PUSH AND PULL PRODUCTION SYSTEMS: ISSUES AND COMPARISONS

MARK L. SPEARMAN and MICHAEL A. ZAZANIS

Northwestern University, Evanston, Illinois

(Received September 1988; revisions received April, November 1989, May, June 1990; accepted August 1990)

Concerns about American manufacturing competitiveness compel new interest in alternative production control strategies. In this paper, we examine the behavior of push and pull production systems in an attempt to explain the apparent superior performance of pull systems. We consider three conjectures: that pull systems have less congestion; that pull systems are inherently easier to control; and that the benefits of a pull environment owe more to the fact that WIP is bounded than to the practice of "pulling" everywhere. We examine these conjectures for analytically tractable models. In doing so, we not only find supporting evidence for our surmises but also identify a control strategy that has push and pull characteristics and appears to outperform both pure push and pure pull systems. This hybrid system also appears to be more general in its applicability than traditional pull systems such as Kanban.

Increased foreign competition has intensified the need for more effective manufacturing. However, the means to accomplish this task has become a subject of controversy. On one hand, much of the practitioner literature suggests that the implementation of Computer Integrated Manufacturing (CIM) is the only means available to regain our position of manufacturing leadership, see, e.g., Vollum (1984) and Berger (1986). Other authors cite the Japanese as having achieved an extremely competitive position while employing limited automation and using simple and decentralized management techniques, e.g., Schonberger (1986).

This debate stems from the clash of two diametrical viewpoints. In one vein, CIM represents the culmination of manufacturing computer involvement that began with material requirements planning (MRP), a suggested improvement over the older reorder point (ROP) system, in the early 1970s. In the opposing vein, the so-called Japanese manufacturing techniques such as just-in-time (JIT) or zero inventories (ZI), make little use of computers and instead place greater responsibility for schedule compliance and quality on the production worker. However, the techniques used to implement JIT and ZI are, in many ways, identical to those found in the "out-dated" ROP systems. *Plus ça change, plus c'est la même chose.*

The terms *push* and *pull* refer to the means for releasing jobs into the production facility. In a push system, a job is started on a *start date* that is computed by subtracting an established *lead time* from the date the material is required, either for shipping or for assembly. A pull system is characterized by the practice of downstream work centers pulling stock from previous operations, as needed. All operations then perform work only to replenish outgoing stock. Work is coordinated by using some sort of signal (or *Kanban*) represented by a card or sign.

One problem with comparing pull and push systems is that terms like JIT have come to mean more than a way to schedule production. JIT includes other features such as short setup times, perfect quality, stockless production, and increased worker involvement. To a certain extent, JIT has come to refer to all that is good in manufacturing. As such, it is difficult to understand *when* and *why* push and pull systems are effective. This paper seeks to address this problem by studying the essence of push and pull in several simple and analytically tractable situations. In particular, we will:

1. address the issues associated with push and pull systems and devise a set of pertinent measures;
2. conjecture reasons for the improved performance of pull systems over push systems;
3. test these conjectures with theoretical comparisons of push and pull systems.

We believe this research has led to a better understanding of how pull systems work. As a result, we are able to propose a system that has characteristics of both push and pull that appears to outperform both pure push and pure pull systems.

Subject classifications: Inventory/production: kanban and other pull systems. Production/scheduling: stochastic.
Area of review: MANUFACTURING, PRODUCTION AND SCHEDULING.

0030-364X/92/4003-0521 $01.25

1. LITERATURE REVIEW

Because of increased concerns about American manufacturing competitiveness, there is new interest in alternative production control systems. Much of the discussion in the literature focuses on the relative merits of push (e.g., MRP) and pull (e.g., Kanban) systems. However, most of the literature dealing with Kanban and other pull systems is descriptive in that few mathematical models have been developed. This fact is noted by Bitran and Chang (1987) and Zangwill (1987).

Hall (1983) provides a good description of how Kanban works and gives some important implementation details. Schonberger suggests that the "variability reduction" found in pull systems is extremely important to overall system effectiveness. This is reiterated by Chen et al. (1988) in a study using queueing networks. Finally, Karmarkar (1986) points out that the number of cards (Kanbans) in the system creates an upper limit on work-in-process (WIP).

The few papers that have provided mathematical models concentrate on deterministic settings. These include work by Kimura and Terada (1981) who develop some basic equations for a Kanban system, and Bitran and Chang (1987) who provide a mathematical programming approach for optimizing a deterministic Kanban system.

One way to avoid the analytic difficulties of modeling Kanban systems is to use simulation and to compare effectiveness in specific instances. An early study by Kimura and Terada compares the effect of fluctuations in demand in push and pull systems and concludes that Kanban tends to dampen these fluctuations. On the other hand, Ritzman and Krajewski (1983) were able to demonstrate that MRP is more effective than ROP in systems having many levels in the bill of material structure and larger lot sizes. More recently, Krajewski et al. (1987) performed an extremely detailed study using a simulation model that has been validated extensively with industry experience. This study involved a great many factors, including customer influences such as forecast error and "specials," vendor influence, buffer mechanisms, product structure, facility design, scrap loss, equipment failures, worker flexibility, inventory accuracy, and lot sizing rules. The principal measures used were percent of past due demand and total inventory. Some of the conclusions of this study were:

> . . . uniform workflows and flexibility to adjust to changing capacity requirements is the key to improving performance. The Kanban system, by itself, is *not crucial* to improving performance (emphasis added).

An important result of this study is the realization that the manufacturing environment itself may have a greater impact on system performance than the type of control strategy used. In light of these conclusions, it is important to separate environmental factors from those related to production control strategies.

