MANUFACTURING EXECUTION: THEORY AND CONCEPTS

Stanley A. Hendryx

Stanley A. Hendryx *is a supervisor in the Manufacturing Systems Engineering Department at AT&T Bell Laboratories in Allentown, Pennsylvania. He has been a consultant to AT&T manufacturing locations in Denver, Richmond, Omaha, North Carolina, Reading, and Singapore on just-in-time and other performance improvements. He joined AT&T in 1982 with an S.B. and S.M. in electrical engineering from the Massachusetts Institute of Technology, Cambridge.*

Modern theory of manufacturing execution holds that time is the most telling indicator of manufacturing performance, that by systematically reducing the manufacturing interval and its variation, a manufacturer can follow a proven path to world class status. Because the method is based on a different philosophy of manufacturing operations, rather than on expensive technology, it requires little initial capital investment. The role of variation is emphasized in this paper, and a quantitative analysis of variation shows how variability occurs in the factory. Batching, quality problems—including low yield—and unreliable equipment and material flows are shown to be major sources of variation and delay. The roles of factory loading, load smoothing, and flexible manufacturing are also shown in quantitative terms. Practical suggestions are given for implementing JIT (just-in-time), including kanban and pull systems, lot sizing and setup time reduction, and compensating for losses from low yields to make JIT work.

Introduction

JIT manufacturing can be defined as a process to achieve continuous improvement by systematically eliminating the three evils of manufacturing: *excess*, *waste*, and *variation*. This article explores the concept of JIT manufacturing to identify principles that might be applied to any manufacturing, not only to high-volume, assembled-to-stock commodity products.

This article is really about *time*: its uses and fundamental significance in manufacturing performance and competitiveness. Particular attention is paid to the insidious effects of variation, and some principal causes of variation are suggested. The main types of variation studied are varying times between arrivals of work at factory workcenters, and varying service times. These variations play a primary role

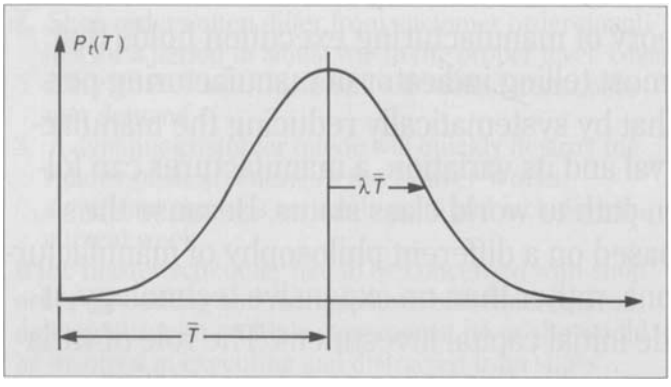

Figure 1. Probability density function.

in factory performance, and are influenced by virtually every action of factory personnel.

Time—The Gauge of Manufacturing Performance

Successful manufacturers combine high quality, low cost, and large market share with the shortest and most nearly constant manufacturing time. To see why, we need only reflect on how manufacturing time is used. We will find many clues to improving operations by examining the causes of waiting time. The total time a product spends in the factory includes both processing time and waiting time, and is called variously *factory time, manufacturing time, manufacturing interval,* or just *interval.*

Before JIT, direct manufacturing process operations might have accounted for only 2 percent of a product's factory time; the remaining 98 percent was *waiting time.* But even some production time might be caused by a faulty operation that spoiled the product, requiring either a total restart or diversion into a rework or repair area. We can account for most of the waiting time by observing that, in most factories, the product was batched into queues, and each unit had to wait its turn. Or, it was processed ahead of schedule to help a department's efficiency rating, but then sat in a storeroom until needed. Perhaps the unit was needed urgently, but was delayed because the next machine had to be fixed. Or maybe a companion unit was late, and "the show could not go on" until it arrived. Finally, the unit may have had to wait for testing or inspection to ensure it conformed to quality standards.

In this pre-JIT environment, it is evident that, except for brief periods when processing adds value to the product, all the rest of the time is the result of excess, waste, or variation. The best manufacturers do not accept these operational penalties as unavoidable. They seek continuous improvement by systematically eliminating the causes of problems in a quest for nothing less than domination over the three evils of manufacturing.

Reducing manufacturing interval and its variation are an acid test to measure improvement in manufacturing performance, and gauge a business' ability to improve its entire operations cycle from order entry through product delivery. As the causes of excess, waste, and variation are eliminated, there is less waiting, and operations become less variable. Consequently, the manufacturing process becomes more reliable, costs come down, and service improves.

