CONWIP: a pull alternative to kanban

MARK L. SPEARMAN†, DAVID L. WOODRUFF†
and WALLACE J. HOPP†

This paper describes a new pull-based production system called CONWIP. Practical advantages of CONWIP over push and other pull systems are given. Theoretical arguments in favour of the system are outlined and simulation studies are included to give insight into the system's performance.

1. Introduction

Effective production control systems are those that produce the right parts, at the right time, at a competitive cost. Some manufacturers, particularly Japanese firms, have reported considerable success meeting these objectives by using 'pull based' production planning and control systems. These systems are often referred to as just-in-time (JIT), kanban, or zero inventory (ZI). Successful implementations of such systems have greatly reduced both inventory levels and lead times. Unfortunately, kanban is not applicable to many production environments. As Hall (1981) points out, 'Kanban is intrinsically a system for repetitive manufacturing. It will not work in a shop controlled by job orders'. For this reason, many applications of the older and arguably less effective material requirements planning (MRP) approach remain.

In this paper we describe a new pull-based production system called CONWIP. CONWIP appears to share the benefits of kanban (e.g., shorter flow times and reduced inventory levels) while being applicable to a wider variety of production environments.

In the next sections we first define 'push' and 'pull' as they relate to production control. We then discuss some of the reasons for the superiority of pull when it can be applied. CONWIP is defined operationally in Section 4. In Section 5 we argue why CONWIP should outperform push systems and how it is more generally applicable than kanban. In this section we also relate CONWIP to the 'drum-buffer-rope' concepts of Goldratt and Fox. We provide results of a simulation study comparing CONWIP and a push system in Section 6. In Section 7 we describe extensions of CONWIP and conclude in Section 8.

2. Push and pull

It is currently popular to divide production control systems into those which 'push' and those which 'pull', even though there are no generally accepted definitions for these terms. For our purposes, push systems will be those where production jobs are scheduled. Pull systems, on the other hand, are those where the start of one job is triggered by the completion of another.

Before discussing these distinctions further, let us make clear the difference between flow time and lead time. Flow time is the time between job release and its completion.

Revision received September 1989.
† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, USA.

0020-7543/90 $3.00 © 1990 Taylor & Francis Ltd.
Flow time is also known as cycle time in some production circles and as sojourn time in queuing. Flow times are typically random. Lead time is a constant used for planning purposes. Service level is defined as the fraction of jobs whose flow time is not greater than their lead time.

Push systems are very popular in the form of materials requirements planning (MRP) and its successor manufacturing resources planning (MRP II). MRP is described in detail in many places, see for example Vollmann et al. (1984). Such systems have been successful in reducing inventory levels and improving customer service levels in many instances. In 1984 Kanet described one such successful implementation at Black and Decker. However, later in 1988 Kanet described a series of problems to be found in MRP and MRP II systems. Briefly, these were: (1) MRP does not always generate feasible plans and this infeasibility is not detected until too late and (2) MRP uses fixed lead times for offsetting that do not depend on capacity utilization. Graves (1988) describes the basic MRP paradigm as a black box model of the manufacturing process in which a job starting at time \( t \) will finish at time \( t + l \), where \( l \) is a known and unchanging lead time. Because of inherent randomness and the need to anticipate congested conditions, this lead time is usually pessimistic. Theoretically most jobs should finish early and then wait in finished goods inventory. However, in most instances, dispatching is performed for 'hot' jobs resulting in highly variable flow times and large WIP and FGI levels.

A pull system, on the other hand, does not schedule the start of jobs but instead, authorizes production. The best known pull system is kanban. Use of parts as components by downstream work centres authorizes the start of the production of more components. If the components have components themselves, then their use triggers their production at upstream work centres and so on. Transmission of the authorization to produced and move parts is done by passing cards between work centres. 'Kanban' is Japanese for the 'card' which is used to authorize production or movement of parts. The number of cards determines the WIP levels in the plant. The analogy to a liquid flow is that removal of parts from the 'end' of the plant pulls component parts forward through the production system.