Another reason it is difficult to compare push and pull systems is that their basic modes of operation are radically different. Push systems control throughput by establishing a Master Production Schedule (MPS) and measure WIP (e.g., input/output control; see Wight 1970) to detect problems in meeting a schedule. Pull systems, on the other hand, control WIP and must measure throughput against required demand. This is typically accomplished using some sort of quota system that represents the amount of production required for each time period. If the quota is always met, no due dates will be missed. However, models that will predict quota shortfalls before the end of a period in a stochastic production facility are rare. Fortunately, Kanban systems offer production foremen great visibility to the status of backlogs (Karmarkar).

1.1. The Contribution of this Paper

The purpose of this paper is not simply to compare Kanban and MRP, but to offer theoretical motivations for the apparent superior performance of pull systems. Along with a general heightened awareness of environmental issues given by Krajewski et al. (1987) and Karmarkar (1986), we submit the following conjectures:

1. There is less congestion in pull systems.
2. Pull systems are inherently easier to control than push systems.
3. The benefits of a pull environment owe more to the fact that WIP is bounded than to the practice of "pulling" everywhere.

To examine these conjectures we first identify basic issues regarding the operation of push and pull systems, particularly, controls and performance measures. We then compare these performance measures in simple models for push and pull systems for which analytic results are available.

The first conjecture regarding congestion is tested using an open queueing network of tandem exponential queues with Poisson arrivals to represent the push system and an "equivalent" closed queueing network (CQN) to represent the pull system. These networks are equivalent in that they have the same stations with the same throughput. In the closed system, jobs are

pulled into the production facility whenever an earlier job is completed and are then pushed between stations. In a Kanban system, jobs are pulled by each station from the previous station. The closed system thus represents a production facility operating under, what we call, a *constant work-in-process* (CONWIP) strategy. The performance of this system sheds light on why pull systems work well and presents an exciting area for further research. Although our results are exact only for exponential processing times and Poisson arrivals, we believe they provide needed insight into other, more realistic, systems.

Our controllability conjecture deals with both practical control considerations and control robustness. In this case, we ignore the improved performance of pull systems implied by the first conjecture and compare instead, the robustness of controlling WIP (as in the pull system) with controlling throughput (as in a push system). For these comparisons, the assumptions are substantially weaker.

To test the last conjecture we compare the performance of Kanban, which pulls everywhere, with CONWIP, which pulls only at the front of the line. Again, these systems involve exponential processing times. Figure 1 presents the three networks used for comparison.

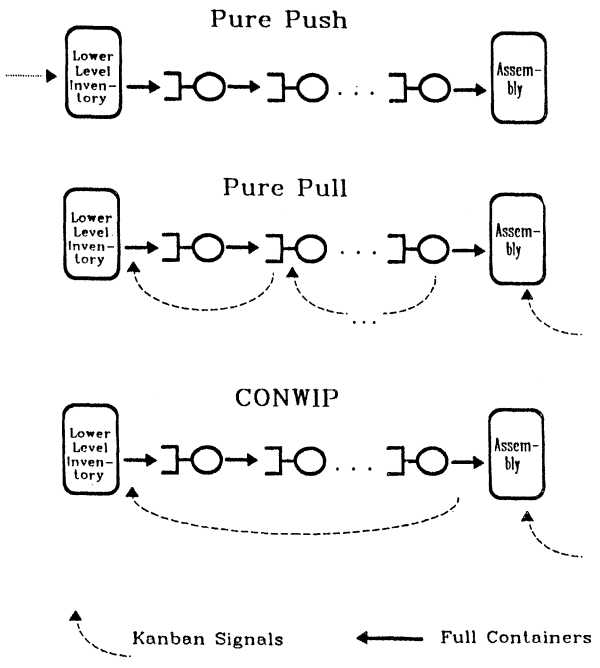

Figure 1. Pure push, pure pull, and constant WIP systems.

2. CONGESTION IN PRODUCTION SYSTEMS

We first address the issues related to macroscopic performance measures of production systems. Before comparing push and pull systems, however, it is necessary to describe the nature of the production settings for which our comparisons are valid and to generalize the production control problem in order to establish relationships between pertinent measures.

2.1. Issues

To be sufficiently general, we do *not* assume the presence of a stationary demand process. Instead, we allow the demand that is seen by the shop floor to be the result of an exogenous demand process combined with a feedback process generated within the firm. Between the actual customer and the shop floor we assume there is some sort of buffer. In many cases, this buffer is the master production schedule. Thus, our results may not hold for situations where jobs are released to the shop as soon as they are received (e.g., a copy shop). However, in many cases, we believe our assumption to be more realistic than assuming a stationary demand process that feeds directly to the shop. In both the build-to-order and the build-to-stock cases, management will usually work to maintain a comfortable level of demand in the MPS. Thus, when the MPS for a given product is filled many weeks out, the sales force will concentrate on other, less popular, products. Likewise, if it appears that the plant will soon catch up with all demand on the MPS, management will respond in one of a variety of ways. It can either "push sales" with discounts or other mechanisms, build anticipation inventory, or reduce production capacity by eliminating shifts or reducing the workforce. In any case, the outside demand process has less impact on the utilization of the individual stations than if jobs arrived independently and proceeded directly to the shop floor.

For these reasons, we assume that there is always demand for the shop. This is important to both push and pull systems. In the push system, we assume that there is always a complete release of work at the start of each period (day, shift, etc.). The decision variables in these systems are how much to release and how often. In the pull system, the decision variable is how much WIP to maintain between various stations. Since the throughput of a pull system is a result of this WIP configuration, the presence of the MPS

buffer means that the pull system will always be able to start work when authorized. Thus, although the *firm* is subject to a nonstationary demand process, the production facility itself is assumed to operate in steady state with a constant Poisson arrival rate in the push system and unlimited product availability in the pull system.

We further generalize the production control problem by considering the production of a part between two adjacent levels within a product structure. This "line" could represent fabrication of a part needed for an assembly operation (in a multilevel bill of material) or the completion of an entire job from raw materials to finished goods (as in a steel mill). All machines in this line are assumed to have independent processing times. We realize that these are ideal conditions for both the push and the pull systems. However, in this section, we are only interested in the relative performance of the two systems and not in robustness of control, which is addressed in the next section.

