Control Point Policy: Efficiency within Make-to-Order Environments

D J Stockton, Riham Khalil, Jason Ardon-Finch

Control Point Policy: Efficiency within Make-to-Order Environments

Journal:	*International Journal of Production Research*
Manuscript ID:	TPRS-2006-IJPR-0230
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	21-Mar-2006
Complete List of Authors:	Stockton, D J; De Montfort University, Faculty of Computing Sciences and Engineering Khalil, Riham; De Montfort University, School of Engineering & Technology Ardon-Finch, Jason; De Montfort University, School of Engineering & Technology
Keywords:	DISPATCHING RULES, SIMULATION
Keywords (user):	

Control Point Policy: Part 1 - Efficiency within Make-to-Order Environments

David John Stockton B.Sc. (Hons), M.Sc., Ph.D., M.I.E.E., C.Eng.
Jason Ardon-Finch B.A. (Hons), Ph.D
Riham Khalil MBA, PhD

De Montfort University
Department of Mechanical and Manufacturing Engineering
The Gateway, Leicester, UK. LE1 9BH.
Tel: +44 (0) 116 2551551 Ext. 8091
Fax: +44 (0) 116 2577052 E-mail: stockton@dmu.ac.uk

Corresponding author: David John Stockton

Abstract

The recognition of the desire for punctual delivery of products has lead to the use of the *service level* as a common performance criterion for measuring the proportion of products that meet due dates specified by the customer. To successfully increase the *service level*, a manufacturing system may respond more quickly to orders by reducing the levels of in-process inventory in the system and hence decrease throughput times. This paper examines the use of the recently developed *Control Point Policy* (CPP) in improving service levels in re-entrant, 'make-to-order' manufacturing systems and compares its effectiveness with that of the *Critical Ratio* scheduling rule. Simulation studies have been undertaken to provide insight into how and when to apply the CPP policy within such environments with results indicating that, in cases requiring small storage areas between machines, the CPP results in better *service level* performance.

KEYWORDS

Simulation, Takt time, lead time, scheduling, dispatching rules, kanbans, CONWIP, performance measures, Control Point Policy, blocking, fixed buffer, material flow, inter-arrival time, buffers.

1. Introduction

The recognition of the desire for punctual delivery of products has lead to the use of the *service level* as a common performance criterion. The service level is a measure of the proportion of products that are completed on time, i.e., that meet due dates specified by the customer. To successfully increase the service level, a manufacturing system must be able to respond more quickly to customers orders, i.e., decrease throughput times. This is typically accomplished by reducing the levels of in-process inventory in the system. Inventory reduction is a function of both the scheduling and inventory control methods in use, i.e. the basic decisions that must be made when a machine becomes available upon finish processing a part are:

a) from which of its upstream buffers should the machine take a part and which part should it take, i.e. these are essentially scheduling decisions, and
b) should the machine take a part from one of its upstream buffers at all, i.e. these are essentially inventory control decisions. Traditionally, this type of decision has been frequently overlooked with the view that an idle machine is wasting available processing time. The concept of permitting machines to remain idle, however, is

becoming more common in practice with the increased use of inventory control policies.

These two questions, therefore, are addressed, respectively, by the use of scheduling techniques and inventory control release policies.

1.1 Scheduling Techniques

There are numerous methods for scheduling parts through a manufacturing system. Most standard texts on production and operations management, express the view that techniques which aim to provide optimal schedules often do so at the expense of over-simplification or excessive computation. The time necessary for developing a technique which produces optimal schedules, often exceeds the time-scale of the project and the resulting strategy can be too complex to be of practical use. Instead, simple *dispatching rules* tend to be adopted in industry in an attempt to promote 'good', rather than 'optimal', product flow. The problem of scheduling is more complex in make-to-order systems since parts or batches of parts must be manufactured within particular time constraints.

Definitions and variations on dispatching rules, for make-to-order systems, appear in the standard literature; examples include Critical Ratio, Least-Slack, Least-Slack-Per-Operation, and Earliest-Due-Date. The choice of dispatching rule is determined by the type and objectives of the production system. In particular the use of Critical Ratio has been widely adopted within make-to-order systems since this ratio can serve as a measure of how urgent it is that a part should progress to the finished goods buffer. At time t the 'expected remaining processing and waiting time' is calculated for the part. This is the sum of all the remaining operation times and expected queuing times between the current production stage and the finished goods buffer. The critical ratio is then given by:

$$\textit{Critical Ratio} = \frac{\text{Due Date of Part} - t}{\text{Expected Remaining Processing and Waiting Time}}$$

The part, in any of the buffers that feed the machine, with the smallest critical ratio, i.e., the most urgent, is selected for processing and is loaded onto the machine immediately. The critical ratio is regarded as a particularly useful indicator of the part's status. For example, if CR > 1 the part is ahead of schedule, if 0 < CR < 1 the part is behind schedule and if CR < 0 the part is already late.

1.2 Inventory Control

A great deal has been written about the use of *kanbans* ('cards' in Japanese) to control product flow. Berkeley (1992) provides a survey of the literature. Kanban control forms the basis of the Just-In-Time philosophy, pioneered by Taiichi Ohno, within the Toyota Production System. The Toyota Production System is described in detail by Ohno (1988) and Monden (1993).

Kanbans are used to signal the removal of inventory from a buffer which, in turn, authorises the production of a part to replace the one just taken. No manufacturing can occur without such authorisation. As a result, each stage is said to produce 'just-in-time' to meet the demand of downstream machines. Production is ultimately regulated by the demand for products at the last manufacturing stage. This leads to the term and description of kanban-controlled systems as *pull* systems. This lies in contrast to *push* systems where production is managed according to forecasts in demand. A comparison of push and pull systems is provided by Tabe et al. (1980).

Arguments for the use of kanbans are particularly appealing, i.e.:

a) the Just-In-Time philosophy aims to ensure that inventory is held at a minimum,
b) information regarding the end product requirements need only be known explicitly at the final work station on an assembly line,
c) the use of kanbans regulates the flow of material through the system without the need for excessive amounts of paperwork using simple visible control signals are used instead,
d) information flow is closely linked to material flow, and
e) deliveries of raw materials from suppliers may be linked to the kanban system.

