MANUFACTURING PERSPECTIVE

Practical Strategies for Lead Time Reduction

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

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208

In this paper we explore the causes of excessive lead time and suggest practical, inexpensive strategies for reducing it. Our recommendations are based on a detailed study of six manufacturing facilities, a survey of current literature, and our on-going research. After describing the relationship between WIP (work in process), mean flow time, flow time variance, and lead time, we systematically review potential methods for reducing lead time by reducing mean flow time and/or flow time variance.

INTRODUCTION

In a recent tour of a plant, we learned that the “raw process time” (the total labor and machine time) for the product added up to about 4 h whereas the average flow time (the time from job start until ready to ship) was around 4 weeks. This kind of disparity is not uncommon; in fact, it is all too often the rule. However, a few companies have reduced lead times (the time from when a customer order is taken until it is shipped) from weeks and months to hours and days. In recognition of its strategic importance, lead time reduction has become an important element of a campaign to increase the competitiveness of American manufacturing (see e.g., [1]).

The importance of lead and flow time reduction is highlighted by the fact that it is one of the few strategies that sales and production departments can agree on. From the perspective of sales, shorter lead times:

1. Offer the ability to quote faster delivery to customers.
2. Lessen the impact of cancelled orders.
3. Reduce the need to make forecasts about future demand.

From the production side, shorter flow times:

1. Improve quality management by reducing the opportunity for work to be damaged and shortening the time between manufacturer and defect detection.
2. Reduce in-process inventories.
3. Decrease disruption of the production process due to engineering change orders.
4. Enable shorter frozen zones in the Master Production Schedule, thereby reducing the dependence on distant forecasts.
5. Allow easier overall management of the facility because there will be fewer jobs to keep track of and fewer special cases (e.g., expedited jobs) to oversee.

Of course, the effects of shorter lead times can only be considered beneficial if they can be achieved without undue sacrifices in other areas such as quality and throughput. Nevertheless, because the benefits are so attractive, many specific suggestions for reducing lead times have been made (see e.g., [2-6]). This previous work has done a great deal to help decrease lead time in specific instances. Additionally, many of the techniques found in the just-in-time (JIT) philosophy also result in reduced lead times. These are discussed in the many good books written on JIT, including [1,7-11]. This paper, however, focuses directly on lead times and on the causes of lead time inflation. In this light, we review lead time reduction techniques found in the JIT literature with an eye toward efficacy and underlying concepts. Most of the techniques we discuss are inexpensive and fairly simple.
Our recommendations are based on our experience with various manufacturing facilities, a survey of the JIT literature, and our on-going research. We have conducted detailed audits of the manufacturing control strategies of six production facilities. These facilities are owned by companies ranging from one-plant firms to giant multinationals. Their operations range from pure fabrication to fabrication-plus-assembly and from job shop through completely repetitive manufacturing. The motivation for lead time reduction varied substantially among them. In one small firm, lead time reduction was needed simply to reduce the amount of precious capital allocated to WIP (work in process). In other firms the needs were strategic—shorter lead times were aimed at gaining a competitive advantage.

With this paper, we also hope to clear up some confusion that has resulted with the reduction of the JIT philosophy into quick "sound bites." For example, the statement, "lead time can be divided into value-added plus non-value-added operations" has led more than one company to attempt lead time reductions by eliminating all labor and machining operations that do not contribute directly to the value of the product. However, we find the relationship between lead time and the "value-added" by operations to be much more subtle than these definitions imply. The whole value-added versus non-value-added controversy can be confusing and even sterile when it comes to reducing lead times. To support this contention, we describe a case where we were able to reduce lead time by increasing non-value-added labor time.

We have found that there are three key points involved in lead time reduction. First, the major components of flow time (and hence lead time) are queueing time and waiting time. Practical strategies for reducing lead times must attack these components to achieve significant results. Second, WIP and flow time are proportional to each other for a given level of throughput. Consequently, causes of excessive lead time can be determined by identifying locations with large inventories. Finally, lead time is related not only to the average of flow time but also to the variance of flow time. Although most strategies that reduce the average flow time also reduce its variance, there are situations where this is not true and lead times may actually increase. We must therefore be aware of the impact of a proposed strategy on both performance measures.

METHODS FOR REDUCING LEAD TIME

We now turn to the main purpose of this paper—identifying simple strategies for reducing lead time. Our strategies fall into five general categories: (1) look for the WIP, (2) keep things moving, (3) synchronize production, (4) smooth the work flow, and (5) eliminate variability. For each of these categories we relate the corresponding JIT philosophy and describe specific instances of implementation. Many of these can be achieved with simple low-cost changes in the production system. The extent to which each is applicable to a specific situation should be apparent in the discussion. In this paper, we focus on the lead time reduction strategies themselves. In a companion paper [12], we develop mathematical models for quantifying the effect of many of these strategies on flow time mean and variance.

Look for the WIP

Suzuki [11] states succinctly that "excess inventory is the root of all evil" (emphasis ours). However, Goldratt [13] points out the importance of inventory to "protect the bottleneck." Thus, extra inventory at a bottleneck should not be considered excessive if it is needed to protect throughput.

To understand the relation between inventories and lead time, first note that flow times and inventories are not independent and that, for a given production rate, the two are directly related. For any given cause, the added inventory and added flow time due to that cause are related by

\[ \text{added inventory} = \text{production rate} \times \text{added flow time} \]

This expression is known as "Little's law" [14] and is widely applicable to almost any queuing situation [15]. For example, if assembly is often held up because not all parts are available, the inventory of parts waiting will be equal to the production rate multiplied by the average waiting time of the parts. Thus, a means of identifying the largest components of lead time is to find the largest inventories (both finished goods and work in process) and work to reduce them [16]. By the same token, any action that reduces flow times will also reduce inventories.

Little's law gives us a relationship between mean flow time and inventories. In many cases, the effects of the variance of flow time on inventories should also be considered. Under reasonable assumptions, it can be shown that, for fixed values of throughput and service, the average waiting inventory (i.e., inventory waiting to be used according to an assembly or shipping schedule) increases linearly with the standard deviation of flow time and does not depend on the mean flow time. A mathematical demonstration of this relationship is given in the appendix; we give an illustration of the reasons here.

Figure 1 shows an example of two flow time densities with equal means and different standard deviations. For the case with the smaller standard deviation, we see that 99% of the jobs finish with flow times that are not greater than 15 days, whereas the larger standard deviation case requires 27 days for the same service level. Note also that the most likely time for the smaller standard deviation is 11 days. Thus, with a lead time of 15 days, jobs are most likely to finish 4 days early yielding a waiting inventory of four days worth of production (Little's law again). In the other case, the most likely flow time is 8 days, yielding a 19-day stay in waiting inventory. Although this case is extreme, it does show that both lead times and waiting inventories depend on the variability associated with flow time and not average flow time alone.

![FIG. 1. Two densities of flow time with unequal variance](image_url)
Although most actions that reduce mean flow time also reduce its variance, there are some cases where the opposite is true. We will discuss one such example in the section on synchronizing production.