We will assume that the part is needed some time l (the lead time) after it is requested. The service level for the line, s, is defined as the probability that the time to complete the part is less than or equal to l. We further assume that any parts arriving before l time units have elapsed will wait in inventory. This is not unusual because in an assembly operation all parts must be present before the assembly can be completed. Likewise, in a single level operation, production is typically coordinated to a shipping schedule (see Kanet and Christy 1984 for a discussion of forbidden early-order departure). We represent the average of this waiting inventory as I_w. Finally, we let n, θ, μ_T, and σ_T designate, respectively, the average work-in-process, the throughput of the line, and the mean and variance of the cycle time in steady state.

We further assume that for n (and, hence, θ) in some range of interest, there exists a distribution function Φ for the cycle time, T, such that

$$F(t) = P\{T \leq t\} = \Phi\left(\frac{t - \mu_T}{\sigma_T}\right).$$

We believe that this assumption is satisfied *approximately* for many systems. Hence, the expression for the average waiting inventory that follows is approximate as well. For further discussion see the Appendix. Define $z_s = \inf\{u: \Phi(u) \geq s\}$. Then for fixed values of throughput θ, and service s, we see that:

1. From Little's law:

   $n = \mu_T \theta$,

   so that WIP depends only on mean cycle time.

2. The average waiting inventory, I_w will be

$$I_w = \theta E[\max(0, l - T)] = \theta \int_0^l (l - t)\, dF(t)$$

$$= \sigma_T \theta \int_0^{z_s} \Phi(z)\, dz, \qquad (1)$$

   so that I_w depends only on the variance of cycle time.

3. The necessary lead time will be,

   $l = \sigma_T z_s + \mu_T$,

   so that l is a linear function of both the mean and standard deviation of cycle time.

Thus, for fixed values of throughput and service we conclude that it is important for cycle time to have both a small mean and a small variance, in that:

1. Smaller mean cycle times allow for:
   - more competitive lead times;
   - smaller reorder points;
   - shorter MPS frozen zone, and thereby, more flexibility;
   - less WIP, and thereby, less exposure to changes and damage;
   - less WIP, and thereby, less inventory investment.
2. Smaller cycle time variance allows for:
   - more competitive lead times;
   - smaller buffers between stages;
   - less waiting (finished) inventory, and thereby, less inventory investment.

2.2. Comparisons

We are able to obtain a comparison of mean cycle time directly by comparing the performance of an open queueing network with an "equivalent" closed queueing network. Consider a closed tandem system made up of K stations with n customers in which each station is composed of a single exponential server with rate μ_i. The throughput, $\theta(n)$, can be written as $G(n-1)/G(n)$, where $G(n)$ is the normalization constant found in Buzen (1973). Consider an "equivalent" open network with Poisson input rate, $\lambda(n) = \theta(n)$. Let EN_i^o and EN_i^c be the expected number of jobs at the ith station of the open and closed systems, respectively.

Theorem 1. *For all K and n, $EN_i^o > EN_i^c$.*

Proof. Clearly

$$EN_i^o = \frac{\lambda(n)/\mu_i}{1 - \lambda(n)/\mu_i} = \sum_{j=1}^{\infty} \left(\frac{\lambda(n)}{\mu_i}\right)^j. \qquad (2)$$

On the other hand,

$$
\begin{aligned}
&EN_i^c \\
&= \sum_{j=1}^{n} P\{N_i \geq j\} \\
&= \sum_{j=1}^{n} \frac{G(n-j)}{G(n)} \left(\frac{1}{\mu_i}\right)^j \\
&= \sum_{j=1}^{n} \frac{G(n-j)\,G(n-j+1)}{G(n-j+1)G(n-j+2)} \cdots \frac{G(n-1)}{G(n)} \left(\frac{1}{\mu_i}\right)^j. \quad (3)
\end{aligned}
$$

But, for all $m > 0$,

$$
\frac{G(n-m)}{G(n-m+1)} \leq \frac{G(n-1)}{G(n)} \quad (4)
$$

with equality holding for $m = 1$. Hence, from (3) we have

$$
\begin{aligned}
EN_i^c &\leq \sum_{j=1}^{n} \left(\frac{G(n-1)}{G(n)}\right)^j \left(\frac{1}{\mu_i}\right)^j \\
&= \sum_{j=1}^{n} \left(\frac{\lambda(n)}{\mu_i}\right)^j < \sum_{j=1}^{\infty} \left(\frac{\lambda(n)}{\mu_i}\right)^j.
\end{aligned}
$$

Corollary 1. *Two results immediately follow from the above theorem.*

i. If EN^o is the expected total number of customers in the system, then $EN^o > n$.

ii. If ET^o is the expected cycle time in the open system and ET^c is the expected cycle time in the closed system, then $ET^o > ET^c$.

2.3. Discussion

The implications of these results are clear—there is less congestion in a closed queueing network than in an equivalent open network. Obviously, a CONWIP system will have less WIP than an equivalent push system operating under these conditions. If the cycle time distributions for the push and CONWIP systems can be approximated by distributions from a family having the aforementioned characteristics, CONWIP has better performance on the measures related to mean cycle time. If the cycle time variance in the closed system is less than that in the open system, we obtain the other competitiveness results. We conjecture that this is true because the number of customers at different servers is negatively correlated in the closed network and is not correlated at all in the open. This notion of "negative dependence" has been formalized by Whitt (1984).

Equally clear is the fact that CONWIP, and indeed all pull systems, require a "better" production environment than do push systems. For instance, in the above example we require the existence of unlimited availability of product and the management of the MPS that allows "work ahead." This more pristine environment is what Krajewski et al. attribute the improved effectiveness of Kanban over MRP.