CONWIP (constant work-in-process) (Spearman et al., 1990) was introduced in an attempt to make Just-In-Time manufacturing applicable to systems with a higher variety of products. In a system controlled using CONWIP no parts are given to the first machine unless the total inventory (made up of all part types) in the system is below a certain limit. A *basestock* policy (Clark and Scarf, 1960), (Kimball, 1988) is one in which the release of parts to all machines (not just the first) is controlled according to the inventory between that machine and the finished goods buffer. The upper bounds on inventory used in these policies are referred to as the *CONWIP Limit* and *basestock levels*, respectively. The implementation of such token-based control policies is discussed by Buzacott and Shanthikumar (1993).

Bonvik et al. (1997) performed a comparison study, by simulation, into the performance of several production control policies on a four-machine flow line. The policies investigated were kanban, minimal blocking (Mitra and Mitrani, 1990), basestock, CONWIP and a hybrid kanban-CONWIP policy (effectively a CONWIP policy implemented with finite buffer sizes). Both constant and changing demand rates were considered and average inventory and service level were used as performance measures to compare the policies. The results show that, when all policies were run with optimal parameter values, the hybrid policy reduces inventory by 10% to 20% over kanban with basestock and with the results of using CONWIP lying in between.

2. The Control Point Policy

The Control Point Policy (CPP), developed by Gershwin, (1999, 2000) for scheduling work through a make-to-stock system has recently been shown by Gzouli (2000) and Yong (2001), to be capable of outperforming Kanban, CONWIP and basestock control policies.

In this work the adopted methods of experimentation and comparison used are the same as those previously used by Bonvik et al. (1997) and are based upon extensive analysis through discrete event simulation. Since the CPP has only recently been developed no precise rules are available concerning how and when to apply the policy. The work reported in this paper is an extension of that by Gzouli (1999) in that as with the work of Yong (2001), it provides an extension of the CPP method into a make-to-order environment. The CPP is compared to *Critical Ratio*, a widely adopted technique for make-to-order systems. An aim of this research is to demonstrate that the CPP performs well (in terms of the service level), particularly when subject to conditions evident in flexible manpower lines, i.e., when buffer sizes and the *Takt* time are required to be small.

Total buffer space was not included in the analyses by Bonvik et al. and Gzouli. One reason for its inclusion in this investigation, is that the graph of service level against total buffer space can reveal information not obvious from that of service level against average total WIP.

Due to its recent formulation, there are currently no techniques available for selecting values for the parameters of the CPP. Simulation studies such as those by Bonvik et al. and

Gzouli are limited by the number of different parameter configurations that can be examined, since each simulation needed to be allowed to run for a considerable amount of time.

The essential features of the CPP are:

a) *blocking before service*, i.e. which describes a system in which a machine is not allowed to load or begin processing a part if the downstream buffer to which the part is destined is full,

b) *buffer selection sequences*, i.e. in which all buffers upstream of a machine are examined in a fixed order of preference when seeking a buffer from which to progress a part, and

c) *hedging times* which dictate when the machine is allowed to remove parts from upstream buffers.

2.1 Blocking Before Service

In the re-entrant system shown in Figure 1, M_1 is not allowed to take a part from B_0 unless the level, n_1, of the buffer B_1 is lower than its capacity, N_1. Similarly, it is not permitted to take a part from B_2 unless $n_3 < N_3$. This ensures than whenever M_1 finishes work on a part, the part will be able to proceed to the downstream buffer. The corresponding assertion of parts being guaranteed free movement to the downstream buffer is not true for machines M_i, $i = 2,\ldots,9$. For example, M_2 and M_3 may both begin working on parts if $n_2 < N_2$. However, if M_2 finishes working (so that $n_2 = N_2$) shortly before M_3 does, then M_3 will be forced to remain idle whilst containing a part until it can be removed, i.e., until a part is removed from B_2.

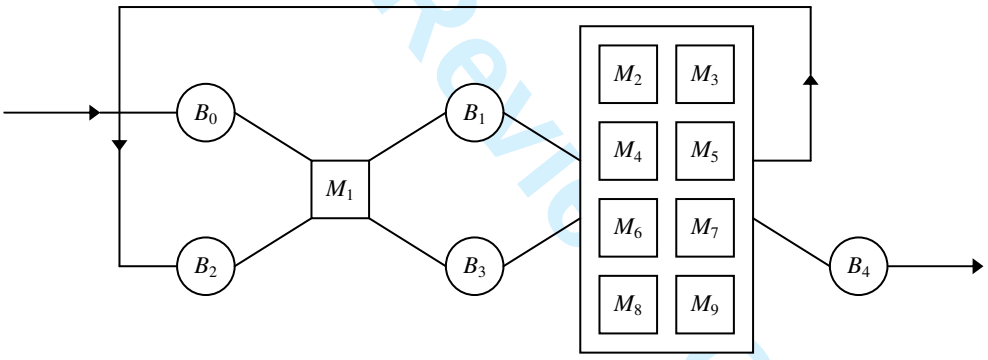

Figure 1: Re-entrant System

2.2 Buffer Selection Sequence

In contrast to the use of Critical Ratio to decide from which buffer a part should be taken (a *dynamic priority* scheme) the CPP employs a *fixed order of priority* to examine buffers. Here all buffers upstream of a machine are examined in the sequence dictated by the fixed order of priority and this order is never altered. If the buffer highest in the sequence is not empty then a part from that buffer will be loaded onto the machine. If it is empty then the buffer next in the sequence is checked and so on. A common preference is to assign a buffer selection sequence such that buffers ahead of a given machine that hold parts closer to their final operations receive higher priority. Often, the aim of this strategy is to decrease the inventory in the system in the hope that reduced congestion will improve product flow. Another possible reason is that the cost of holding inventory may be higher toward the end of the system, in which case reducing the level of a buffer further downstream would be more cost efficient. This is often the case since

parts are considered as having value added to them during production. Holding costs, however, are not of concern for these comparisons and are viewed as constant throughout the system. In particular, Gzouli (2000) showed that giving priority to the buffer furthest downstream gave the best results for this system.

2.3 Hedging Times

In addition to blocking before service and establishing a fixed buffer selection sequence, the key notion needed to define the CPP is that of a hedging time. Hedging times are assigned to a machine, one for each of it upstream buffers, and dictate when the machine is allowed to remove parts from those buffers. For example, in the system illustrated in Figure 1, M_1 has two hedging times associated with it: H_0 and H_2. H_0 is used to prohibit the removal of a part from B_0 until the current time, t, is 'close enough' to (i.e., within H_0 time units of) the part's due date. In other words, M_1 is not permitted to take a part from B_0 unless Due Date of Part in B_0 - $t \leq H_0$. Similarly, M_1 is not allowed to take a part from B_2 unless Due Date of Part in B_2 - $t \leq H_2$.