Keep Things Moving

Schonberger [1] describes how the JIT approach at IBM came to be known as “continuous flow manufacturing” or CPM. The basic idea is if the product is always moving toward completion, both flow times and inventories will decrease. This is particularly important since 90-95% of the time spent in a factory is spent waiting [17]. Breaking the flow time down into its components reveals why:

flow time = run time + setup time + move time + queue time + wait-for-parts time + wait-to-move time

Run time is the total processing time at work centers required to complete the job. Setup time is the sum total of all of the internal setups involved in processing the job. Move time is the time required to move the job between work centers. Queue time is the time spent waiting in line for work centers to become available. Wait time has two components: wait-for-parts time, which is time spent waiting for other subassemblies so that an assembly operation can begin, and wait-to-move time, which is time spent waiting for other parts in a batch to be completed so the batch can be moved to the next work center. Note that a job waiting for a resource to accomplish the move, such as a forklift, is not incurring wait-to-move time in our terminology. This situation is exactly analogous to waiting for a machine for processing and hence is appropriately included in queue time.

The above breakdown of flow time is the same as that given by Wight [18], except that the two components of wait time are broken out explicitly. This is important because the two forms of delay have very different causes and hence are amenable to different types of remedy.

Since total run, setup, and move times typically make up only a fraction of the total flow time, a large percentage is made up of waiting in queues, waiting for parts, and waiting to move. Thus, it makes sense to focus efforts to reduce the flow time associated with these components. Buying equipment to increase production rates or undertaking setup time reduction programs are associated with significantly reduce average flow time unless queueing and waiting times have been reduced to extremely low levels. This is not to deprecate conventional JIT wisdom that setup time reduction is an important goal [7,9,11]. Setup time reduction is important for increasing capacity and in reducing flow time variance, but does little to decrease mean flow time. There are, however, some simple strategies that are effective.

Splitting Jobs. Where setup times cannot be reduced to insignificant levels, large batch sizes may be required at some work centers to achieve needed capacity. However, it is not always necessary to use large batch sizes on nonbottleneck operations. If a nonbottleneck operation can keep up with the rate dictated by the bottleneck with lot sizes of 1, then a lot size of 1 is appropriate. Otherwise, lot sizes should be as small as practical. Following the lead of Goldratt, many authors are beginning to make the distinction between a process batch and a transfer batch (see e.g., [19]). At the bottleneck, where capacity is critical, reducing the lot size (process batch) may not be practical. However, the lot size processed by the bottleneck does not have to equal the lot size that is transferred (transfer batch). Forcing the entire lot to wait until the last piece is finished can be a significant source of wait time. Therefore, large lots should be used only in front of bottlenecks. Elsewhere, the process lot should be split into transfer lots as small as can be practically handled. This process is sometimes called “overlapping” in the JIT literature [1]. Good plant layout (e.g., work cells) can make frequent moves easy to accomplish [8].

We were recently reminded of the importance of transfer lot size when one of the plants in our study set up a manufacturing cell that used flexible equipment to reduce setup times in an attempt to shorten lead times for crankshafts. Unfortunately, the lot size used was about 10,000 and was not split for the purpose of moving work. As a result, although the time to complete the first crankshaft was markedly reduced, the average flow time for the entire lot changed very little.

Sharing Transfer Batches (or The Value of “Non-Value-Added”). Many times an inherently slow process will require many parallel tools to provide sufficient capacity. If incoming transfer batches are assigned to individual tools, lead time can be excessive. If the batches are shared by more than one tool, the average flow time can be reduced. Multiple tools will complete the work on the shared transfer batch faster than a single tool and the entire batch can then move more quickly to the next work center.

Of course, sharing transfer batches may require additional material handling as the shared batch may not be near all of the work stations. It may seem somewhat counterintuitive that adding operations that are clearly “non-value-added” will result in reduced lead times, but it is clearly the case. Adding such operations, however, can reduce the effective capacity of the work center and may require the hiring of an additional worker to serve in a material handling function. Whether to do this represents an obvious tradeoff between increased labor cost and shorter lead times.

In some cases, however, the workers are not highly utilized. In one of the plants studied that produced raw (i.e., with no components) circuit boards, decreasing flow time had become an important priority. The company had just undertaken a “non-value-added reduction drive” and flow times remained excessive. The sequence of operations to circuitize a board involved two conveyor type operations separated by a manual expose operation. Because this operation was necessarily long, there were 32 expose machines operating in parallel. Boards were collected at the end of the first conveyor in carts and then moved to the expose operation, with one cart per expose machine. The flow time through the entire process center was about 2 days. Although reluctant at first, the plant manager agreed to utilize a single outbound cart in the expose area, reducing the average wait-to-move time to about 1/12th of its original value. The result was a 30% reduction in flow time for the process center with no increase in head count.

By sharing carts, the manager has increased the number of “no-value-adding steps, that is, more move operations. This appears to be in conflict with some conventional JIT slogans. However, perhaps it is more precise to say that conventional JIT slogans are not always consistent with the JIT wisdom of Ohno [10, p. ix] who advises that we remove “the non-value-added wastes.” The key word is waste. If we examine the cart sharing example more closely we see that by adding the waste of “material handling” we have eliminated the waste of “wait-to-move time.” A different solution might be to place the machines into closer proximity or to provide some type of material handling device that would eliminate the additional handling. In our case, however, such costly modifi-
cations were not warranted since the operation was not the bottle-
neck of the line and the operators could easily keep up with the re-
quired production rates. To summarize, we recognize that making
needless costly modifications is also a form of “waste.”

Queue Control. An obvious means to reduce queue time is to
maintain shorter queues. In a Kanban \[7,9,11\] or CONWIP \[20,21\]
system this is straightforward since, in these pull systems, WIP is
controlled directly. If such a system is used, WIP can be reduced
until further reductions would lead to disruptions. If the cause of
the potential disruption can be determined and eliminated, then
queues can be further reduced. For instance, if WIP reduction
results in starvation at a bottleneck and there are unreliable or
extremely variable machines in the line, the system can probably be
improved by addressing these problems. If this variability cannot
be reduced, extra WIP must be maintained.

In an MRP system, controlling queues is more difficult. How-
ever, techniques such as input/output control \[18\] do provide added
control of queue lengths. The WIP levels in each work center
should be evaluated to see where reductions can be made in order
to reduce safety lead times. Also, production releases coordinated
to bottleneck rates will avoid producing parts that are not immen-
tly needed.

There is a limit to how far WIP levels can be reduced before
throughput will invariably be decreased. “Zero inventories” will
result in zero throughput, regardless of the popularity of the slogan.
A recent paper by Conway et al. \[22\] indicates that most U.S.
firms operate with far more WIP than is needed. They suggest that
the WIP that is present be placed in front of the bottleneck for
unbalanced lines and near the middle for balanced lines.

Synchronize Production

Because a part assembly cannot be completed until all components
are available, synchronization between fabrication and assembly is
extremely important. The synchronization of component fabrica-
tion with assembly schedules is also difficult. It is quite common
for an assembly line to produce based on what is available rather
than what is needed. Harmon and Peterson liken productivity
improvement opportunities in assembly systems to the proverbial
pot of gold at the end of the rainbow \[8\]. In this same vein, efforts to
synchronize component fabrication with assembly can have
dramatic effects on waiting inventories and therefore lead times.

In Kanban and CONWIP systems, synchronization occurs
naturally, since work is pulled only as needed. MRPII systems
would, in theory, also synchronize production if flow times were
constant. However, because flow times are not constant, the lead
times used to compute job release times from due dates are inflated
to values much larger than average flow times to accommodate
capacities. Such practice builds inefficiency into MRPII
systems because inflated lead times cause jobs to finish early (on
average), thereby increasing waiting inventory.