Kanban's operation and merits have been described many times, see for example Hall (1981), Kimura and Terada (1981) and Schonberger (1986). Unfortunately kanban, a production control system, is often confused with JIT, a manufacturing philosophy. The JIT philosophy encompasses not only kanban but also total quality control, set-up reduction, and worker participation. Among the basic advantages of JIT are reduced WIP levels and shorter flow times allowing for lower production costs and greater customer responsiveness. Some of these improvements appear to be due to the JIT environment while others may be attributable more directly to kanban.

In this paper we are concerned primarily with only the production control aspects of JIT. Some authors describe kanban as behaving like a siphon pipe with parts flowing through a facility. It is really more accurate to say that kanban requires that the production process behave as a nearly frictionless pipe. Scrap loss, significant set-ups and expediting are not tolerated.

Push and pull are not mutually exclusive approaches. For instance, MRP and kanban can be combined. An interesting example of this is provided by synchro-MRP which is described by Hall (1981). In this system, work is scheduled by the MRP system but cannot be started without a kanban as authorization as well. As we shall see, this system is similar in some ways to CONWIP and shares some of its advantages.
3. Superiority of pull systems

MRP is generally considered to be applicable to many more manufacturing firms than is kanban. But kanban seems to produce superior results when it can be applied. We divide the reasons for this superiority into environmental, queueing, and control effects.

3.1. Environmental effects

Kanban is often cited as being applicable only in certain environments (see e.g. Hall 1981). In some cases the environment can be shaped to conform with kanban's needs; in other cases this is not possible so kanban cannot be used. Kanban is difficult, or impossible to use when there are (see e.g., Monden 1983, p. 64):

1. job orders with short production runs, or
2. significant set-ups, or
3. scrap loss, or
4. large, unpredictable fluctuations in demand.

MRP, on the other hand, can be used in almost any discrete part production environment.

The fact that kanban forces the shaping of the production environment can be an important benefit. One of the key tenets of kanban practitioners is that the number of kanbans be inexorably reduced. This reduces WIP and makes any environmental problems more noticeable. Proponents of kanban are fond of an analogy between WIP in a plant full of environmental problems and water in a stream full of rocks. Reducing the WIP shows production problems just as reducing the water level shows the rocks in a stream.

Extensive simulation studies by Krejewski et al. (1987) indicate that environmental considerations may be the main cause for differences in the performance of push and pull production systems. A significant portion of the reason for kanban's apparent superiority over push systems may be its requirement for, and facilitation of, environmental improvement. Kanban's reputation may also be enhanced by the fact that it is typically used in environments, where there are not serious environmental impediments to successful production management.

3.2. Queueing effects

We can gain insight into some of the advantages of pull over push systems by modelling a push system as an open queueing network and a pull system as a closed queueing network. These models are appropriate since push systems schedule throughput and measure WIP (e.g., input/output control), while pull systems set the WIP levels and measure throughput. As we will see later, the closed queueing model is an extremely good representation of a CONWIP production line.

Spearman and Zazanis (1988) compared equivalent closed and open systems composed of exponential machines. These systems are equivalent in that they involve the same machines and throughput of the closed system is set equal to the Poisson input stream in the open system. Their comparisons proved that the closed system has less average WIP at every station than the open system. Consequently, the total system WIP and the average flow time will be less in a pull system than in a push system. Obviously, the conditions for these comparisons to hold exactly (i.e., exponential process times and Poisson arrivals) are not met in most well-managed production facilities. However, many authors (e.g. Karmakar (1987) who cites Solberg (1977) and
Suri (1983)) have suggested that macroscopic results obtained using these assumptions will hold under many realistic conditions. Furthermore, if a work centre processes a wide variety of parts, then the processing times may appear to be nearly exponential due to variability in the production mix rather than variability in the processing itself.

There is also evidence that the variance of flow time will be less in pull system than in an equivalent push system. This stems from the fact that the number of jobs at each work station is negatively correlated in a pull system and not correlated at all in a push system. This notion of 'negative dependence' has been formalized by Whitt (1984). We have also observed this phenomenon in simulation studies.