In the case of machines M_2 to M_9 there are two hedging times, H_1 and H_3, assigned to the work centre as a whole. If the due date of a part in B_1 is such that Due Date of Part in B_1 - $t \leq H_1$ then the part is removed from B_1 and placed in any one of the machines that happens to be available. If not, then the part remains in B_1.

If none of the parts in the highest priority upstream buffer meet these criteria, then the next highest priority buffer is checked and so on. If none of the parts in any of the upstream buffers are close enough to their due dates to be given to M_1 then the machine is forced to remain idle. The hedging times in the CPP effectively form time-based release policies for every buffer in the system and forcing machines, in this way, to remain idle unless they need to be working on parts plays an important role in the performance of this policy.

There are several reasons why hedging times should be used to prevent parts from being operated on, even when a machine would be idle, i.e.:

a) they ensure that downstream machines remain available if more important or urgent parts are coming along, and
b) they enable inventory costs to be reduced, especially when parts that are more highly processed are assigned greater value.

3. Description of system used to compare CPP and CR

The work system, Figure 2, used to compare CPP with Critical Ratio contains 6 buffers in which parts might accumulate. The re-entrant nature of the system is such that a 'self-regulatory' effect in terms of preventing the build-up of inventory is not likely to occur, ie the sequential nature of the system results in any one machine being less directly influenced by events at another. Different processing times and repair probabilities are assigned to different machines in an attempt to introduce further imbalance in the system.

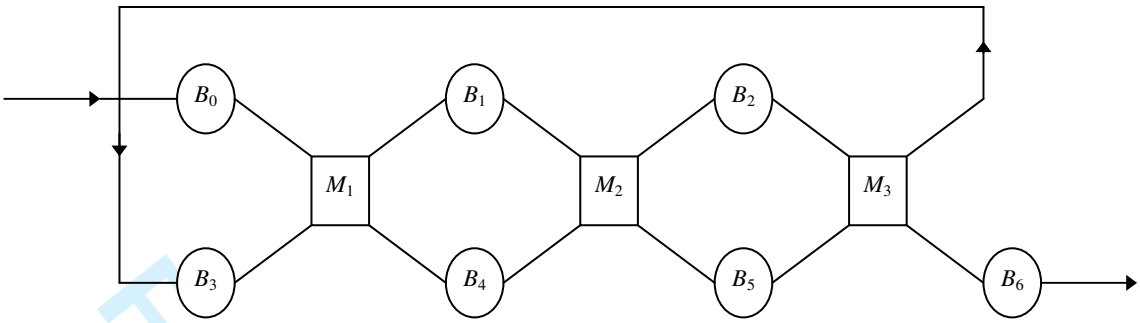

Figure 2: A simple re-entrant system comprising three unreliable machines

The system is also a make-to-order system and, in such environments where parts are assigned due dates, it is appropriate to use policies based on time rather than WIP. For example, it may be beneficial not to release a part into the system until the difference between its due date and the current time is below a certain value. This value should be chosen such that there is a degree of confidence of being able to manufacture the part within that period of time. Wein (1988) has shown that the use of a release policy at the entry point, i.e. termed the 'entry policy' of the system, is of particular importance.

The way in which due dates are assigned represents an unpredictable and variable customer demand process so the urgency to manufacture parts will vary greatly among customers. Scheduling techniques which aim to minimise job tardiness are particularly well suited to this kind of environment. Critical Ratio, a method within this class is expected to perform well under the conditions prevailing.

A key feature of the CPP is that it includes both real time scheduling and release policies within its structure. Critical Ratio is solely a method for scheduling so comparing the performance of Critical Ratio without an entry policy against the performance of the CPP would be unfair. Therefore, in what follows, Critical Ratio has been implemented with the addition of the same entry policy inherent in the CPP. This entry policy appears in the form of the *hedging time* of the first machine.

3.1 Material Flow and Machine Characteristics

Since a re-entrant system is being considered, for any given machine, there is more than one upstream buffer. The raw materials buffer, B_0, has infinite capacity and N_i, the size of buffer B_i, is finite for all $i = 1,\ldots,6$. Parts travel through the system in the sequence:

$$B_0 - M_1 - B_1 - M_2 - B_2 - M_3 - B_3 - M_1 - B_4 - M_2 - B_5 - M_3$$

and finally to the finished goods buffer, B_6 until its *due date* is reached, i.e., until the current time, t, is equal to the due date of the part. If, on arrival at B_6, the due date of a part is less than t the part is removed immediately.

Machines M_1, M_2 and M_3 have processing times of:

$$\tau_1 = 1 + \varepsilon_1$$
$$\tau_2 = 0.95 + \varepsilon_2$$
$$\tau_3 = 1 + \varepsilon_3$$

where the error $\varepsilon_i \in$ [-0.02, 0.02] and is biased towards 0 for i = 1,2,3, ie $\varepsilon_i = 0.08|j - 0.5|(j - 0.5)$ where j is chosen at random from [0, 1]. The probabilities of machines failing, (i.e. breaking down), on receiving parts are $p_i = 0.005$ for i = 1,2,3 and repairs are performed according to geometric distributions with parameters r_i, where:

$$r_1 = 0.035, \qquad r_2 = 0.035 \qquad \text{and} \qquad r_3 = 0.04.$$

The fact that failures can only occur when machines receive parts places them in the category of *operation dependent failures* as opposed to *time dependent failures* (Buzacott and Hanifin, 1978). These values result in, for example M_3 failing, on average, once in every $1/p_3$ = 200 operations with an average repair time of $1/r_3$ = 25 time units.

It can be shown that the *isolated production rates* are $r_1/\bar{\tau}_1(r_1 + p_1)$ for M_i for i = 1, 2, 3 (Gershwin, 1994) where $\bar{\tau}_i$ is the average value of τ_i. The *expected isolated processing times* are the inverses of the isolated production rates; for example $\bar{\tau}_1(r_1 + p_1)/r_1$ for M_1. The machines are, however, not isolated and when the buffers in the system are not of infinite capacity the expected processing times will be larger than when in isolation.

3.2 Inter-Arrival Times and Due Dates

Parts enter B_0 at an average rate of 1 every Takt time units. The inter-arrival time is ε_{Takt} time units, (the unit of time is arbitrary in all that follows so shall frequently be omitted), where $\varepsilon_{Takt} \in [0, 2\times Takt]$ and is biased towards the centre of the interval, ie the arrival time of a part is given by adding ε_{Takt} to the arrival time of the previous part as follows:

$\varepsilon_{Takt} = Takt\{1 + 4|j - 0.5|(j - 0.5)\}$; j is chosen at random from a uniform [0, 1] distribution.