However, it is important to note that there are many ways that the production environment can be improved and that pull systems are better equipped to exploit these improvements than push systems. For example, as discussed above, most firms do have some control over due dates (someone must agree to them) and have some control over production rates. The ability of pull systems to work ahead allows them to take advantage of this improved environment. Of course, push systems could also work ahead but this would require a manual override of the installed release mechanism and some sort of indicator based on shop conditions. However, as we modify the push to take advantage of downstream information it begins to take on characteristics of a pull system.

We point out, however, that in the absence of the ability to work ahead, the CONWIP system may have *more* WIP, on average, than the push system. Consider a system subject to external Poisson demands with no MPS or other buffering mechanism, i.e., all demands go directly to the shop floor. A CONWIP-like system would hold jobs in a buffer outside the shop if the WIP level for the shop was at the desired level. The push system, on the other hand, would always allow arriving jobs into the shop. It is clear then that cycle times in the push system would be less than in the pull system because work is never started later, and would therefore have less average WIP.

This example highlights the importance of the buffering mechanism that separates the external demand process from the shop floor and the ability of pull systems to exploit the opportunity to work ahead. Given these environmental conditions, CONWIP lines tend to have shorter and more predictable cycle times that push lines. Consequently, CONWIP lines can make better use of these additional controls than push systems. We now further develop this issue of controllability.

3. CONTROLLABILITY

As stated earlier, a push system controls throughput and measures WIP while a pull system controls WIP and measures throughput. We will discuss the issues surrounding controllability and then compare the two systems.

3.1. Issues

There are two issues related to controllability. The first deals with practical implementation considerations while the second deals with robustness, i.e., the sensitivity of an optimal strategy to errors in control. Regarding the first issue, the pull system is clearly superior for three reasons:

1. WIP is easier to control than throughput because it can be observed directly.
2. Throughput is typically controlled with respect to *capacity*. As such, capacity must be estimated and such estimates must include details, such as process time, setup time, random outages, worker efficiency, and rework.
3. Throughput is controlled by specifying an *input* rate. If the input rate is less than the capacity of the line, then throughput is equal to input. If not, throughput is equal to capacity and WIP builds without bound. By incorrectly estimating capacity, input can easily exceed the true capacity. This is particularly true when seeking high utilization rates.

In addition to the practical control issues just discussed, we now show that it is inherently easier to control WIP than to control throughput. First note that by using an appropriate function relating the specified WIP (throughput) to the resulting throughput (WIP), we can control a pull (push) system by specifying a throughput rate (WIP level) and then computing and setting the corresponding WIP level (throughput rate). Our point here is not so much that pull systems are easier to control, but that controlling WIP is more robust than controlling throughput. The fact that it is natural to control WIP in pull systems leads to the conclusion that such systems are easier to control.

3.2. Comparisons

The robustness comparison comes from computing the sensitivity of an optimal strategy to errors in setting control levels. Our comparison involves a simple static optimization problem where the control variable is set to an optimal value. Of course, not allowing for feedback and a "closed loop" control policy is an oversimplification of the real situations. However, we believe that even this naive model sheds some light on the robustness issues that arise when controlling push and pull systems.

We consider the optimization problem of balancing the cost of lost production with the cost of added WIP. Lost production costs take the form of missed sales opportunities, while WIP costs include not only inventory carrying costs but also the cost of longer lead times and less flexibility. We denote the (average) WIP by n and the steady-state throughput by θ and assume that there exists a twice continuously differentiable function, f, such that $\theta = f(n)$. We further assume f to be nondecreasing and strictly concave everywhere and note that Shanthikumar and Yao (1988) have shown that this is true for Jackson-like networks in which the service rate at each station is a nondecreasing concave function of the queue length. Since this function is one-to-one,

$$\theta = f(n)$$

$$n = f^{-1}(\theta).$$

If p represents the marginal profit per piece and c the associated carrying cost per period per piece, then the profit per unit time, Z, associated with a particular policy will be

$$Z(n) = pf(n) - cn \quad \text{controlling WIP.} \tag{5}$$

$$\hat{Z}(\theta) = p\theta - cf^{-1}(\theta) \quad \text{controlling throughput.} \tag{6}$$

Note that

$$\hat{Z}(\theta) = \hat{Z}(f(n)) = Z(n).$$

We define θ^* and n^* to be the values of θ and n that maximize Z. These values exist and are unique because f is concave everywhere.

Robustness is related to the sensitivity of an optimal strategy to small changes in the control. A Taylor expansion of Z at the optimum control level shows that this sensitivity is characterized by the second derivative of Z. To avoid dimensional problems we define these in terms of the optimal values themselves. Hence we want to compare

$$\theta^{*2} \frac{d^2\hat{Z}(\theta)}{d\theta^2}\bigg|_{\theta^*} \text{ versus } n^{*2} \frac{d^2Z(n)}{dn^2}\bigg|_{n^*}.$$

Lemma 1

$$\frac{d^2\hat{Z}(\theta)}{d\theta^2}\bigg|_{\theta^*} \left(\frac{d^2Z(n)}{dn^2}\bigg|_{n^*}\right)^{-1} = \frac{p^2}{c^2}.$$

Proof. Elementary calculus is applied to the definitions of Z and f. (The assumption of the strict concavity of f guarantees that d^2Z/dn^2 is strictly negative.)

The following theorem shows that controlling WIP is more robust than controlling throughput.

Theorem 2

$$\theta^{*2}\frac{d^2\hat{Z}(\theta)}{d\theta^2}\bigg|_{\theta^*} > n^{*2}\frac{d^2Z(n)}{dn^2}\bigg|_{n^*}.$$

Proof. We assume that Z^* is positive, otherwise we would quickly go out of business, so that,

$$Z(n^*) = pf(n^*) - cn^* > 0.$$

Then

$$p\theta^* > cn^*$$

$$\frac{\theta^{*2}p^2}{n^{*2}c^2} > 1.$$

The desired result comes directly from the lemma.