Reduced flow time means and variances have intrinsic benefits, but perhaps more important is that they can imply smaller WIP and FGI for a given lead time (see Hopp et al. (1988) for further discussion). We can illustrate this concept by considering two production systems whose flow times have equal means and different variances (see Fig. 1). We see that if the lead time is fifteen days, then the lower variance production system will have more jobs that finish before fifteen (and therefore a higher service level). Also, more jobs will finish closer to fifteen and therefore spend less time in FGI. With equal throughput rates this implies less average FGI for the system with smaller variance.

3.3. Control effects

Another advantage of pull over push systems stems from the control of WIP (pull) versus the control of throughput (push). As we shall see, the WIP and throughput fluctuations in a push system result in violations of the assumption that flow times (and therefore lead times) are constant. Also, WIP is inherently easier to optimize than throughput.

To see this, note that Little’s law gives us

\[ \text{Average flow time} = \frac{\text{Average WIP}}{\text{Average throughput}}. \]

which demonstrates the invalidity of the 'black box' assumption regarding flow times. Flow times will not be constant, but will vary with WIP and throughput. For plants

![Figure 1. Density of flow times for two production systems.](image)
operating near capacity, throughput remains nearly constant because it is bounded above. WIP, on the other hand, can grow to dangerously high levels if left unchecked. As WIP grows, so do flow times until some corrective action is taken to reduce the WIP (e.g. overtime). WIP is bounded in a pull system, so this problem is avoided.

Beyond providing a check on WIP growth, pull systems are inherently easier to control for two reasons. First, WIP is directly observable while capacity which is needed to appropriately release work in a push system, must be estimated. Because capacities depend on numerous efficiency factors and may fluctuate with changes in product mix, such estimates are difficult to make. Second, there is evidence that WIP is a more robust control that throughput (see Spearman and Zazanis (1988) for the necessary conditions and a proof). In other words, errors in setting WIP levels will degrade the performance of pull systems less than errors in estimating capacity will hurt the performance of push systems. We will observe corroborating evidence in the simulation study presented in Section 6.

4. CONWIP

Our goal is to develop a system that possesses the benefits of a pull system and can be used in a wide variety of manufacturing environments. The CONWIP (CONstant Work In Process) approach offers the promise of achieving this goal.

4.1. Operation

CONWIP is described here for a single production line. Extension to control of multiple lines is discussed in Section 7. Our description of CONWIP assumes that parts are moved in standard containers, each of which contains roughly the same amount of 'work'. That is, the total process time at the bottleneck for each container is approximately the same.

CONWIP is a generalized form of kanban. Like kanban, it relies on signals. Although these signals could be electronic, we will refer to them as cards for the remainder of this paper. In a CONWIP system, the cards traverse a circuit that includes the entire production line. A card is attached to a standard container of parts at the beginning of the line. When the container is used at the end of the line, the card is removed and sent back to the beginning where it waits in a card queue to eventually be attached to another container of parts.

In a kanban system, each card is used to signal production of a specific part. CONWIP production cards are assigned to the production line and are not part number specific. Part numbers are assigned to the cards at the beginning of the production line. The numbers are matched with the cards by referencing a backlog list. When work is needed for the first process centre in the production line, the card is removed from the queue and marked with the first part number in the backlog for which raw materials (or components) are present. The time of the part number match is also noted on the card as the system entry time. Fig. 2 illustrates the operation of a CONWIP system.

Maintenance of the backlog is the responsibility of production and inventory control staff. In many cases the backlog will be generated from a master production schedule. In other cases, firm orders may be added to the backlog as they are received. Expeditors are allowed to rearrange the backlog and/or add part numbers to it. Under no circumstances are expeditors allowed to force the start of work without a card present, even if the first process centre in the production lines is idle.
The queue discipline used at all process centres in the line is 'first in system first served' (FSFS). In other words, work with the lowest system entry time is started first. The only exception is rework, which is given the highest priority.