In this case, $Takt$ = 3. This arrival process ensures that the average inter-arrival time, over a long simulation period, is equal to $Takt$ but permits us the possibility of periods in which arrivals occur in quick succession or with lower than average frequency. On arrival at B_0 a part is given a due date. In order to form a customer demand process with high variability, parts are assigned due dates according to:

$$\text{Due Date} = \text{Arrival Time} + \text{Average Customer Lead Time} + \varepsilon_{CustLeadTime}$$

where:

$$\varepsilon_{CustLeadTime} \in [-\{\text{AvgCustLeadTime} - 4(\sum_{i=1}^{3}\frac{\bar{\tau}_i(r_i + p_i)}{r_i})\},\{\text{AvgCustLeadTime} - 4(\sum_{i=1}^{3}\frac{\bar{\tau}_i(r_i + p_i)}{r_i})\}]$$

and is biased towards 0, ie $\varepsilon_{CustLeadTime} = 4|j - 0.5|(j - 0.5)\{\text{AvgCustLeadTime} - 4(\sum_{i=1}^{3}\frac{\bar{\tau}_i(r_i + p_i)}{r_i})\}$.

This definition of $\varepsilon_{CustLeadTime}$ results in an average difference between the due date of a part and its arrival time of:

Average(Due Date - Arrival Time) = Average Customer Lead Time

In this case, Average Customer Lead Time = 130. This serves to ensure that:

Average(Due Date - Arrival Time) = Average Customer Lead Time

and that the smallest possible customer lead time is $4(\sum_{i=1}^{3}\frac{\bar{\tau}_i(r_i + p_i)}{r_i})$.

As stated $\bar{\tau}_i(r_i + p_i)/r_i$ are the expected isolated processing times for each machine. Therefore, this covers the 'worst case scenario' where, if priority is given to a part at the raw materials buffer, there should (if the buffers are sufficiently large) be enough time available for processing and enough slack should the part have to wait for machines to become available at each stage of production. As the sizes of the buffers in the system decrease, this claim will become less justifiable since the expected isolated processing times of the machines will become larger.

There are two main differences in this system from that used by Gzouli:

1. The system is more erratic or vulnerable to random events. This is due to:
   a. larger variation in both inter-arrival time and due date allocation, and
   b. a lower *Takt* time which in turn, results in higher machine utilisation.

   Although increased instability would seem an undesirable property for a manufacturing system to have, it enables the benefits of one production control policy over another to be seen more clearly; if a manufacturing system is congested and subject to disruptions the control policy will have more 'controlling' to do.

2. This is a 'purely' *make-to-order* system (Schonberger and Knod, 1994). By 'purely' make-to-order we mean that each individual part is ordered by and delivered to a specific customer. For this reason, it is of no use analysing the performance of control policies which deal with safety stocks such as CONWIP, basestock and, to a lesser extent, kanban as in the work by Gzouli. These policies assume that parts stored in the finished goods buffer will be useful for satisfying orders that have become unexpectedly difficult to meet due either to an unusual rise in customer demand or to decreased system productivity. Systems in which this assumption is true are termed *make-to-stock*. The control parameters in make-to-stock policies are based on inventory levels with the view that parts are interchangeable.

The characteristics used to discuss system performance will be average total work-in-process, total buffer space and *service level*. These are calculated at the end of each simulation run as follows:

$$Average\ Total\ WIP = \sum_{i=1}^{4} \bar{n}_i \quad \text{where } \bar{n}_i \text{ is the average value of } n_i, \text{ the level of the buffer } B_i$$

$$Total\ Buffer\ Space = \sum_{i=1}^{4} N_i \quad \text{where } N_i \text{ is the size or capacity of } B_i$$

$$Service\ Level = 1 - \frac{\text{Number of Parts Finished Late + Number of Late Parts Still in the System}}{\text{Number of Parts that have Entered the System}}$$

where a part that was 'finished late' is one that arrived at the finished goods buffer after its due date and a 'late part still in the system' is one whose due date has already expired but that has not yet reached the finished goods buffer.

Essentially, the nature of these characteristics is that a higher service level (benefit) can be achieved with a larger average total WIP or total buffer space (costs). Decreasing the average total WIP or total buffer space (benefits) will result in a lower service level (cost). There is a trade-off, therefore, between WIP, total buffer space and service level which is usually resolved by management constraints (e.g., the service level must be at least 95% or the average total WIP must not exceed 100 parts) or by capacity constraints (e.g., there may only be enough room on the factory floor for 200 units of storage). The trade-off between average total WIP and service level is demonstrated for two hypothetical production control policies S_a and S_b in Figure 3.

Each of the curves in Figure 3 is known as the *convex hull* of a set of data. There are many points above and to the left of these curves in the original data (see Figures 4 to 11). For example, many different system configurations (choices of buffer sizes and control policy parameter values) will result in the same service level. Of those configurations, the point on the convex hull, is that which creates the least WIP. In short, the convex hulls in Figure 3 represent the best system configurations (the best 'WIP to service level' pairs) for policies S_a and S_b. S_a is described as being better than policy S_b since S_a can achieve any given service level with a lower average total WIP than S_b. Conversely, for any given value of average total WIP, S_a can achieve a higher service level than S_b.