The problem of deciding which jobs should be performed next
takes place in the shop floor control module in the MRPII hier-
carchy and is associated with “dispatching.” A commonly used tech-
nique involves a “dispatching rule” which allows an operator or
foreman to choose the next job by considering those currently wait-
ing in queue. Typical dispatching rules are “shortest processing
time,” “earliest due date,” and “critical ratio,” a rule that involves
both job size and due date \[23\] for a survey of dispatching rules.

It is currently popular in the academic literature to propose
new, presumably better, dispatching rules. Unfortunately, many of
these rules are better in the sense that they improve some myopic
measure of performance which considers only fabrication lines.
Use of rules which do not take into account their effect on waiting
inventory may cause tremendous increases in flow time for finished
assemblies.

The SPT rule (process the job with the shortest processing
time first) is a good example. This rule will provide the best per-
formance for a single machine in the sense that average flow time
for all parts will be as low as possible. Simulation studies have
also shown SPT to be effective in some job shops \[24\]. Unfortu-
nately, performing the smallest job first will have the obvious effect
of increasing the variance of flow time; large jobs simply wait
longer. Since lead times depend not only on average flow time but
also on flow time variance, implementation of SPT could have the
effect of increasing lead times.

Increases in flow time variance are particularly disastrous in
an assembly system. If an assembly requires two components, one
which requires a long processing time and one with a short
processing time, then SPT will result in excessive waiting inven-
tory. An example of a dispatching rule which does take into
account the effect on waiting inventory is EDD (process the job
with the earliest due date first). The corresponding rule for pull
systems is “first in system, first served” (process the job request
that arrived first).

A second problem is to match production rates of fabrication
with consumption rates at assembly. Kanban, OPT \[19\], and
CONWIP systems will naturally equalize these production rates
even in unbalanced lines. Many of the problems with excessive
lead times which we have encountered are caused by the perceived
need to “keep busy.” In one of the plants studied, we observed
workers who run their machines even when parts are being
produced much faster than they are being used at downstream
machines; they stop only when there is no physical space left for
additional WIP. Clearly, it is important for management to under-
stand the need to balance flow and to communicate this to workers
in the form of appropriate performance measures.

One important point to note is that synchronization becomes
much easier as the variability in the production process is reduced.
If production times are very predictable, it will be unnecessary to
build in large safety lead times and subassemblies will be much
more likely to arrive at assembly simultaneously. Hence, measures
that reduce variability of flow time are important in improving
synchronization.

Smooth the Work Flow

Almost all of the works on JIT implementation describe the need to
smooth work flow in order to reduce inventories and lead times
\[7,9,10,11\]. We list some easily implemented methods.

Leveling Work Releases. The goal of leveling work releases is
to maintain an even workload. Low workloads lead to small
queues and quick turnaround, while high workloads cause long
queues and long flow times. Overall, uneven workloads increase
both flow time mean and variance. When discussing workloads, it
is important to note that the time scale we are considering is on the
order of days and not months as in discussions of using a chase or
level aggregate production plan \[e.g., 25, p. 214\]. Gradual
increases in release rates, accompanied by increases in capacity,
can facilitate a chase production strategy. What is important is to
release frequently and in small amounts. That way, there are
smaller queues at release points and less likelihood for a “bulge” of
WIP that alternately overloads and starves production resources.

Ultimately, releases should be tied to production at bottleneck
work centers. In this way, a “pipeline” to the bottleneck is

Hopp, Spearman and Woodruff: Practical Strategies for Lead Time Reduction
established and is always full. More frequent releases only create WIP in front of the bottleneck with no increase in production. Also, planning changes in the form of order cancellations and engineering change orders are less disruptive is the work is not yet released than if the work is already in process and somewhere on the floor. In pull systems such as Kanban and CONWIP, the release rates (pull rates) are naturally equal to downstream production rates. In push systems, such as MRP this is more difficult, although more frequent releases (such as twice daily rather than weekly) combined with some form of WIP control (e.g., input/output control) can do a great deal to reduce flow time variance and congestion on the shop floor. The drum-buffer-rope system proposed by Goldratt [13] is an example of a push system that directly ties work releases to bottleneck production.

**Establishing a Uniform Work Flow.** Changes in the product mix lead to setups, changeovers, and disruption of factory rhythm. By establishing routings for families of parts [26] so that setups between families are avoided, and by reducing setup times for the instances where setups cannot be avoided [9, pp. 76-84] many of these problems can be alleviated. Also, since some changes are inevitable, cross-training of workers will minimize the disruption of these changes.

Product standardization, involving the use of common lower levels, can reduce the number of different parts that must be produced, tracked, and inventoried. Also, by customizing the product late in production the system is less sensitive to changes in customer orders. Both of these changes were particularly relevant for one of the firms that we studied. This company conducted a product standardization review that focused on limiting the range of customization options and developed ways to delay customization until the end of the manufacturing process. This strategy enabled them to offer maximum flexibility to their customers at little cost to themselves.

**Rationalizing Line Balancing.** We have found that systems with all centers operating independently and at similar rates will tend to have longer average flow times than systems with a distinct bottleneck [27]. Any time one work center works somewhat faster than the preceding work center (due to random variations), it will become starved unless WIP is excessive. Because of this dependency, the periods of high production do not make up for low periods. Although the usual reason given for balancing a production line is to "break bottlenecks" the practice has the opposite effect—all stations become bottlenecks and therefore each must be protected with additional WIP.

This is not to suggest that line balancing should not be employed on paced assembly lines. On such lines, the rhythm of the system prevents starvation and WIP buildup. Line balancing was never intended to balance capacity between work centers. Goldratt [28] states this succinctly, "balance the flow, not the capacity." During the inevitable slack periods at nonbottleneck stations, preventive maintenance and other housekeeping tasks can be performed.

**Eliminate Variability.**

Unlike Suzuki, Schonberger [1] insists that the root of all evil is variability. Variability in processing times caused by rework, downtime, and lack of consistency in production methods increase both mean and the variance of flow time. Unlike inventory, there are no instances where more variability is good.

Some strategies for reducing variability include the following.

**Reducing Rework.** Rework operations can have a tremendous impact on flow time mean and variance, particularly if jobs requiring rework compete with regular jobs for resources. Tight quality control and SPC should be used to eliminate rework wherever possible. Where large lot sizes are used, quality checks should be made before completion of the lot. This will help avoid the situation where the entire lot is completed before it can be checked. Where possible, quality checks should be located in front of the bottleneck or other time-consuming operations. Quality control in front of the bottleneck will avoid wasting its important capacity on defective parts. Quality checks in front of lengthy operations will help decrease the amount of rework. If rework is disruptive and means for eliminating it cannot be found, a dedicated line for rework may be required. This will help smooth the flow of work through the facility and speed the recombinant of jobs requiring rework with the rest of the jobs. However, rework lines can also serve as "psychological crutches" and remove some of the stigma attached to quality problems. Elimination of the problems at the source should always be considered before resorting to rework lines.