3.3. Discussion

The conditions for WIP to be controlled more easily than throughput are rather weak: 1) throughput must be a one-to-one concave function of WIP, and 2) there must exist a control level at which the business is profitable.

While in theory for any system either throughput or (average) WIP can be the control variable, in practice control policies for one of these variables are typically more natural and more easily implemented. Comparisons of control policies for different systems do not fall in the framework of Theorem 2.

Consider a simple example. Both systems are composed of five identical exponential servers with unit processing times and cost coefficients of $p = 100$ and $c = 1$. For one system we allow Poison arrivals, in the other we employ CONWIP. Since these strategies are different the function relating WIP to throughput in the CONWIP system will not be the inverse of the function relating throughput to average WIP in the push system. We therefore need two functions, viz.

$$f(n) = \frac{n}{n+4} \quad \text{controlling WIP}$$

$$g(\theta) = \frac{5\theta}{1-\theta} \quad \text{controlling throughput.}$$

Figure 2 compares the two profit curves as a function of percent of optimal. Note that the difference in congestion (implied by our first theorem) yields a small gap at the optimal. More important is the fact that CONWIP outperforms optimal push with WIP levels as small as 40% too low and as large as 60% too high. Also, since throughput is bounded by capacity, the push profit falls sharply above the optimal and

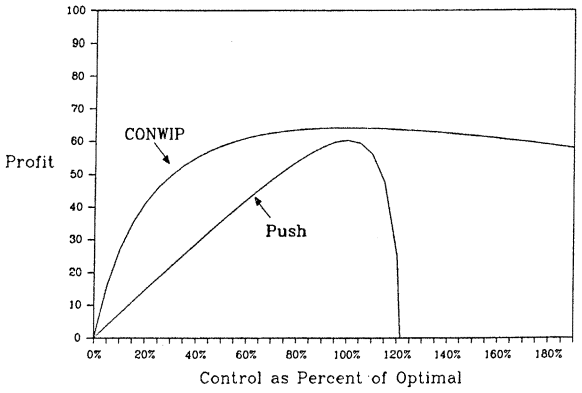

Figure 2. Profit functions for exponential lines with identical stations.

becomes negative approximately 20% above optimal. CONWIP continues to have positive profits until WIP levels have reached 600% above optimal.

Note that this type of stationary analysis is not always a good representation of reality. It does, however, give us an indication of why pull systems are easier to control. In a real production system there would be feedback under both push and pull production scenarios. Consider the case of overestimating capacity in the push system by more than 20%. Although the steady-state model would predict a negative profit level, it is unlikely that the plant management would allow conditions to become so bad. Instead, as WIP levels and cycle times begin to rise some emergency action, such as overtime or cancellation of orders, would be taken. Although this may not be as bad as negative profit, it is certainly not desirable.

Also note that in a push system there is a strong temptation to overload the MPS by optimistically estimating capacities. Unfortunately, this is often not recognized until WIP levels have become excessive. If the quota used in a pull system is inflated, it will be discovered almost immediately.

Finally, a conservative pull user will err high when setting WIP levels. She will then gradually reduce WIP levels until starvation of bottleneck resources occur. Figure 2 shows the danger of erring on either side of optimal when setting input rates in a throughput controlled system.

4. THE MAGIC OF PULL

In this section, we motivate the conjecture that the bounding of WIP in a pull system can be more important than pulling at every station. We do this by

comparing the throughput of a system in which WIP is bounded but allows "pushing" between stations (CONWIP) with an equivalent system which pulls at each station (Kanban). Our comparison involves exponential CONWIP and Kanban systems subject to infinite demand with an infinite supply of raw materials. Before doing this, however, we discuss some of the practical benefits associated with CONWIP.

4.1. Issues

Monden (1983) describes the basic philosophy behind the Toyota Production System as a means to "... produce the kind of units needed, at the time needed, and in the quantities needed." Kanban represents a subset of this larger system and is used to maintain just-in-time production. However, Monden warns that:

> Unless the various prerequisites of this system are implemented perfectly (i.e., design of processes, standardization of operations and smoothing of production, etc.), then just-in-time will be difficult to realize, even though the Kanban system is introduced.

The conditions necessary for Kanban to work well are (Monden, p. 4):

1. "Smooth" production involving a stable product mix;
2. Short setups;
3. Proper machine layout;
4. Standardization of jobs;
5. Improvement activities;
6. Autonomation (autonomous defects control).

If these conditions are met, Kanban provides "an information system to harmoniously control the production quantities at every process."

We believe, however, that such harmonious control is largely due to the fact that WIP is bounded, thereby creating shorter and less variable cycle times, and not from the fact that the Kanban system pulls at every station. If we relax the pulling requirement, it appears that we can achieve the benefits of Kanban in less pristine production environments.

For instance, we no longer require a stable product mix. In a CONWIP system the pull signal specifies only a certain *routing*. Any part that uses that routing can be started. Synchronization of assembly operations is accomplished by starting jobs in a predetermined *sequence*.

Secondly, CONWIP addresses the problem of having many part numbers on a single line. For instance, a circuit board operation typically has a small set of product *types* (i.e., dimensions of the boards) and a large number of unique part numbers (i.e., different "artwork"). In such an environment it is almost impossible to implement a Kanban system because some WIP is required for each active part number. Using CONWIP, a new job is started whenever an existing job is completed. Consequently, the new job must share the same routing as the completed job but does not have to be for the same part number.

Also, CONWIP is simpler to operate since only one value of WIP must be specified for an entire line versus a Kanban system in which the number of cards must be specified for each station.

Finally, robust queueing models are available to predict the performance of CONWIP systems (Gordon and Newell 1967, Reiser and Kobayashi 1975, Denning and Buzen 1978, and Reiser and Lavenberg 1980). On the other hand, modeling stochastic Kanban systems is extremely difficult.