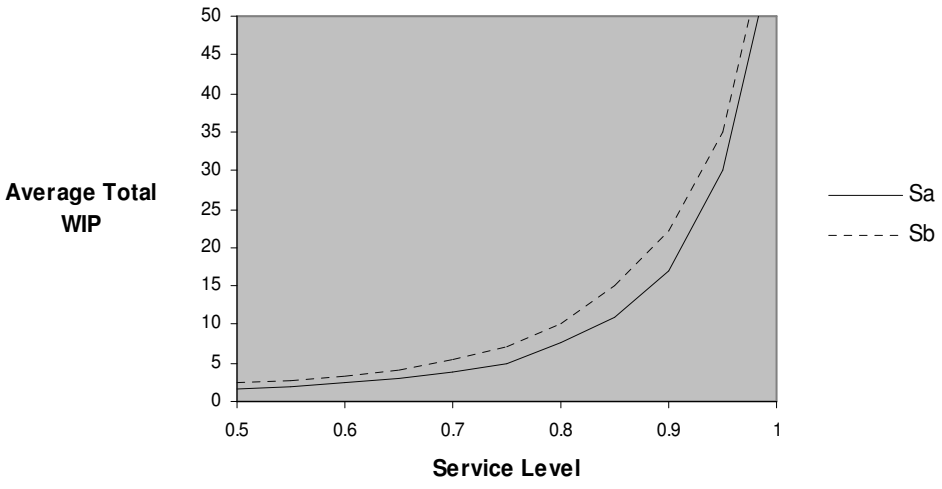

Figure 3: The trade-off between service level and average total WIP

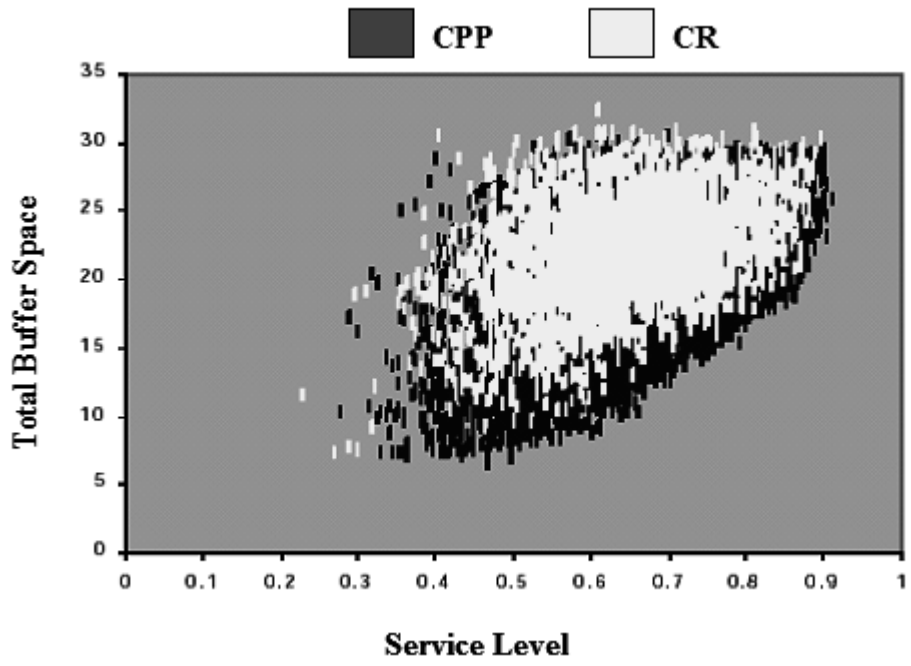

Figure 4: The trade-off between service level and average total WIP (Takt = 3.25, Buffer sizes = 3, 6 or 14)

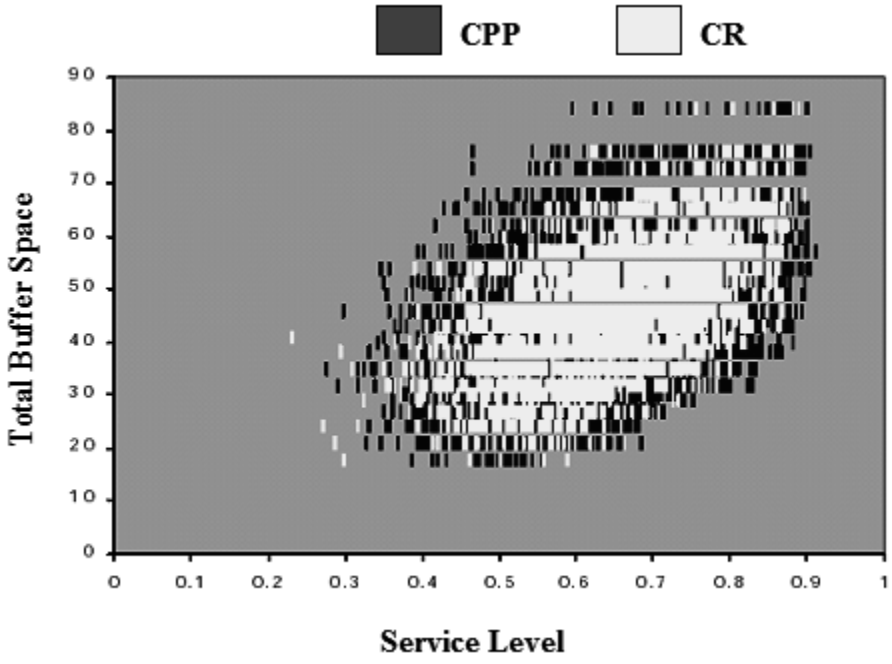

Figure 5: The trade-off between service level and total buffer space (Takt = 3.25, Buffer sizes = 3, 6 or 14)

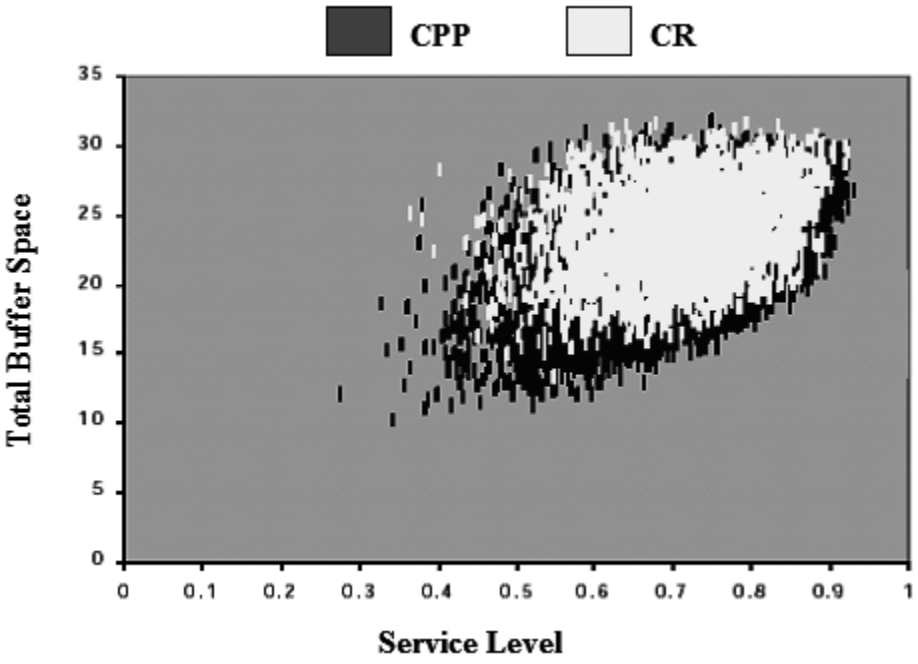

Figure 6: The trade-off between service level and average total WIP (Takt = 3.25, Buffer sizes = 5, 8 or 16)