**Improving Machine Reliability.** An important source of flow time variance can be machine down time. Although many companies track machine availability (A), some do not track the mean time between failures (MTBF) and the mean time to repair (MTTR). The relation between these quantities is

\[
A = \frac{MTBF}{MTBF + MTTR}
\]

Although only availability is needed to determine the capacity of a machine, it is not sufficient to determine the overall congestion in the line. Long and infrequent outages are more disruptive than short and frequent ones even if the availabilities are equal. This is because short outages require relatively small buffer stocks between machines to prevent starvation whereas longer outages require correspondingly more. Of course, downtime at bottleneck resources is particularly important, not only because of its effect on system capacity, but also because it has a strong effect on flow time variance. Regularly scheduled preventive maintenance can be used to minimize machine downtimes. Replacing or overhauling unreliable machines, particularly if they are at the bottleneck, can also be used to reduce flow time variance.

We developed a queueing model of the production facility of one of our clients in order to estimate the impact of improving MTBF and MTTR. With this model we have been able to justify variability reduction projects that would have been impossible to justify using ordinary accounting models.

**Planning for Yield Losses.** Although the best strategy to deal with yield loss is to eliminate it, this is not always possible. A reasonable alternative is to consider carefully an appropriate strategy for dealing with yield loss.

A common procedure for deciding on the size of a job to start, given a desired finish quantity and some knowledge of yield loss, is to inflate the job size by dividing the desired quantity by the average yield (i.e., the proportion of good parts that are completed). If either the yields are low or the job sizes are large, it is unlikely that the resulting finish quantity will be equal to the desired quantity. However, by applying the above simple strategy, we imply that to end up with one piece short is just as bad as having one piece too many. This is seldom the case, since short pieces must be expedited through the factory to catch up with the rest of the job whereas additional pieces typically sit in finished goods inventory until the next order. Both outcomes are undesirable but the first usually has more severe consequences. The flow time clearly increases for the order involved and for other jobs as well if setups must be broken to
facilitate the expediting. Rework also increases the variance of flow time because of the congestion it causes.

In one of the plants we studied, we proposed a change to the way jobs were "yielded." Each job was started with a quantity large enough such that the probability of having sufficient parts was equal to a prespecified value, say 90%. This procedure resulted in a slight increase in finished goods while greatly reducing the number of small, "hot" jobs that had been both robbing capacity (through more setups) and increasing lead time.

**Vendor Variability.** Variability is an important factor in purchased, as well as fabricated, parts. One of the firms in our study assembled final products almost entirely from purchased parts. Prior to our involvement, they had based their decisions on how much buffer lead time to build into the schedule entirely on cost (i.e., inexpensive parts were brought in well ahead of the times they were needed, whereas expensive parts were ordered with very little slack in the schedule). After our review, we concluded that variance of vendor lead time was as important as the cost of the part (i.e., highly variable vendors require more buffer lead time than very dependable ones). On the basis of this realization, we set up a system for establishing appropriate buffer lead times and tracking vendor performance that directly incorporated variance.

**CONCLUSION**

We have argued that the key methods for reducing lead time are those that reduce mean flow time and flow time variance. As indicated in our discussion, there are many ways to accomplish these reductions with little or no cost.

We have also seen the danger of implementing the JTT philosophy through the use of slogans. We have seen instances of "good" inventory and valuable "non-value-added" operations. However, we do believe in one slogan. *Don't look back; you never know who's gaining on you.*

**REFERENCES**


**APPENDIX**

To understand completely the relationships between lead times, flow times, WIP, and finished goods inventories, we consider the production of a part between two adjacent levels within a product structure. This production could represent fabrication of a part needed for an assembly operation (in a multilevel bill of material) or the completion of an entire part from raw materials to finished goods (as in a steel mill). Suppose 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 $l$ has elapsed will wait in inventory. This is not unrealistic, since at assembly operations 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. We represent the average of this waiting inventory as $\bar{W}$. Finally, we let $\bar{W}$, $\bar{T}$, and $\Theta$ designate, respectively, the average work in process (not including waiting inventory), flow time, and throughput for the line. Then from Little’s law we see that

$$\bar{W} = \bar{T} \Theta$$

We further assume that the distribution of the flow time $T$ can be approximated by a distribution whose mean serves as a location parameter and is equal to $T$ and whose standard deviation functions as a scale parameter and is equal to $\sigma_T$. Examples of such distributions include the normal, uniform, and symmetrical triangular. We denote this distribution function as $F$ and define $Z = (T - T)/\sigma_T$ having a distribution function $F_T$ such that $F_T(0) = s$. Then the necessary lead time for a given service level $s$ will be,

$$l = \sigma_T^2 z + T$$

Hence, lead time increases linearly in both average flow time and the standard deviation of flow time. Let $\phi$ be the density corre-
sponding to $\Phi$. In cases where the normal is a good approximation, the average waiting inventory $I_w$ will be

\[
I_w = \bar{\theta} E[\max(0, l - T)]
\]
\[
= \bar{\theta} \int_0^l F(t) \, dt
\]
\[
= \sigma_r \bar{\theta} \int_{-\infty}^{z_r} \phi(z) \, dz
\]
\[
= \sigma_r \bar{\theta} [\Phi(z_r) + \Phi(z_r)]
\]

The quantity in the brackets in the last equation is constant for any given service level. Although this model is too simple to represent a real production facility, it does point out that, for fixed values of throughput and service, the expected waiting inventory increases linearly with the standard deviation of the flow time but does not depend on the mean flow time.

Hence, as we stated above, reduction of average flow times allow for smaller reorder points, more competitive lead times, a shorter frozen zone and thereby more flexibility. The smallest possible value for average flow time is the raw processing time and this occurs only when the variance of the processing times are zero [24]. Small variance in flow time will reduce waiting inventories as well as lead times.

**Wallace J. Hopp** is an associate professor in the Department of Industrial Engineering and Management Sciences at Northwestern University and performs research in probability models of manufacturing systems, forecast horizons in nonhomogeneous Markov decision processes, and the development of alternative production control systems. He teaches courses in stochastic processes and a senior design sequence. He has research articles in *Operations Research, Management Science, IIE Transactions, Journal of Optimization Theory and Applications, Mathematical Programming, Naval Research Logistics Quarterly, Queueing Systems,* and the *International Journal of Production Research*. He is a senior member of IIE, a full member of ORSA, and a member of TIMS and SIAM. Professor Hopp is also an active consultant to manufacturing firms.

**Mark L. Spearman** is an assistant professor in the Department of Industrial Engineering and Management Sciences at Northwestern University where he teaches in the areas of production and inventory control and systems simulation. Professor Spearman’s research addresses the need for better manufacturing models and ways to incorporate JIT techniques within hierarchical production planning systems. This research has been supported by grants from both private industry and the National Science Foundation. He is an area editor of *Production and Operations Management* and has research papers appearing in and/or has served as a referee for *Engineering Economist, IIE Transactions, International Journal of Flexible Manufacturing Systems, International Journal of Production Research, Journal of Manufacturing and Operations Management, Management Science* and *Operations Research*. He is a senior member of IIE, a full member of ORSA, a member of TIMS, POMS and is certified in production and inventory by APICS.