4.2. Comparisons

Aside from the above practical concerns, we believe that pulling everywhere is actually less efficient than maintaining a constant WIP level. By efficient we mean the throughput obtained for a given number of Kanbans. We were first led to this conclusion when observing in simulation studies that WIP would naturally accumulate before bottleneck processes in CONWIP systems. This, of course, tends to increase the utilization of the bottleneck and, thereby, increases the throughput of the entire line. We can demonstrate this effect by comparing the throughput of a CONWIP system composed of exponential servers with that of an equivalent (i.e., having the same number of cards) Kanban system (see Figure 3).

For our comparison, we consider systems having K stations, each station with an exponential machine having a processing rate of μ_i. The Kanban system works as follows. Containers of parts with Kanbans attached are queued at "stock points" at each station (indicated by ∇ in Figure 3). Authorizing Kanbans indicating demand from the next station are stored in a Kanban box (indicated by $\sqcup$ in the figure). Both systems have unlimited raw material (shaded ∇) and are subject to unlimited demand (shaded $\sqcup$). Consequently, the stock point at the first station *always* has one container of parts and the Kanban box at station K *always* has one Kanban.

If station i has a container of parts at its stock point *and* a Kanban in its Kanban box, a worker will remove the Kanban from the container and send it back to station $i-1$ (indicating a replenishment demand) and begin work on the parts. When finished, the worker removes the authorizing Kanban from the box,

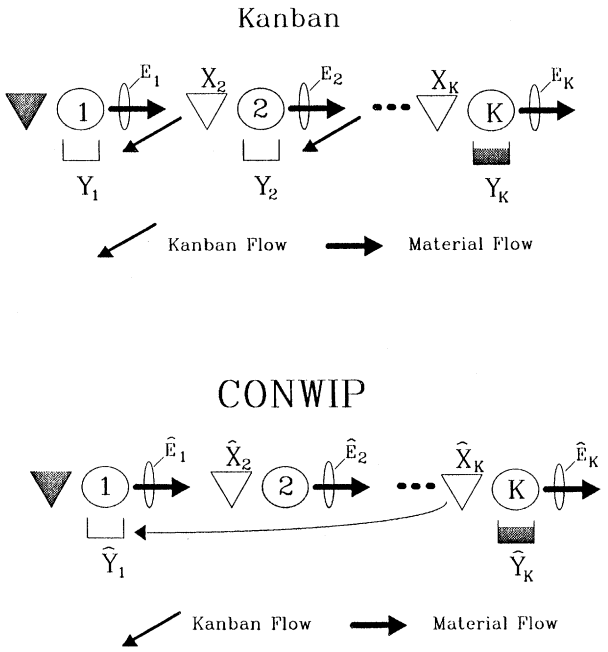

Figure 3. Equivalent Kanban and CONWIP systems.

attaches it to the container and then sends both to station $i + 1$. (Note that in most real Kanban systems there is an inbound and an outbound stock point and two Kanbans, one authorizing production and one authorizing conveyance. Our system is equivalent in that the movement of material can, itself, be considered as a process between two stockpoints.)

The first station in the CONWIP system works in the same fashion as the Kanban system. At the other stations the only requirement to begin work is that a container is available in the stock point. Kanbans flow from station K to station 1 whenever a container is started at K. Otherwise, Kanbans flow along with the material from station to station.

We begin the comparison with all the Kanbans attached to containers and n_i containers at station i, $1 < i \leqslant K$ and no Kanbans in any box except at station K. Because of the one Kanban at K, work will begin immediately at all stations. We now use uniformization (e.g., see Ross 1983) to construct sample paths of the two systems on the same probability space in order to compare their outputs.

Let $(t_u)_{j=1,2,\ldots}$ denote the epochs in a Poisson counting process with rate $\sum_i \mu_i$. Event ϵ_i occurs at epoch t_j with probability $\mu_i/\sum_i \mu_i$ and corresponds to the potential completion of a container of parts. Completion depends on the status of the system, i.e., whether there is work to complete and whether there is an authorizing Kanban. Let $Y_i(t_j)[\hat{Y}_i(t_j)]$ denote the number of Kanbans in the box and $X_i(t_j)[\hat{X}_i(t_j)]$ denote the number of containers at station i at epoch t_j for the Kanban [CONWIP] system. Finally, denote the number of containers completed and moved from station i immediately after epoch t_j by $E_i(t_j)[\hat{E}_i(t_j)]$.

At $t = 0$, we note that

$$Y_i(0) = 0,\ 1 \leqslant i < K$$

$$\hat{Y}_i(0) = 0,\ 1 < i < K$$

$$X_i(0) = \hat{X}_i(0) = n_i > 0,\ 1 < i \leqslant K$$

$$E_i(0) = \hat{E}_i(0) = 0,\ 1 < i \leqslant K. \tag{7}$$

Also,

$$Y_K(t) = \hat{Y}_K(t) = 1,\ t \geqslant 0$$

$$X_1(t) = \hat{X}_1(t) = 1,\ t \geqslant 0.$$

For $n = 1, 2, \ldots$, note the following relationships: For event ϵ_1:

$$E_1(t_j) = E_1(t_{j-1}) + I\{Y_1(t_{j-1}) > 0\} \tag{8}$$

$$\hat{E}_1(t_j) = \hat{E}_1(t_{j-1}) + I\{\hat{Y}_1(t_{j-1}) > 0\} \tag{9}$$

$$Y_1(t_j) = E_1(t_j) - E_1(t_j) + I\{X_2(t_j) > 0\} I\{Y_2(t_j) > 0\} \tag{10}$$

$$\hat{Y}_1(t_j) = \hat{E}_K(t_j) - \hat{E}_1(t_j) + I\{\hat{X}_K(t_j) > 0\}, \tag{11}$$

where $I\{.\}$ is the indicator function.