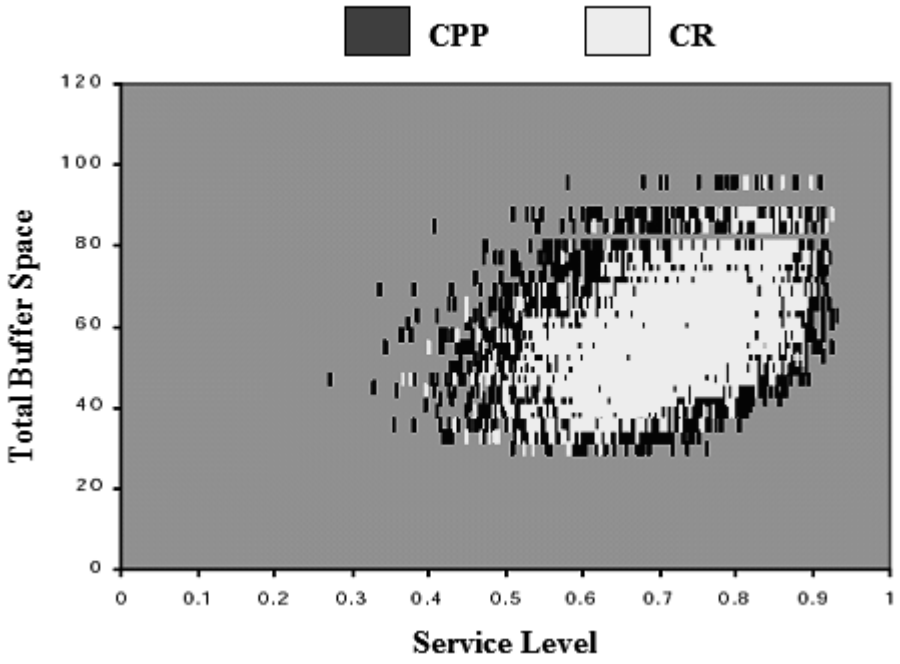

Figure 7: The trade-off between service level and total buffer space (Takt = 3.25, Buffer sizes = 5, 8 or 16)

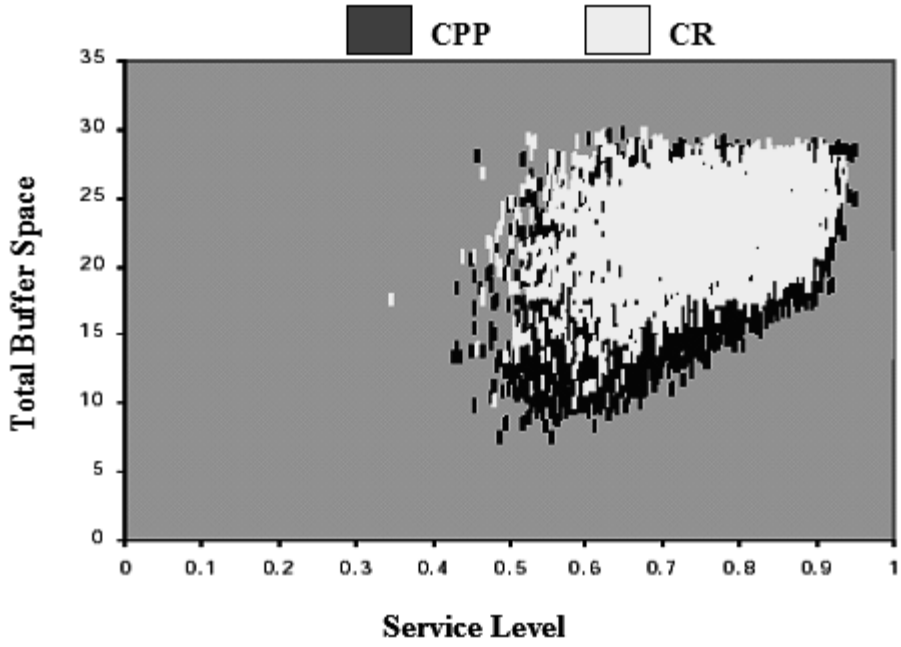

Figure 8: The trade-off between service level and average total WIP (Takt = 3.5, Buffer sizes = 3, 6 or 14)

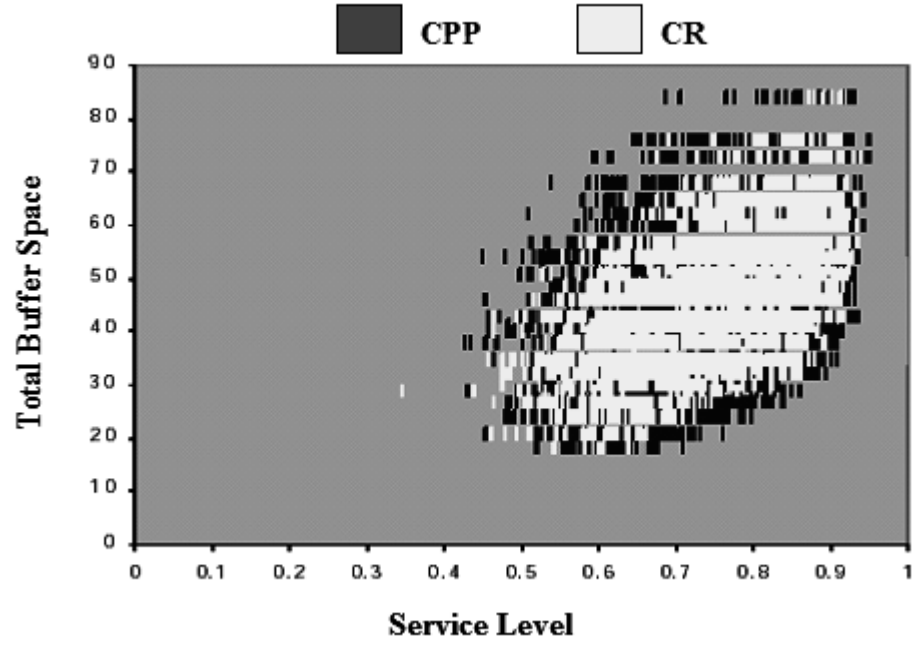

Figure 9: The trade-off between service level and total buffer space (Takt = 3.5, Buffer sizes = 3, 6 or 14)