**David L. Woodruff** is currently a research associate at Northwestern University but soon will be an assistant professor at the University of California at Davis. His past work has dealt with sequencing jobs in a generalized pull system. Mr. Woodruff has research articles appearing in *International Journal of Production Research* and the *Journal of Manufacturing and Operations Management*. He has eight years experience in software development and general management. He is a member of ORSA, TIMS, IIE and POMS.
MANUFACTURING PERSPECTIVE

Practical Strategies for Lead Time Reduction

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

Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208

In this paper we explore the causes of excessive lead time and suggest practical, inexpensive strategies for reducing it. Our recommendations are based on a detailed study of six manufacturing facilities, a survey of current literature, and our ongoing research. After describing the relationship between WIP (work in process), mean flow time, flow time variance, and lead time, we systematically review potential methods for reducing lead time by reducing mean flow time and/or flow time variance.

INTRODUCTION

In a recent tour of a plant, we learned that the “raw process time” (the total labor and machine time) for the product added up to about 4 hours whereas the average flow time (the time from job start until ready to ship) was around 4 weeks. This kind of disparity is not uncommon; in fact, it is all too often the rule. However, a few companies have reduced lead times (the time from when a customer order is taken until it is shipped) from weeks and months to hours and days. In recognition of its strategic importance, lead time reduction has become an important element of a campaign to increase the competitiveness of American manufacturing (see e.g., [1]).

The importance of lead and flow time reduction is highlighted by the fact that it is one of the few strategies that sales and production departments can agree on. From the perspective of sales, shorter lead times:

1. Offer the ability to quote faster delivery to customers.
2. Lessen the impact of cancelled orders.
3. Reduce the need to make forecasts about future demand.

From the production side, shorter flow times:

1. Improve quality management by reducing the opportunity for work to be damaged and shortening the time between manufacturer and defect detection.
2. Reduce in-process inventories.
3. Decrease disruption of the production process due to engineering change orders.
4. Enable shorter frozen zones in the Master Production Schedule, thereby reducing the dependence on distant forecasts.
5. Allow easier overall management of the facility because there will be fewer jobs to keep track of and fewer special cases (e.g., expedited jobs) to oversee.

Of course, the effects of shorter lead times can only be considered beneficial if they can be achieved without undue sacrifices in other areas such as quality and throughput. Nevertheless, because the benefits are so attractive, many specific suggestions for reducing lead times have been made (see e.g., [2-6]). This previous work has done a great deal to help decrease lead time in specific instances. Additionally, many of the techniques found in the just-in-time (JIT) philosophy also result in reduced lead times. These are discussed in the many good books written on JIT, including [1,7-11]. This paper, however, focuses directly on lead times and on the causes of lead time inflation. In this light, we review lead time reduction techniques found in the JIT literature with an eye toward efficacy and underlying concepts. Most of the techniques we discuss are inexpensive and fairly simple.
Our recommendations are based on our experience with various manufacturing facilities, a survey of the JIT literature, and our on-going research. We have conducted detailed audits of the manufacturing control strategies of six production facilities. These facilities are owned by companies ranging from one-plant firms to giant multinationals. Their operations range from pure fabrication to fabrication-plus-assembly and from job shop through completely repetitive manufacturing. The motivation for lead time reduction varied substantially among them. In one small firm, lead time reduction was needed simply to reduce the amount of precious capital allocated to WIP (work in process). In other firms the needs were strategic — shorter lead times were aimed at gaining a competitive advantage.

With this paper, we also hope to clear up some confusion that has resulted with the reduction of the JIT philosophy into quick “sound bites.” For example, the statement, “lead time can be divided into value-added plus non-value-added operations” has led more than one company to attempt lead time reductions by eliminating all labor and machining operations that do not contribute directly to the value of the product. However, we find the relationship between lead time and the “value-added” by operations to be much more subtle than these definitions imply. The whole value-added versus non-value-added controversy can be confusing and even sterile when it comes to reducing lead times. To support this contention, we describe a case where we were able to reduce lead time by increasing non-value-added labor time.

We have found that there are three key points involved in lead time reduction. First, the major components of flow time (and hence lead time) are queuing time and waiting time. Practical strategies for reducing lead times must attack these components to achieve significant results. Second, WIP and flow time are proportional to each other for a given level of throughput. Consequently, causes of excessive lead time can be determined by identifying locations with large inventories. Finally, lead time is related not only to the average of flow time but also to the variance of flow time. Although most strategies that reduce the average flow time also reduce its variance, there are situations where this is not true and lead times may actually increase. We must therefore be aware of the impact of a proposed strategy on both performance measures.

METHODS FOR REDUCING LEAD TIME

We now turn to the main purpose of this paper — identifying simple strategies for reducing lead time. Our strategies fall into five general categories: (1) look for the WIP, (2) keep things moving, (3) synchronize production, (4) smooth the work flow, and (5) eliminate variability. For each of these categories we relate the corresponding JIT philosophy and describe specific instances of implementation. Many of these can be achieved with simple low-cost changes in the production system. The extent to which each is applicable to a specific situation should be apparent in the discussion. In this paper, we focus on the lead time reduction strategies themselves. In a companion paper [12], we develop mathematical models for quantifying the effect of many of these strategies on flow time mean and variance.

Look for the WIP

Suzuki [11] states succinctly that “excess inventory is the root of all evil” (emphasis ours). However, Goldratt [13] points out the importance of inventory to “protect the bottleneck.” Thus, extra inventory at a bottleneck should not be considered excessive if it is needed to protect throughput.

To understand the relation between inventories and lead time, first note that flow times and inventories are not independent and that, for a given production rate, the two are directly related. For any given cause, the added inventory and added flow time due to that cause are related by

\[
\text{added inventory} = \text{production rate} \times \text{added flow time}
\]

This expression is known as “Little’s law” [14] and is widely applicable to almost any queueing situation [15]. For example, if assembly is often held up because not all parts are available, the inventory of parts waiting will be equal to the production rate multiplied by the average waiting time of the parts. Thus, a means of identifying the largest components of lead time is to find the largest inventories (both finished goods and work in process) and work to reduce them [16]. By the same token, any action that reduces flow times will also reduce inventories.

Little’s law gives us a relationship between mean flow time and inventories. In many cases, the effects of the variance of flow time on inventories should also be considered. Under reasonable assumptions, it can be shown that, for fixed values of throughput and service, the average waiting inventory (i.e., inventory waiting to be used according to an assembly or shipping schedule) increases linearly with the standard deviation of flow time and does not depend on the mean flow time. A mathematical demonstration of this relationship is given in the appendix; we give an illustration of the reasons here.

Figure 1 shows an example of two flow time densities with equal means and different standard deviations. For the case with the smaller standard deviation, we see that 99% of the jobs finish with flow times that are not greater than 15 days, whereas, at the larger standard deviation case requires 27 days for the same service level. Note also that the most likely time for the smaller standard deviation is 11 days. Thus, with a lead time of 15 days, jobs are most likely to finish 4 days early yielding a waiting inventory of four days worth of production (Little’s law again). In the other case, the most likely flow time is 8 days, yielding a 19-day stay in waiting inventory. Although this case is extreme, it does show that both lead times and waiting inventories depend on the variability associated with flow time and not average flow time alone.

![FIG. 1. Two densities of flow time with unequal variance](image)
Although most actions that reduce mean flow time also reduce its variance, there are some cases where the opposite is true. We will discuss one such example in the section on synchronizing production.

