Research 29 (1991) 695–712.

[17] P.J. Egbelu, Machining and material flow system design for minimum cost production, International Journal of Production Research 28 (1990) 353–368.

[18] W.G. Truscott, Scheduling production activities in multistage batch manufacturing systems, International Journal of Production Research 23 (1985) 315–328.

[19] M.M. Srinivasan, Y.A. Bozer, Which one is responsible for WIP: the workstations or the material handling system?, International Journal of Production Research 30 (1992) 1369–1399.

[20] C. Commault, A. Semery, Taking into account delays in buffers for analytical performance evaluation of transfer lines, IIE Transactions 22 (1990) 133–142.

[21] Y. Dallery, S.B. Gershwin, Manufacturing flow line systems: a review of models and analytical results, Queueing Systems 12 (1992) 3–94.

[22] 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.

[23] H.S. Lau, G.E. Martin, The effects of skewness and kurtosis of processing times in unpaced lines, International Journal of Production Research 25 (1987) 1483–1492.

[24] S.G. Powell, D.F. Pyke, Buffering unbalanced assembly systems, IEE Transactions 30 (1998) 55–65.

[25] E.M. Goldratt, J. Cox, The Goal, North River Press, New York, 1986.

[26] K.R. Baker, S.G. Powell, D.F. Pyke, The performance of push and pull systems: a corrected analysis, International Journal of Production Research 28 (1990) 1731–1736.

[27] K.R. Baker, S.G. Powell, A predictive model for the throughput of simple assembly systems, European Journal of Operational Research 81 (1995) 336–345.