Since pull systems set the WIP level and measure throughput, there may need to be some mechanism to ensure 'correct' throughput levels. Of course, it is also possible to let the line run as fast as it can. This may be a viable option if the line produces products whose demand exceeds the maximum capacity of the line. If the line can't be allowed to run unchecked, then the control parameters must be established which determine a target level of production and the point at which actions are taken if production is expected to be above or below the target.

4.2. Parameters

The parameters to be established for a CONWIP line are:

- \( m \), the card count. This determines the maximum WIP level for the line.
- \( q \), the production quota. This is the target production quantity for a period.
- \( n \), the maximum work ahead amount. If \( q + n \) is produced during a period, the line is stopped until the start of the next period.
- \( r \), a capacity shortage trigger. This is a function of the actual production up to sometime \( t \), \( A(t) \). The trigger function indicates that additional capacity must be utilized (e.g., overtime must be scheduled). An example of a simple trigger function is a constant allowable shortfall. In that case capacity additions are triggered at the end of the production period of length \( T \) if \( A(T) < q - r \). A more sophisticated trigger function would involve the use of the probability distribution of \( A(T) \) given \( A(t) \) and the state of the production line (e.g., machine failure status).

Optimization of these parameters is the subject of ongoing research. Finding the optimal values will clearly involve economic trade-offs. Increasing \( n \) and/or \( m \) will tend to increase service levels at the expense of higher inventories. Increasing \( q \) will tend to increase expected revenues while decreasing service. For a given \( q \), the trigger function helps determine the service level and the costs associated with capacity additions (e.g., overtime costs can be balanced with service level). A cautious approach to CONWIP implementation would be the same as that offered by kanban practitioners: set the inventory parameters \( m \) and \( n \) higher than seems to be optimal and gradually lower them over time.
4.3. Discussion

In some ways, CONWIP may be considered to be input-output control carried to its logical extreme. In Wight (1970) states that (THE INPUT TO A SHOP MUST BE EQUAL TO THE OUTPUT) (emphasis his). His paper discusses implementing input/output control by the production control function. CONWIP provides a practical method of implementing input/output control at the shop floor level.

CONWIP is also similar to a technique used in air traffic control. On days with heavy air traffic, a departing plane will sometimes be held on the ground at the originating airport rather than be allowed to take off and remain in a holding pattern at the congested destination airport. Because the object is to avoid delays at the destination airport (in the air), planes are held even if take-off runways are free at the originating airport. The result is greater safety and lower fuel consumption with no added delay.

Under CONWIP, a job will not be started unless a place in the system has been vacated for it. The same balancing act that air traffic controllers face must be addressed by user of CONWIP. Enough jobs (planes) must be placed in the line (en route to arrive) so that the bottleneck station (the runway) is seldom idle but not so much that jobs (planes) wait a great deal of time. If this is done well, then the system will attain maximum throughput without excessive flow time or WIP.

CONWIP will naturally also comply with Goldratt's admonition to 'balance the flow, not the capacity' (Goldratt 1986). The operation of a CONWIP line is regulated by the bottleneck resource. Its utilization determines the capacity of the line, while all other resources must periodically be idle (this is often cited in arguments against the use of machine utilization as a performance measure). Assuming there is sufficient demand for the output of the line, a CONWIP system with the correct number of cards will maintain just enough WIP to keep the bottleneck busy. If work begins to pile up behind the bottleneck, then cards will not be carried to the end of the line and new work will not be started. On the other hand, if the bottleneck is finishing work very quickly, then cards will be recycled quickly.

5. Comparisons

In this section, we compare CONWIP with kanban and with push based production control of a single production line. CONWIP differs from kanban in three main ways:

1. use of a backlog to dictate the part number sequence,
2. cards are associated with all parts produced on a line rather than individual part numbers, and
3. jobs are pushed between workstations in series once they have authorized by a card to start at the beginning of the line.

The differences between CONWIP and push control stem largely from the built in feedback of the CONWIP system.

5.1. CONWIP more general than kanban

All manufacturing environments are not well-suited to the use of kanban. In particular, production lines which produce many different parts face serious practical problems. There simply isn't enough room to have a standard container of each part number present, and even if there were, WIP levels would be higher than necessary.
CONWIP can solve this problem in many cases because the backlog allows explicit control over which parts are produced and in which sequence. Under CONWIP, WIP is not maintained for each part number.