**Keep Things Moving**

Schonberger [1] describes how the JIT approach at IBM came to be known as “continuous flow manufacturing” or CFM. The basic idea is if the product is always moving toward completion, both flow times and inventories will decrease. This is particularly important since 90-95% of the time spent in a factory is spent waiting [17]. Breaking the flow time down into its components reveals why:

\[
\text{flow time} = \text{run time} + \text{setup time} + \text{move time} + \text{queue time} + \text{wait-for-parts time} + \text{wait-to-move time}
\]

Run time is the total processing time at work centers required to complete the job. Setup time is the sum total of all of the internal setups involved in processing the job. Move time is the time required to move the job between work centers. Queue time is the time spent waiting in line for work centers to become available. Wait time has two components: wait-for-parts time, which is time spent waiting for other subassemblies so that an assembly operation can begin, and wait-to-move time, which is time spent waiting for other parts in a batch to be completed so that the batch can be moved to the next work center. Note that a job waiting for a resource to accomplish the move, such as a forklift, is not incurring wait-to-move time in our terminology. This situation is exactly analogous to waiting for a machine for processing and hence is appropriately included in queue time.

The above breakdown of flow time is the same as that given by Wight [18], except that the two components of wait time are broken out explicitly. This is important because the two forms of delay have very different causes and hence are amenable to different types of remedy.

Since total run, setup, and move times typically make up only a fraction of the total flow time, a large percentage is made up of waiting in queues, waiting for parts, and waiting to move. Thus, it makes sense to focus efforts to reduce the flow time associated with these components. Buying equipment to increase production rates or undertaking setup time reduction programs are unlikely to significantly reduce average flow time unless queueing and waiting times have been reduced to extremely low levels. This is not to deprecate conventional JIT wisdom that setup time reduction is an important goal [7,9,11]. Setup time reduction is important for increasing capacity and in reducing flow time variance, but does little to decrease mean flow time. There are, however, some simple strategies that are effective.

**Splitting Jobs.** Where setup times cannot be reduced to insignificant levels, large batch sizes may be required at some work centers to achieve needed capacity. However, it is not always necessary to use large batch sizes on nonbottleneck operations. If a nonbottleneck operation can keep up with the rate dictated by the bottleneck with lot sizes of 1, then a lot size of 1 is appropriate. Otherwise, lot sizes should be as small as practical. Following the lead of Goldratt, many authors are beginning to make the distinction between a *process batch* and a *transfer batch* (see e.g., [19]). At the bottleneck, where capacity is critical, reducing the lot size (process batch) may not be practical. However, the lot size processed by the bottleneck does not have to equal the lot size that is transferred (transfer batch). Forcing the entire lot to wait until the last piece is finished can be a significant source of wait time. Therefore, large lots should be used only in front of bottlenecks. Elsewhere, the process lot should be split into transfer lots as small as can be practically handled. This process is sometimes called “overlapping” in the JIT literature [1]. Good plant layout (e.g., work cells) can make frequent moves easy to accomplish [8].

We were recently reminded of the importance of transfer lot size when one of the plants in our study set up a manufacturing cell that used flexible equipment to reduce setup times in an attempt to shorten lead times for cranks. Unfortunately, the lot size used was about 10,000 and was not split for the purpose of moving work. As a result, although the time to complete the first crankshaft was markedly reduced, the average flow time for the entire lot changed very little.

**Sharing Transfer Batches (or The Value of “Non-Value-Added”).** Many times an inherently slow process will require many parallel tools to provide sufficient capacity. If incoming transfer batches are assigned to individual tools, lead time can be excessive. If the batches are shared by more than one tool, the average flow time can be reduced. Multiple tools will complete the work on the shared transfer batch faster than a single tool and the entire batch can then move more quickly to the next work center.

Of course, sharing transfer batches may require additional material handling as the shared batch may not be near all of the work stations. It may seem somewhat counterintuitive that adding operations that are clearly “non-value-added” will result in reduced lead times, but it is clearly the case. Adding such operations, however, can reduce the effective capacity of the work center and may require the hiring of an additional worker to serve in a material handling function. Whether to do this represents an obvious tradeoff between increased labor cost and shorter lead times.

In some cases, however, the workers are not highly utilized. In one of the plants studied that produced raw (i.e., with no components) circuit boards, decreasing flow time had become an important priority. The company had just undertaken a “non-value-added reduction drive” and flow times remained excessive. The sequence of operations to circuitize a board involved two conveyor type operations separated by a manual exposure operation. Because this operation was necessarily long, there were 12 expose machines operating in parallel. Boards were collected at the end of the first conveyor in carts and then moved to the expose operation, with one cart per expose machine. The flow time through the entire process center was about 2 days. Although reluctant at first, the plant management agreed to utilize a single outbound cart in the expose area, reducing the average wait-to-move time to about 1/12th of its original value. The result was a 30% reduction in flow time for the process center with no increase in head count.

By sharing carts, the manager has increased the number of non-value-adding steps, that is, more move operations. This appears to be in conflict with some conventional JIT slogans. However, perhaps it is more precise to say that conventional JIT slogans are not always consistent with the JIT wisdom of Ohno [10, p.11] who advises that we remove “the non-value-added wastes.” The key word is waste. If we examine the cart sharing example more closely we see that by adding the waste of “material handling” we have eliminated the waste of “wait-to-move time.” A different solution might be to place the machines into closer proximity or to provide some type of material handling device that would eliminate the additional handling. In our case, however, such costly modifi-
cations were not warranted since the operation was not the bottle-
neck of the line and the operators could easily keep up with the re-
quired production rates. To summarize, we recognize that making
needless costly modifications is also a form of “waste.”

Queue Control. An obvious means to reduce queue time is to
maintain shorter queues. In a Kanban [7,9,11] or CONWIP [20,21]
system this is straightforward since, in these pull systems, WIP is
controlled directly. If such a system is used, WIP can be reduced
until further reductions would lead to disruptions. If the cause of
the potential disruption can be determined and eliminated, then
queues can be further reduced. For instance, if WIP reduction
results in starvation at a bottleneck and there are unreliable or
extremely variable machines in the line, the system can probably
be improved by addressing these problems. If this variability cannot
be reduced, extra WIP must be maintained.

In an MRP system, controlling queues is more difficult. How-
ever, techniques such as input/output control [18] do provide added
control of queue lengths. The WIP levels in each work center
should be evaluated to see where reductions can be made in order
to reduce safety lead times. Also, production releases coordinated
to bottleneck rates will avoid producing parts that are not imminently
needed.

There is a limit to how far WIP levels can be reduced before
throughput will invariably be decreased. “Zero inventories” will
result in zero throughput, regardless of the popularity of the slogan.
A recent paper by Conway et al. [22] indicates that most U.S.

firms operate with far more WIP than is needed. They suggest that
the WIP that is present be placed in front of the bottleneck for
unbalanced lines and near the middle for balanced lines.

Synchronize Production

Because a part assembly cannot be completed until all components
are available, synchronization between fabrication and assembly is
extremely important. The synchronization of component fabrica-
tion with assembly schedules is also difficult. It is quite common
for an assembly line to produce based on what is available rather
than what is needed. Harmon and Peterson liken productivity
improvement opportunities in assembly systems to the proverbial
pot of gold at the end of the rainbow [8]. In this same vein, efforts to
synchronize component fabrication with assembly can have
dramatic effects on waiting inventories and therefore lead times.