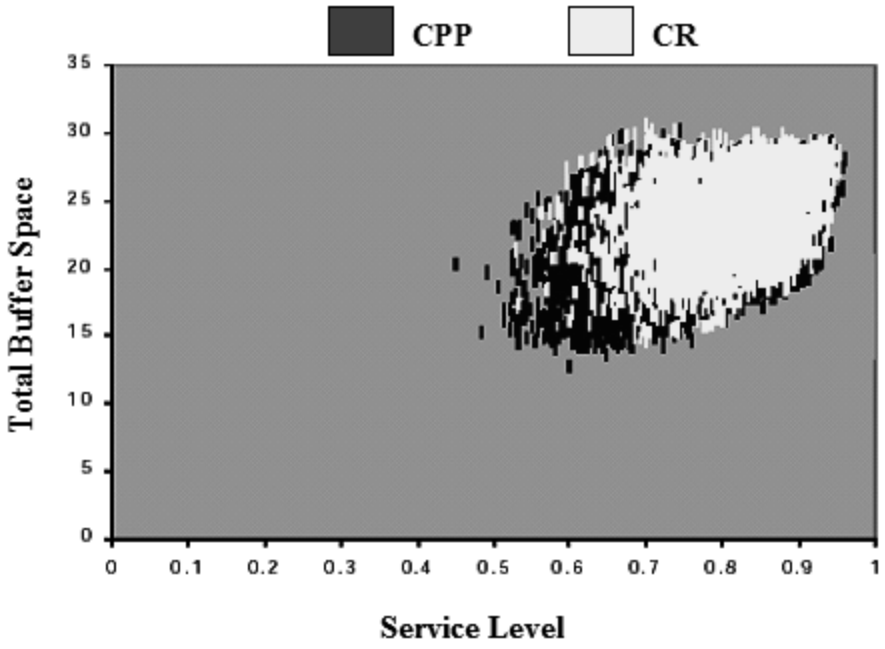

Figure 10: The trade-off between service level and average total WIP (Takt = 3.5, Buffer sizes = 5, 8 or 16)

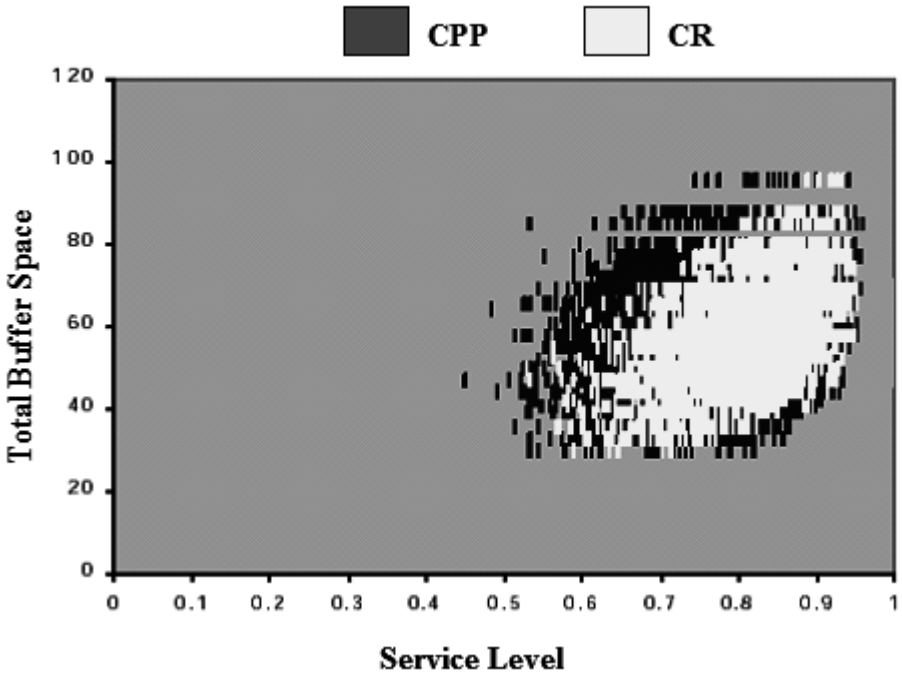

Figure 11: The trade-off between service level and total buffer space (Takt = 3.5, Buffer sizes = 5, 8 or 16)

In both this system and the system examined by Gzouli, the value for the average customer lead time is fairly high when compared to the amount of processing time necessary to

manufacture parts. A large average customer lead time ensures a high service level, provided that the capacity of the system is such that the rate of arrival of raw material is feasible, that demand can be met and that the buffer sizes are sufficiently large. A high service level is necessary under these favourable conditions for meaningful comparison of the different control policies. This is because measures will be taken to reduce the average total WIP and total buffer space in the system and then observe by how much the service level is diminished. The fact that the service level is diminished means a lenient average customer lead time should be chosen to ensure a high initial service level.

4. Experimentation and Results

Simulations were run over 100,000 time units using both the CPP and Critical Ratio with warm-up periods of 5,000. This warm-up period was chosen by directly observing how service levels changed with time throughout a range of simulation experiments. Simulations were carried out using two values for the *Takt* time; 3.25 and 3.5. Different combinations of hedging times and buffer sizes were used for each simulation run. Buffer sizes were allowed to assume small, intermediate or large values. Two sets of results were obtained for each choice of *Takt* time; one set of results where the possible buffer sizes had capacities of 3, 6 and 14 and another where the available buffer sizes were 5, 8 and 16. Hedging times, H_i were permitted to assume values of 95, 105, 115, 125 or 135 such that $H_0 \geq H_1 \geq H_2 \geq H_3 \geq H_4 \geq H_5$. Averages of 5 sets of results were taken for each choice of parameters. This resulted in a total of 18,225 simulations for each set of results using the CPP and 3,645 using Critical Ratio. The results, shown in Figures 4 to 11, therefore, required 364,500 and 72,900 simulations, respectively. Simulations were checked at t = 10,000; if the service level at that time was less than 0.7 the run was stopped and discarded.

These results clearly show that the CPP performs better, albeit only slightly, than Critical Ratio on this system. It is important to note that, of the two sets of results where *Takt* = 3.25, namely Figures 4 and 7, the distinction is more pronounced in Figures 4 and 5, i.e., the set with the smaller buffer sizes. This is also true for the results where *Takt* = 3.5, i.e., the difference is clearer in Figures 8 and 9 than in Figures 10 and 11. This strongly supports the view that the benefits of using the CPP become more evident when the system is required to run with small buffer sizes.