In production environments where it is not economically feasible to eliminate significant set-ups, kanban is generally thought to be inappropriate (Hall 1983). CONWIP, on the other hand, allows explicit sequencing of part production so set-ups can be incorporated in the planning process.

The use of a backlog by CONWIP allows sequencing of jobs to be done by production control personnel when appropriate. This is in contrast to kanban where the sequencing is done on the shop floor. Sequences may need to be controlled when jobs have different priorities.

There are theoretical reasons to believe that CONWIP will result in lower WIP levels than a kanban system with the same throughput (Spearman and Zazanis 1988). These are easy to see for a system with a distinct bottleneck. In a kanban system, there will be generally be WIP at process centres upstream from the bottleneck at all times. In a CONWIP system, WIP will tend to collect at the bottleneck. Hence, CONWIP will tend to produce higher utilization of the bottleneck, and therefore greater throughput than kanban.

For the purpose of controlling a single production line, kanban and CONWIP share many characteristics. Many of the environmental and psychological effects of kanban will be present in a CONWIP system. Most of these effects can be attributed to WIP control. Keeping WIP low has the following beneficial effects:

1. The chances for early detection of quality problems are improved. If WIP levels are lower, so are flow times. If a process is producing defective items, the first one reaches a subsequent operation where the defect will be noticed before too many other defects have been produced.

2. There is simply less clutter on the shop floor. When WIP levels are low, operators waste less time searching through WIP for the next job to process. The chances for damage and mishaps are also decreased.

3. A 'tight ship' mentality is promoted. Reduced WIP makes it harder to cover up or tolerate: machine failures, defects, yield losses, theft, and unnecessary idle time.

In short, if kanban is good, CONWIP is better, since it provides the benefits of kanban to a wider variety of situations.

5.2. CONWIP more effective than push

CONWIP is expected to outperform push systems in many production situations for reasons given in Section 3. Those reasons are shared by kanban and other pull based control systems. There are other reasons which are unique to CONWIP.

Effectiveness can be quantitatively evaluated by considering work in process (WIP), finished goods inventory (FGI) levels, and the fraction of jobs that are late (1—service level). This section will compare the performance of CONWIP with push systems using these three measures. It is clear that lower FGI and WIP coupled with higher service level is 'good'.

CONWIP seems to be particularly superior to a push system when it is desirable to run the plant at the highest possible output rates. In practice, it is difficult to operate a plant near full capacity because it is difficult to measure capacity precisely. An
CONWIP: a pull alternative to kanban

advantage of CONWIP over a push system is that if production in excess of capacity is inadvertently scheduled, the pull system leaves planners with more flexibility. The push system will be clogged with WIP when the mistake is discovered (assuming the bottleneck operation is not the release point). On the other hand, the CONWIP system will simply have a long backlog of part numbers to be started. A list of part numbers is much easier to manipulate than WIP.

Of course, MRP II offers something of a solution to this problem for the push system. MRP II contains a capacity check module which warns the scheduler if an attempt is made to schedule more work than the capacity of the line. Unfortunately, the capacity used for comparison must necessarily be an estimated constant, while capacity is really a random variable whose realization depends on the complex interaction of a number of factors. The feedback of cards in CONWIP provides a means of avoiding excess WIP based on the actual performance of the production line. It can be thought of as a ‘natural capacity check’ which prevents the start of excess work when the line is experiencing lower than anticipated capacity. On the other hand, CONWIP can facilitate taking advantage of unusually good luck with respect to capacity by allowing more work to start if jobs are being finished quickly.

CONWIP also appears to alleviate a problem found in many push systems called the ‘overtime vicious cycle’ which unfolds as the following.

(1) Production planning compute the capacity available for production in a given period and report it to sales. This is typically computed as the amount of production time available at the bottleneck process centre less lost production for (a) repairs, (b) operator inefficiency, (c) set-ups, (d) breaks, as well as (e) yield losses and (f) rework.