In Kanban and CONWIP systems, synchronization occurs
naturally, since work is pulled only as needed. MRP II systems
would, in theory, also synchronize production if flow times were
constant. However, because flow times are not constant, the lead
times used to compute job release times from due dates are inflated
to values much larger than average flow times to accommodate
contingencies. Such practice builds inefficiency into MRP
systems because inflated lead times cause jobs to finish early (on
average), thereby increasing waiting inventory.

The problem of deciding which jobs should be performed next
takes place in the shop floor control module in the MRP II hierar-
chy and is associated with “dispatching.” A commonly used tech-
nique involves a “dispatching rule” which allows an operator or
foreman to choose the next job by considering those currently wait-
ing in queue. Typical dispatching rules are “shortest processing
time,” “earliest due date,” and “critical ratio,” a rule that involves
both job size and due date (see [23] for a survey of dispatching rules).

It is currently popular in the academic literature to propose
new, presumably better, dispatching rules. Unfortunately, many of
these rules are better in the sense that they improve some myopic
measure of performance which considers only fabrication lines.
Use of rules which do not take into account their effect on waiting
inventories may cause tremendous increases in flow time for finished
assemblies.

The SPT rule (process the job with the shortest processing
time first) is a good example. This rule will provide the best per-
formance for a single machine in the sense that average flow time
for all parts will be as low as possible. Simulation studies have
also shown SPT to be effective in some shop shops [24]. Unfortu-
nately, performing the smallest job first will have the obvious effect
of increasing the variance of flow time; large jobs simply wait
longer. Since lead times depend not only on average flow time but
also on flow time variance, implementation of SPT could have the
effect of increasing lead times.

Increases in flow time variance are particularly disastrous in
an assembly system. If an assembly requires two components, one
which requires a long processing time and one with a short
processing time, then SPT will result in excessive waiting inven-
tory. An example of a dispatching rule which does take into
count the effect on waiting inventory is EDD (process the job
with the earliest due date first). The corresponding rule for pull
systems is “first in system, first served” (process the job request
that arrived first).

A second problem is to match production rates of fabrication
with consumption rates at assembly. Kanban, OPT [19], and
CONWIP systems will naturally equalize these production rates
even in unbalanced lines. Many of the problems with excessive
lead times which we have encountered are caused by the perceived
need to "keep busy." In one of the plants studied, we observed
workers who run their machines even when parts are being
produced much faster than they are being used at downstream
machines; they stop only when there is no physical space left for
additional WIP. Clearly, it is important for management to under-
stand the need to balance flow and to communicate this to workers
in the form of appropriate performance measures.

One important point to note is that synchronization becomes
much easier as the variability in the production process is reduced.
If production times are very predictable, it will be unnecessary to
build in large safety lead times and subassemblies will be much
more likely to arrive at assembly simultaneously. Hence, measures
that reduce variability of flow time are important in improving
synchronization.

Smooth the Work Flow

Almost all of the works on JIT implementation describe the need to
smooth work flow in order to reduce inventories and lead times
[7,9,10,11]. We list some easily implemented methods.

Leveling Work Releases. The goal of leveling work releases is
to maintain an even workload. Low workloads lead to small
queues and quick turnaround, while high workloads cause long
queues and long flow times. Overall, uneven workloads increase
both flow time mean and variance. When discussing workloads, it is
important to note that the time scale we are considering is on the
order of days and not months as in discussions of using a chase or
level aggregate production plan (e.g., [25, p. 214]). Gradual
increases in release rates, accompanied by increases in capacity,
can facilitate a chase production strategy. What is important is to
release frequently and in small amounts. That way, there are
smaller queues at release points and less likelihood for a “bulge” of
WIP that alternately overloads and starves production resources.

Ultimately, releases should be tied to production at bottleneck
work centers. In this way, a “pipeline” to the bottleneck is
established and is always full. More frequent releases only create WIP in front of the bottleneck with no increase in production. Also, planning changes in the form of order cancellations and engineering change orders are less disruptive is the work is not yet released than if the work is already in process and somewhere on the floor. In pull systems such as Kanban and CONWIP, the release rates (pull rates) are naturally equal to downstream production rates. In push systems, such as MRP this is more difficult, although more frequent releases (such as twice daily rather than weekly) combined with some form of WIP control (e.g., input/output control) can do a great deal to reduce flow time variance and congestion on the shop floor. The drum-buffer-rope system proposed by Goldratt [13] is an example of a push system that directly ties work releases to bottleneck production.

Establishing a Uniform Work Flow. Changes in the product mix lead to setups, changeovers, and disruption of factory rhythm. By establishing routings for families of parts [26] so that setups between families are avoided, and by reducing setup times for the instances where setups cannot be avoided [9, pp. 76-84] many of these problems can be alleviated. Also, since some changes are inevitable, cross-training of workers will minimize the disruption of these changes.

Product standardization, involving the use of common lower levels, can reduce the number of different parts that must be produced, tracked, and inventoried. Also, by customizing the product late in production the system is less sensitive to changes in customer orders. Both of these changes were particularly relevant for one of the firms that we studied. This company conducted a product standardization review that focused on limiting the range of customization options and developed ways to delay customization until the end of the manufacturing process. This strategy enabled them to offer maximum flexibility to their customers at little cost to themselves.

Rationalizing Line Balancing. We have found that systems with all centers operating independently and at similar rates will tend to have longer average flow times than systems with a distinct bottleneck [27]. Any time one work center works somewhat faster than the preceding work center (due to random variations), it will become starved unless WIP is excessive. Because of this dependency, the periods of high production do not make up for low periods. Although the usual reason given for balancing a production line is to "break bottlenecks" the practice has the opposite effect — all stations become bottlenecks and therefore must each be protected with additional WIP.

This is not to suggest that line balancing should not be employed on paced assembly lines. On such lines, the rhythm of the system prevents starvation and WIP buildup. Line balancing was never intended to balance capacity between work centers.

Goldratt [28] states this succinctly, "Balance the flow, not the capacity." During the inevitable slack periods at nonbottleneck stations, preventive maintenance and other housekeeping tasks can be performed.

Eliminate Variability

Unlike Suzuki, Schonberger [1] insists that the root of all evil is variability. Variability in processing times caused by rework, downtime, and lack of consistency in production methods increase both mean and the variance of flow time. Unlike inventory, there are no instances where more variability is good.

Some strategies for reducing variability include the following.

Reducing Rework. Rework operations can have a tremendous impact on flow time mean and variance, particularly if jobs requiring rework compete with regular jobs for resources. Tight quality control and SPC should be used to eliminate rework whenever possible. Where large lot sizes are used, quality checks should be made before completion of the lot. This will help avoid the situation where the entire lot is completed before it can be checked. Where possible, quality checks should be located in front of the bottleneck or other time-consuming operations. Quality control in front of the bottleneck will avoid wasting its important capacity on defective parts. Quality checks in front of lengthy operations will help decrease the amount of rework. If rework is disruptive and means for eliminating it cannot be found, a dedicated line for rework may be required. This will help smooth the flow of work through the facility and speed the recombination of jobs requiring rework with the rest of the jobs. However, rework lines can also serve as "psychological crutches" and remove some of the stigma attached to quality problems. Elimination of the problems at the source should always be considered before resorting to rework lines.