For event ϵ_i, $0 < i \leqslant K$:

$$E_i(t_j) = E_i(t_{j-1}) + I\{Y_i(t_{j-1}) > 0\} I\{X_i(t_{j-1}) > 0\} \tag{12}$$

$$\hat{E}_i(t_j) = \hat{E}_i(t_{j-1}) + I\{\hat{X}_i(t_{j-1}) > 0\} \tag{13}$$

$$Y_i(t_j) = E_{i+1}(t_j) - E_i(t_j) + I\{Y_{i+1}(t_j) > 0\} I\{X_{i+1}(t_j) > 0\} \tag{14}$$

$$X_i(t_j) = n_i + E_{i-1}(t_j) - E_i(t_j) \tag{15}$$

$$\hat{X}_i(t_j) = n_i + \hat{E}_{i-1}(t_j) - \hat{E}_i(t_j). \tag{16}$$

The following lemma is needed to obtain our comparison.

Lemma 2. *For all i and j*

$$E_i(t_j) \leqslant \hat{E}_i(t_j) + n_{i+1}.$$

Proof. The proof involves two cases: $i = 1$ and $1 < i \leqslant K$.

Case 1 ($i = 1$). At $t = 0$ the inequality of Lemma 2 is true by hypothesis. Assume that is is true at t_{j-1}. We only need to examine the case where the inequality is

not strict, namely

$$E_1(t_{j-1}) = \hat{E}_1(t_{j-1}) + n_2.$$

Using (10) and (11) and dropping the t_{j-1}

$$\hat{Y}_1 - Y_1 = \hat{E}_K - E_2 - \hat{E}_1 + E_1 + I\{\hat{X}_K > 0\} - I\{X_2 > 0, Y_2 > 0\}. \quad (17)$$

Using (15) allows E_2 to be written as,

$$E_2 = E_K - \sum_{j=3}^{K-1} (n_j - X_j).$$

So that (17) can be written as

$$\hat{Y}_1 - Y_1 = \hat{E}_K - E_K + \sum_{j=3}^{K-1} (n_j - X_j) + n_2 + I\{\hat{X}_K > 0\} - I\{X_2 > 0\} I\{Y_2 > 0\}.$$

The sum of the first two terms on the right-hand side is positive by hypothesis, since there can be only one event per epoch. The summation term is nonnegative since $n_j \geq X_j$. The sum of the last three terms is nonnegative since $n_2 > 0$. The desired result is obtained by considering (8) and (9).

Case 2 ($1 < i \leq K$). Again, the only interesting case is when

$$E_i(t_{j-1}) = \hat{E}_i(t_{j-1}) + n_{i+1}.$$

From (16) and (15) and again dropping t_{j-1},

$$\hat{X}_i - X_i = \hat{E}_{i-1} - E_{i-1} - \hat{E}_i + E_i = \hat{E}_{i=1} - E_{i-1} + n_{i+1}. \quad (18)$$

Since the right-hand side is greater than zero by hypothesis, the desired result is obtained by considering (12) and (13).

The following theorem provides our throughput comparison.

Theorem 3. *The throughput of an exponential Kanban system will not exceed that of an equivalent CONWIP system, almost surely.*

Proof. This is immediate consequence of the above lemma and the fact that $n_{K+1} = 0$.

This result is not surprising because the Kanban system is equivalent to a closed queueing network with finite queue space and *blocking*. As such, one would expect it to have less throughput than the same system with infinite queue space.

We also note that the above analysis considers systems in which the "value-added" nature of WIP is not considered, i.e., the cost of carrying WIP is largely the cost of decreased responsiveness as well as the attendant management overhead associated with tracking more parts. In such systems, the WIP level as represented by the number of pieces (as opposed to cost) is all that needs to be considered. However, in systems where the inventory carrying costs are significant and where value is added to the product in terms of increased variable cost (e.g., purchased components), the comparison is not as clear. For such cases, we must consider the WIP *value*. Since, for certain card assignments in the Kanban system, the WIP value may be less than that under CONWIP, there could be cases in which it is better to run Kanban than CONWIP. This occurs when one is essentially willing to sacrifice throughput (revenue) for the sake of reducing inventory investment. In such cases, and particularly when the bottleneck station is near the end of the line, a Kanban system could provide higher profit levels.

5. CONCLUSIONS

We have conjectured that less congestion results in pull systems because WIP levels are limited and WIP variability is reduced. Our demonstration of this for the exponential case provides some theoretical justification for this supposition. We also suggest that the effectiveness of pull systems does not result from pulling but from limiting WIP and WIP variability. Our comparison of a CONWIP system to a Kanban system offers credibility to this hypothesis. Finally, we have shown, both from a practical standpoint and with respect to an optimal policy, that a pull system is inherently easier to control than a push system.

We point out that there are many remaining issues to be resolved. Our congestion result was obtained using a push system that does not measure WIP to a pull system that does not measure throughput. An important area of further study is to determine ways to track throughput in a pull system to external demand. Other issues include defining routings in CONWIP systems and to determine the number and size of jobs to be used on those routings. Detailed simulation studies are needed to determine effective implementation strategies for CONWIP. Finally, an industrial test site is required before CONWIP can be considered as an alternative to Kanban or MRP.

APPENDIX

The Distribution of Cycle Times

For our comparisons regarding congestion we required that distribution of cycle time for both the push and

the pull systems be suitably approximated by a family of distributions whose mean is a location parameter and whose standard deviation is a scale parameter. This assumption affects our comparison of required lead times and average waiting inventory but not the comparison of WIP levels (which depends only on Little's law).

Note that the normal distribution fits our assumption and appears to work especially well for systems with cycle times whose mean is significantly larger than the standard deviation. The normal is clearly appropriate in push systems with a large number of exponential stations because response times in each station are independent. It also appears appropriate for some closed systems of suitable size as indicated by Wong (1979) and from our own simulation experience.

When making comparisons, however, a more subtle assumption is made; namely, that the distribution family for the push and the pull systems is the same. If the normal is appropriate for both systems, then this is not a problem. Otherwise, several of our results are not valid.