(2) Sales personnel ‘book’ this capacity by committing to customer orders and thereby establishing the master production schedule (MPS).

(3) Work is released to the plant by the MRP system from the MPS. The average rate of release is equal to the average capacity (if there is sufficient demand).

(4) For some reason or another, the bottleneck centre becomes starved—the one contingency not included in step 1.

(5) Since work continues to be put into the system when the bottleneck was starved, WIP increases.

(6) As WIP increases, flow times increase.

(7) As flow times increase, jobs become late.

(8) The shop floor becomes congested and customers start to complain.

(9) Management authorizes overtime to work down the backlog.

This cycle is often exacerbated by the fact that production planners are often optimistic about available capacity. Under CONWIP, the temptation to book additional work is lessened because overbooking becomes apparent sooner. Since WIP is constant, the flow time is more predictable. It is easy to see when work will not finish on time. When this occurs, overtime will still be necessary, but the condition of the facility will not have deteriorated as in the push system.

A related issue is the scheduling of low priority work. In many cases, a production line is ‘kept busy’ by producing low priority parts. This is intended to make use of capacity not employed in the production of high priority parts. In a push system, the production control staff must use a projected capacity in order to decide if low priority work should be scheduled. This is a problem if the bottleneck operation for the line is not the first operation. Low priority work can be scheduled and processed by the initial
work centres in the line before it is realized that the actual capacity of the bottleneck will be lower than expected. This problem is less severe in a CONWIP line because the arrival of cards to authorize the start of production is controlled by the actual speed of the line.

5.3. CONWIP and 'drum-buffer-ropes'

Although developed along very different reasoning, there are many similarities between CONWIP and the more general construct known as drum-buffer-ropes (DBR) proposed by Goldratt and Fox (1986). DBR is more general than CONWIP in that it can be applied to a pure job-shop environment whereas CONWIP cannot. However, when applied to flow lines, DBR and CONWIP result in similar systems.

Under DBR a 'drumbeat' for the rest of the plant is maintained by sequencing work to be done at the bottleneck operation. The drumbeat is then protected by maintaining a time buffer for parts going to the bottleneck. Non-bottleneck operations are then scheduled to maintain this buffer. Finally, a 'rope' is tied from the bottleneck to material release points to ensure that material is released only at the rate that it is used by the bottleneck thereby preventing increase in inventory. These ropes are represented by 'earliest release times' of raw material.

Under CONWIP, an inventory buffer in front of the bottleneck appears naturally as long as the established WIP level is sufficient. Since the bottleneck is (nearly) always utilized the 'drumbeat' for the line is maintained. Inventory 'ropes' are the consequence of releasing jobs only when another is completed. Thus while a major consideration under DBR is to establish appropriate time buffers that protect the bottleneck, the equivalent decision under CONWIP regards the WIP level to be maintained. For instance if under DBR a time buffer of $T_b$ is to be maintained in a system that has a bottleneck rate of $r_b$ and an average of cycle time of $T$ not counting the time buffer, then the equivalent CONWIP level would be $r_b(T + T_b)$.

The main difference between DBR and CONWIP is that under DBR, decisions must be made regarding the time of work releases, while under CONWIP the releases occur at the bottleneck rate automatically. Thus, given the relative effectiveness of controlling WIP versus controlling release rates (see Spearman and Zazanis (1988) and the simulation results in Section 6) it would appear that CONWIP will be more robust to errors in capacity data. Because the WIP buffer will naturally accumulate in front of the bottleneck, CONWIP will be robust to errors in determining the bottleneck and changes in product mix which move the bottleneck.

It is possible, and may be attractive in certain cases, to modify CONWIP to resemble DBR very closely. When there is a distinct and stable bottleneck with downstream machines subject to long outages, it would be advisable to pull from the bottleneck by making the bottleneck last machine in the CONWIP loop and push work to the downstream stations. This would avoid the problem of cards piling up in front of a failed machine and resulting in starvation of the bottleneck. In this configuration, CONWIP would look very much like DBR, since work authorizations in both cases would emanate from the bottleneck. However, CONWIP would still require a WIP decision in place of release timing decisions.