Improving Machine Reliability. An important source of flow time variance can be machine down time. Although many companies track machine availability (A), some do not track the mean time between failures (MTBF) and the mean time to repair (MTTR). The relation between these quantities is

\[ A = \frac{MTBF}{MTBF + MTTR} \]

Although only availability is needed to determine the capacity of a machine, it is not sufficient to determine the overall congestion in the line. Long and infrequent outages are more disruptive than short and frequent ones even if the availabilities are equal. This is because short outages require relatively small buffer stocks between machines to prevent starvation whereas longer outages require correspondingly more. Of course, downtime at bottleneck resources is particularly important, not only because of its effect on system capacity, but also because it has a strong effect on flow time variance. Regularly scheduled preventive maintenance can be used to minimize machine downtimes. Replacing or overhauling unreliable machines, particularly if they are at the bottleneck, can also be used to reduce flow time variance.

We developed a queueing model of the production facility of one of our clients in order to estimate the impact of improving MTBF and MTTR. With this model we have been able to justify variability reduction projects that would have been impossible to justify using ordinary accounting models.

Planning for Yield Losses. Although the best strategy to deal with yield loss is to eliminate it, this is not always possible. A reasonable alternative is to consider carefully an appropriate strategy for dealing with yield loss.

A common procedure for deciding on the size of a job to start, given a desired finish quantity and some knowledge of yield loss, is to inflate the job size by dividing the desired quantity by the average yield (i.e., the proportion of good parts that are completed). If either the yields are low or the job sizes are large, it is unlikely that the resulting finish quantity will be equal to the desired quantity. However, by applying the above simple strategy, we imply that to end up with one piece short is just as bad as having one piece too many. This is seldom the case, since short pieces must be expedited through the factory to catch up with the rest of the job whereas additional pieces typically sit in finished goods inventory until the next order. Both outcomes are undesirable but the first usually has more severe consequences. The flow time clearly increases for the order involved and for other jobs as well if setups must be broken to
facilitate the expediting. Rework also increases the variance of flow time because of the congestion it causes.

In one of the plants we studied, we proposed a change to the way jobs were "yielded." Each job was started with a quantity large enough such that the probability of having sufficient parts was equal to a prespecified value, say 90%. This procedure resulted in a slight increase in finished goods while greatly reducing the number of small, "hot" jobs that had been both robbing capacity (through more setups) and increasing lead time.

**Vendor Variability.** Variability is an important factor in purchased, as well as fabricated, parts. One of the firms in our study assembled final products almost entirely from purchased parts. Prior to our involvement, they had based their decisions on how much buffer lead time to build into the schedule entirely on cost (i.e., inexpensive parts were brought in well ahead of the times they were needed, whereas expensive parts were ordered with very little slack in the schedule). After our review, we concluded that variance of vendor lead time was as important as the cost of the part (i.e., highly variable vendors require more buffer lead time than very dependable ones). On the basis of this realization, we set up a system for establishing appropriate buffer lead times and tracking vendor performance that directly incorporated variance.

**CONCLUSION**

We have argued that the key methods for reducing lead time are those that reduce mean flow time and flow time variance. As indicated in our discussion, there are many ways to accomplish these reductions with little or no cost.

We have also seen the danger of implementing the JIT philosophy through the use of slogans. We have seen instances of "good" inventory and valuable "non-value-added" operations. However, we do believe in one slogan. **Don't look back, you never know who's gaining on you.**

**REFERENCES**


**APPENDIX**

To understand completely the relationships between lead times, flow times, WIP, and finished goods inventories, we consider the production of a part between two adjacent levels within a product structure. This production 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). Suppose the part is needed at 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 unrealistic, since at assembly operations 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. We represent the average of this waiting inventory as *T*. Finally we let *W*, *T*, and *θ* designate, respectively, the average work in process (not including waiting inventory), flow time, and throughput for the line. Then from Little's law we see that

\[
W = T \theta
\]

We further assume that the distribution of the flow time *T* can be approximated by a distribution whose mean serves as a location parameter and is equal to *T* and whose standard deviation functions as a scale parameter and is equal to *σ*.* Examples of such distributions include the normal, uniform, and symmetrical triangular. We denote this distribution function *F* and define \(Z = (T - T)/\sigma\) having a distribution function *Φ* such that *Φ(σZ) = s*. Then the necessary lead time for a given service level *s* will be,

\[
l = \sigma Z + \bar{T}
\]

Hence, lead time increases linearly in both average flow time and the standard deviation of flow time. Let *θ* be the density corre-
sponding to \( \Phi \). In cases where the normal is a good approximation, the average waiting inventory \( \bar{I}_w \) will be

\[
\bar{I}_w = \bar{\Theta} E[\max(0, I - T)]
\]
\[
= \bar{\Theta} \int_0^T F(t) \, dt
\]
\[
\equiv \sigma_r \bar{\Theta} \int_{-\infty}^{\infty} \Phi(z) \, dz
\]
\[
= \sigma_r \bar{\Theta}[\Phi(\sigma_r) + \Phi(-\sigma_r)]
\]

The quantity in the brackets in the last equation is constant for any given service level. Although this model is too simple to represent a real production facility, it does point out that, for fixed values of throughput and service, the expected waiting inventory increases linearly with the standard deviation of the flow time but does not depend on the mean flow time.

Hence, as we stated above, reduction of average flow times allow for smaller reorder points, more competitive lead times, a shorter frozen zone and thereby more flexibility. The smallest possible value for average flow time is the raw processing time and this occurs only when the variance of the processing times are zero [24]. Small variance in flow time will reduce waiting inventories as well as lead times.

Wallace J. Hopp is an associate professor in the Department of Industrial Engineering and Management Sciences at Northwestern University and performs research in probability models of manufacturing systems, forecast horizons in nonhomogeneous Markov decision processes, and the development of alternative production control systems. He teaches courses in stochastic processes and a senior design sequence. He has research articles in Operations Research, Management Science, IIE Transactions, Journal of Optimization Theory and Applications, Mathematical Programming, Naval Research Logistics Quarterly, Queueing Systems, and the International Journal of Production Research. He is a senior member of IIE, a full member of ORSA, and a member of TIMS and SIAM. Professor Hopp is also an active consultant to manufacturing firms.

Mark L. Spearman is an assistant professor in the Department of Industrial Engineering and Management Sciences at Northwestern University where he teaches in the areas of production and inventory control and systems simulation. Professor Spearman’s research addresses the need for better manufacturing models and ways to incorporate JIT techniques within hierarchical production planning systems. This research has been supported by grants from both private industry and the National Science Foundation. He is an area editor of Production and Operations Management and has research papers appearing in and/or has served as a referee for Engineering Economist, IIE Transactions, International Journal of Flexible Manufacturing Systems, International Journal of Production Research, Journal of Manufacturing and Operations Management, Management Science and Operations Research. He is a senior member of IIE, a full member of ORSA, a member of TIMS, POMS and is certified in production and inventory by APICS.

David L. Woodruff is currently a research associate at Northwestern University but soon will be an assistant professor at the University of California at Davis. His past work has dealt with sequencing jobs in a generalized pull system. Mr. Woodruff has research articles appearing in International Journal of Production Research and the Journal of Manufacturing and Operations Management. He has eight years experience in software development and general management. He is a member of ORSA, TIMS, IIE and POMS.