For distributions lacking the location and scaling property of the first two central moments, we can no longer write the required lead time as a simple linear function of the mean and standard deviation. We can, however, obtain bounds with these moments using Chebyshev's inequality,

$$P\{T \leq l\} = s.$$

Writing $l = \mu + z\sigma$ yields

$$P\left\{\frac{T-\mu}{\sigma} \leq z\right\} = s.$$

Chebyshev's inequality yields $1 - 1/z^2$ as a lower bound for service. If system A has a cycle time distribution with mean μ_1 and variance σ_1^2, and system B has a cycle time mean of μ^2 and variance of σ_2^2, then we can compute lead times l_1 and l_2 that will guarantee at least a service level of $1 - 1/z^2$. If $\mu_1 > \mu_2$ and $\sigma_1 > \sigma_2$, then $l_1 > l_2$.

The robustness of the average waiting inventory formula can be examined by comparing

$$\frac{\int_0^{l_1} (l_1 - t)\, dF_1(t)}{\int_0^{l_2} (l_2 - t)\, dF_2(t)} \text{ versus } \frac{\sigma_1}{\sigma_2}.$$

Writing l as $\mu + z\sigma$ allows the ratio of the integrals to be written as

$$\frac{\mu_1 + z\sigma_1 - \int_0^{l_1} t\, dF_1(t)}{\mu_2 + z\sigma_2 - \int_0^{l_2} t\, dF_2(t)},$$

and we note that the service levels are no longer the same for a given value of z. If, however, we let $s \to 1$ by allowing $z \to \infty$, the ratio becomes, σ_1/σ_2.

ACKNOWLEDGMENT

The authors thank Gabriel Bitran for his encouragement in this endeavor, and Uday Karmarkar for his insightful comments. We also acknowledge the contributions of Wallace Hopp in the development of CONWIP. Finally, we are grateful to two anonymous referees and the associate editor whose comments significantly improved both the clarity and the precision of the paper. This work has been supported, in part, by a grant from IBM Corporation and by grants DDM-8905638 and ECS-881103 from the National Science Foundation.

REFERENCES

BERGER, G. 1986. Managing Integrated Manufacturing Technologies. In *Proceedings 29th APICS Annual International Conference*, St. Louis.

BITRAN, G. R., AND L. CHANG. 1987. A Mathematical Programming Approach to a Deterministic Kanban System. *Mgmt. Sci.* **33,** 427–441.

BUZEN, J. P. 1973. Computational Algorithms for Closed Queueing Networks With Exponential Servers. *Comm. ACM* **16,** 527–531.

CHEN, H., M. HARRISON, A. MANDELBAUM, A. VAN ACKERE AND L. M. WEIN. 1988. Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication. *Opns. Res.* **36,** 202–215.

DENNING, P. J., AND J. P. BUZEN. 1978. The Operational Analysis of Queueing Network Models. *ACM Comput. Surv.* **10,** 255–261.

GORDON, W. J., AND G. F. NEWELL. 1967. Closed Queueing Systems With Exponential Servers. *Opns. Res.* **15,** 254–265.

HALL, R. W. 1983. *Zero Inventories.* Dow Jones-Irwin, Homewood, Ill.

KANET, J. J., AND D. P. CHRISTY. 1984. Manufacturing Systems With Forbidden Early Order Departure. *Int. J. Prod. Res.* **22,** 41–50.

KARMARKAR, U. S. 1986. Kanban Systems. Working Paper Series No. QM8612, Center for Manufacturing and Operations Management, The Graduate School of Management, University of Rochester, Rochester, N.Y.

KIMURA, O., AND H. TERADA. 1981. Design and Analysis of Pull System: A Method of Multi-Stage Production Control. *Int. J. Prod. Res.* **19,** 241–253.

KRAJEWSKI, L. J., B. E. KING, L. P. RITZMAN AND D. S. WONG. 1987. Kanban, MRP, and Shaping the Manufacturing Environment. *Mgmt. Sci.* **33,** 39–57.

MONDEN, Y. 1983. *Toyota Production System: Practical Approach to Management*. Industrial Engineering and Management Press, Norcross, Ga.

REISER, M., AND H. KOBAYASHI. 1975. Queueing Networks With Multiple Closed Chains: Theory and Computational Algorithms. *IBM J. Res. Dev.* **19,** 283–294.

REISER, M., AND S. S. LAVENBERG. 1980. Mean-Value Analysis of Closed Multichain Queueing Networks. *J. ACM* **27,** 313–322.

RITZMAN, L. P., AND L. J. KRAJEWSKI. 1983. Comparison of Material Requirements Planning and Reorder Point Systems. In *Simulation in Inventory and Production Control*, H. Bekiroglu (ed.). Society for Computer Simulation, La Jolla, Calif.

ROSS, S. 1983. *Stochastic Processes.* John Wiley, New York.

SCHONBERGER, R. J. 1986. *World Class Manufacturing: The Lessons of Simplicity Applied.* The Free Press, New York.

SHANTHIKUMAR, J. G., AND D. D. YAO. 1988. Second-Order Properties of the Throughput of a Closed Queueing Network. *Math. Opns. Res.* **13,** 524–534.

VOLLUM, R. 1984. Computer Integrated Manufacturing: A Matter of Survival. In *Proceedings 27th APICS Annual International Conference*, St. Louis.

WHITT, W. 1984. Open and Closed Models for Networks of Queues. *AT&T Bell Lab. Tech. J.* **63,** 1911–1978.

WIGHT, O. 1970. Input/Output Control a Real Handle on Lead Time. *Prod. Inv. Mgmt.* **11-3,** 9–31.

WONG, J. W. 1979. Response Time Distribution of the $M/M/m/N$ Queueing Model. *Opns. Res.* **27,** 1196–1202.

ZANGWILL, W. 1987. From EOQ Towards ZI. *Mgmt. Sci.* **33,** 1209–1223.