† This is because the bottleneck is the slowest station in, what is effectively, a tandem closed queueing network. It has been noted (see e.g., Bondi and Whitt (1986)) that in such systems any 'excess' WIP will tend to accumulate in front of the slowest station in the line.
6. Simulation results

We have conducted numerous simulations to make comparisons between
CONWIP and push based systems. It is difficult to get results from a simulation which
can be safely used to make broad generalizations. This is particularly true for pull
systems where many of the benefits are due to easier control. For example, in a steady
state simulation one does not typically observe the 'overtime vicious cycle'. In short, it
is hard to produce a fair simulation that models the behaviour of management. Many
of the problems with push systems (e.g., MRP and MRP II) are really 'management'
problems (see Kanet 1988).

In spite of these difficulties, simulation can offer some insight into the relative
performance of complex systems. Below we discuss the results of a simulation study
that illustrates some of the advantages of CONWIP over a push system.

6.1. The environment

The business environment simulated is a simplified version of a fairly typical
situation. The manufacturing facility performs high profit jobs which are in response to
firm orders which arrive at random. The due date for these class one jobs is determined
by adding an officially quoted lead time to the order date and there is a significant
penalty for tardiness. The facility is designed to have more than enough capacity to
service class one orders on time, which of course, means that there must be excess
capacity. A portion of the excess capacity can be used to produce lower profit items
during periods when the demand for high profit items is low. The due date for class two
jobs is determined by adding a lead time to the date when the order is released. (Since
CONWIP releases work continuously, while the push system has a periodic release
system, the low priority job lead time for the push system is one period greater than for
the CONWIP system, thereby giving the push system approximately a one half period
'head start' on average). The penalty for low priority job tardiness is not severe (hence
the 'head start' is not very important).

The simulation makes use of arbitrary periods that are divided into a hundred time
units. The work centres modelled were assumed to have processing time distributions
that are normal with a coefficient of variation of 0.5; the mean processing times and
number of identical servers in each work centre are listed in Table 1. The costs and
other parameters are shown in Table 2 (remember that the lead time quoted for class

<table>
<thead>
<tr>
<th>Work centre number</th>
<th>Number of servers</th>
<th>Mean server process time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 1. Production line characteristics.
two jobs is longer for push than for CONWIP). The marginal profit and cost parameters are used to compute an objective value (a profit) at the end of each simulation run.

6.2. Methods

The simulation runs were for five hundred periods. In order to reduce transient effects, the first hundred periods were ignored in the analysis. The processing times for each work centre were drawn from a pseudo-random number stream unique to the work centre so that the sample paths followed for push and CONWIP systems would be as similar as possible.

For the push system, the control variable is the release quantity for each period. The CONWIP system is controlled by establishing a work in process (WIP) level. A quota, overtime and work ahead were not used. The controls for both systems were optimized by exhaustive search using three simulation replicates for each value of the control. Each time the simulation was started with a new value of the control, the random number streams were reset so that the comparisons among control values followed very similar sample paths. After optimization, the systems were each simulated to produce ten replicate pairs. Within each pair, the random number streams began with same seeds, but between pairs, the initial seeds were changed.

6.3. Findings

The optimal release quantity for the push system is nineteen jobs, and the optimum WIP level for the CONWIP system is 51 jobs. Figure 3 gives a comparison between push and CONWIP for each of the ten replicate pairs when both systems are controlled optimally. The difference can be summarized as statistically, and somewhat economically, significant. A paired t-test reveals that the difference is significant at better than 0.001; however, the means differ by only 3.2%.

The economic significance of CONWIP can be appreciated by studying Fig. 4. This is a manifestation of the more general controllability result alluded to earlier and presented in Spearman and Zazanis (1988); CONWIP is easier to control than a push system. The x-axis is scaled as the fraction of optimum control because the controls come from two different domains (release quantity for push and WIP for CONWIP). In a simulation study it is easy to find optimum controls because one can experiment with the actual system that will make use of the controls. In a real factory, optimization is much more difficult because parameters are unknown, subject to change and estimates can be subject to large biases. As we see in this example, CONWIP performs very well with control parameters in excess of fifteen per cent away from the optimum while the
push system performance degrades much more rapidly as the control diverges from optimum.