Also, the difference between the CPP and Critical Ratio is larger when *Takt* = 3.25 than when *Takt* = 3.5. This upholds the argument that the advantages, offered by the CPP, in controlling a production system are better exploited when the system is under greater pressure. An important observation must, however, be made on this point. The benefits of allowing machines to remain idle are not realised when the *Takt* time is large and the system is under little pressure. This is because buffers are not likely to reach their capacities on a frequent basis. The advantages are more readily seen when the *Takt* time is reduced, as is shown by Figures 4 to 11. However, this is only true up to a certain point. If the *Takt* time is further decreased so that the system is under a great deal of pressure then there will simply not be enough capacity to afford the luxury of idle time. There seems to be a band of values for the *Takt* time in which it is beneficial to allow machines to remain idle. Developing a method for identifying this range of values for the *Takt* time is a useful topic for further work.

5. Discussion and Conclusions

The intended advantage of using the Critical Ratio over the CPP is to quantify how urgent it is to get parts through the system in order to assist part selection. Fixed buffer selection sequences do not make use of this information. However, the use of a hedging time by CPP attempts to ensure that a part is not loaded onto a machine until it is 'sufficiently urgent' to do so. In other words, not only does the CPP aim to move parts through the system such that they do not arrive at machines late (as with other scheduling policies) but also such that they do not arrive too early. In addition, if the scheduling of parts is done well, questions regarding which part is most urgent should not be necessary; parts should arrive at each stage of production at the appropriate time. In particular, the use of an entry policy to ensure that a part is introduced into the manufacturing system at the correct time can be very influential with regards to successful scheduling (Wein, 1988).

The results of the simulation experiments provided in Figures 4 to 11 show that the recently developed Control Point Policy (Gershwin 1999, 2000) performs better, in terms of the service level, than the popular Critical Ratio technique on a simple, re-entrant production systems. The results also show that the benefits of the CPP can be seen more clearly in a make-to-order environment and in environments where the assignment of due dates to parts results in a high level of re-sequencing by the chosen scheduling policy. Critical Ratio specifically aims to perform this re-sequencing successfully by making use of any available due date information. The CPP uses hedging times to reorder parts just as effectively despite its use of fixed buffer selection sequences. In short, more is demanded from a scheduling or control policy when the system is under greater pressure and when parts need frequent re-sequencing. Under such circumstances, the qualities of a policy will be exposed. It has been shown that, under these exact circumstances, the qualities of the CPP *are* exposed.

In addition, the CPP outperforms Critical Ratio by allowing machines to remain idle. The advantage of this strategy becomes more evident when buffer sizes and the *Takt* time are small. It has been noted, however, that the strategy of allowing machines to remain idle is only appropriate for a range of values of the *Takt* time.

Despite the numerous and lengthy simulations it must be pointed out that the results serve only as an indication of the performances that can be achieved using the CPP and Critical Ratio. This is because the buffer sizes and hedging times were only permitted to assume certain values. However, the simulation experiments performed paid particular attention to cases with small buffer sizes and *Takt* times; characteristics typical of flexible manpower lines. The results indicate that the CPP lends itself well to these situations and, as such, provides an ideal candidate for the control of a manufacturing system type that is growing in popularity, ie multi-part type flexible manpower lines.

6. Acknowledgements

The authors wish to thank the UK's Engineering and Physical Science Council for sponsoring this research work through its Innovative Manufacturing Initiative (EPSRC Grant No. GR/M58818).

7. References

[1] Berkeley, B.J., 1992, A review of the kanban production control research literature, *Production and Operations Management*, vol. 1, no. 4, 393-411.

[2] Bonvik, A.M, Couch, C.E. and Gershwin, S.B., 1997, A comparison of production-line control mechanisms, *International Journal of Production Research*, vol. 35, no. 3, 789-804.

[3] Buzacott, J.A. and Shanthikumar, J.G., 1993, *Stochastic Models of Manufacturing Systems*, Prentice Hall, 1993.
[4] Clark, A.J. and Scarf, H., 1960, Optimal policies for the multi-echelon inventory problem, *Management Science*, vol. 6, no. 4, 475-490.
[5] Gershwin, S.B., 1994, *Manufacturing Systems Engineering*, Prentice Hall.
[6] Gershwin, S.B., 1999, System analysis, design and control: Unification and decomposition, *Second Aegean International Conference on Modelling of Manufacturing Systems,* Tinos Island, Greece.
[7] Gershwin, S.B., 2000, ``Design and Operation of Manufacturing Systems --- The Control-Point Policy," *IIE Transactions*, Volume 32, Number 2, pp. 891-906.
[8] Gzouli, O., 2000, *Comparison of Scheduling Policies by Simulation*, M.Sc. thesis, Massachusetts Institute of Technology.
[9] Kimball, G., 1988, General principles of inventory control, *Journal of Manufacturing and Operations Management*, vol. 1, no. 1, 119-130.
[10] Mitra, D. and Mitrani, I., 1990, Analysis of a kanban discipline for cell co-ordination in production lines, Part I, *Management Science*, vol. 36, no. 12, 1548-1566.
[11] Monden, Y., 1993, *Toyota Production System: An integrated approach to Just-In-Time*, 2nd edition, Industrial Engineering and Management Press, Norcross, GA.
[12] Ohno, T., 1998, *Toyota Production System*, Productivity Press, Cambridge, MA.
[13]Schonberger, R.J. and Knod, E.M., 1994, *Operations Management: Continuous Improvement,* 5th Edition, IRWIN Inc.
[14] Spearman, M.L., Woodruff, D.L. and Hopp, W.J., 1990, CONWIP: a pull alternative to kanban, *International Journal of Production Research*, vol. 18. no. 2, 245-257.
[15] Tabe, T., Murumatsu, R. and Tanaka, Y., 1980, Analysis of production ordering quantities and inventory variations in a multi-stage production ordering system, *International Journal of Production Research*, vol. 18, no. 2, 245-257.
[16] Wein, L.M., 1988, Scheduling semiconductor wafer fabrication, *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, no. 3, 115-130.
[17]Yong, M. S., 2001, “Simulation of real-time scheduling policies in multi-product, make-to-order semiconductor fabrication facilities”, MSc thesis, Massachusetts Institute of Technology.