The push system WIP has a mean of 48 jobs, which is actually slightly less than the constant WIP in the optimally controlled CONWIP system (although Fig. 4 demonstrates that in this example, the CONWIP system would outperform the push system with a lower WIP). The important difference is that the push system WIP varies wildly as can be seen in Fig. 5. This results in high tardiness costs some of the time and high FGI costs at other times. In this case, the superiority of CONWIP is not due to WIP reduction; it is due to WIP control. The WIP variability in the push system also has strategic implications, because job flow time varies with the WIP and is clearly unpredictable.

Of course, one should not try to draw general conclusions from a simulation study. Cost, service and arrival parameters were varied and the same qualitative results held. This result was expected. As we argued in Section 4, CONWIP does a better job of 'using up' excess capacity.

7. Extensions

So far we have been concerned mostly with the use of CONWIP for the control of a single line. In this section we outline the issues associated with using CONWIP to control plants that produce all parts in a bill of materials. There are design problems, such as inter-line buffer sizing, whose solution depends on the characteristics of the manufacturing firm and its markets so we will not be able to specify complete general solutions here.

One of the main advantages that CONWIP offers is that the flow times of CONWIP lines are fairly predictable because the WIP levels are nearly constant. It is
Figure 4. Controllability.

Figure 5. Push system WIP versus time with optimum controls.
much easier to coordinate production in a line with constant WIP than one where the WIP levels cannot be known a priori.

When a product requires the production and assembly of components, the production control system must coordinate the operation of multiple lines. For a system of CONWIP production lines, this means sequencing the backlog for all the lines. The objective is to sequence the backlogs so that components come together for assembly at roughly the same time, as close as possible to the 'right' time as determined by the due date for the finished product.

An obvious solution is to use an mrp-style explosion of the bill of materials to generate the backlogs for the multiple lines. Note that lead times are not required since each CONWIP line will run at the rate at which parts are being assembled. After an initial transient, the lines will produce, on average, at exactly the same rate. An example of this type of system for two fabrication lines feeding assembly is shown in Fig. 6.

This scheme will work well if the assembly operation is the bottleneck or there are not significant set-ups. If a fabrication line is the bottleneck and there are significant set-ups on that line's bottleneck, then the backlog must be sequenced to assure adequate capacity for the line. This backlog must be propagated to the other fabrication lines and to the assembly operation. In many cases, backlog creation could become very complex; algorithms for sequencing backlogs when there are significant set-ups on bottleneck fabrication lines is the subject of ongoing research.

8. Conclusions

We have described a pull-based production control strategy that offers the possibility of significant improvements over other production control systems. This seems to be particularly true at high levels of plant utilization and in environments with distinct bottleneck operations. Many of the benefits of CONWIP can be attributed to the fact that it is a pull-based production system.

The system does offer some distinct advantages over kanban. One of them is that it can be used in some production environments where kanban is impractical because of
too many part numbers or because of significant set-ups. By allowing WIP to collect in front of the bottleneck, CONWIP can function with lower WIP than kanban. Also, CONWIP makes use of backlog of part numbers, which can allow job sequencing to be done by production control personnel.

While the backlog affords the opportunity for control, it also provides a tremendous challenge. The backlog sequence is the key to assuring adequate capacity when there are significant set-ups and to optimizing synchronization of production of part components. These issues are the formidable research challenge presented to us by CONWIP.

CONWIP is not necessarily the optimal means of controlling production in every situation. However, for those firms attempting to operate their production lines near capacity it is certainly worthy of further research. We have already begun implementation of the CONWIP approach in a large circuit board plant. We will describe this experience in a subsequent paper.

References


Hall, W. R., 1981, Driving the Productivity Machine: Production Planning and Control in Japan (Falls Church, Virginia: American Production and Inventory Control Society).