Speed is an important source of competitive advantage. Faster, more constant competitors need less inventory for a given level of customer service, have a lower risk of obsolete inventory, and generally enjoy lower costs and higher profits. They have greater flexibility and more options for innovative approaches to customer service. Rapid product development and rapid manufacturing execution provide an enormous advantage in winning market share by enabling the company to bring product innovations to market sooner than competitors.[1]

Managing a manufacturing business by managing *time* is the essence of JIT.

The Just-In-Time Paradigm

JIT, as the term implies, tries to streamline the manufacturing process so, at each stage, production

and consumption (by the next stage) are matched over short time periods—typically a few minutes to a day. Implicit in JIT, and essential to it, is the continuing proper utilization of the business' resources. To achieve both synchronous operation and high utilization is a formidable goal, attained by systematically finding and eliminating excess, waste, and variation. To pursue this goal, priorities are set by the problems that most disrupt synchronous operation, or that have the most adverse economic effects, e.g., low yield. This is the JIT paradigm.

Variation and Manufacturing Interval

Understanding variability is the key to understanding what needs to be improved. Thus, we will look more closely at *variation*, particularly in interarrival times at workcenters, and in service times: how to measure variation, its causes, and its influence on interval.

Measures of Variation. Before proceeding, we need precise measures of variation. We will use *standard deviation* and *coefficient of variation*. Standard deviation, commonly denoted σ (sigma), is a measure of how a probability density function—showing the relative frequency of occurrence of different values—is distributed around the mean. Wider density functions have higher standard deviations (see Figure 1). Standard deviation has the same dimensional units as the quantity under discussion. The square of the standard deviation, σ^2, is called the *variance*.

The ratio of the standard deviation to the mean

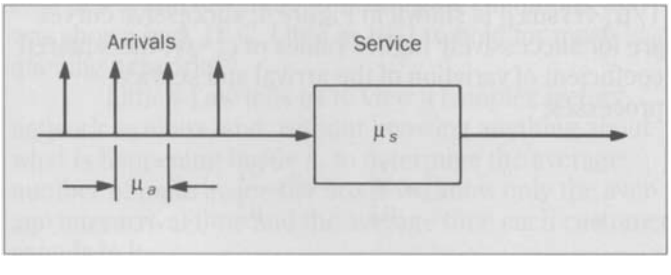

Figure 2. Single server service facility.

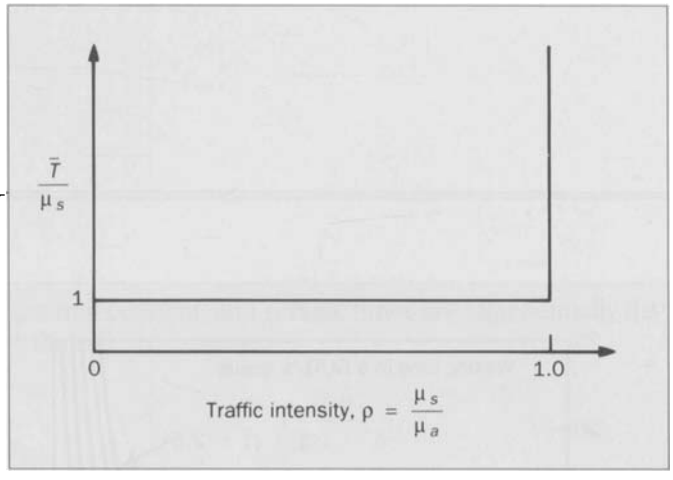

Figure 3. Time in the facility, no variation in service or arrival times.

is called the *coefficient of variation*, and is shown as c: $c=\sigma/\mu$, where μ (mu) denotes the mean. Coefficient of variation is defined only for processes having a non-zero mean, and is useful when *percentage* deviation from the mean is relevant; c is dimensionless. The squared coefficient of variation, c^2, appears frequently.

Readers interested in the mathematical details are referred to the standard literature in queuing theory[2] and stochastic processes.[3,4]

Variation, Delay, and Utilization. Consider a service facility with a single server, where service time, μ_s, is the same for all customers. The *capacity* of the facility is then $1/\mu_s$ customers per unit time. If the time between customer arrivals is invariably μ_a, and as long as the service time is less than the time between arrivals, then no one has to wait for service, because the previous customer is always done before the next one arrives. This is true even if the interarrival time is arbitrarily close to the service time—as long as μ_a is greater than or equal to μ_s—because we assumed that neither μ_s nor μ_a ever varies.

However, if μ_a is even *slightly* less than μ_s, the server will get farther behind—by time $\mu_s - \mu_a$—as each customer arrives. In the long run, the queue in front of the server will become infinite, and the average customer will never get served, even though the server is busy all the time. This situation is illustrated in Figures 2 and 3.

The ratio of average service time, μ_s, to average interarrival time, μ_a, is called *traffic intensity*, a term from telephone switching, which figured strongly in the early

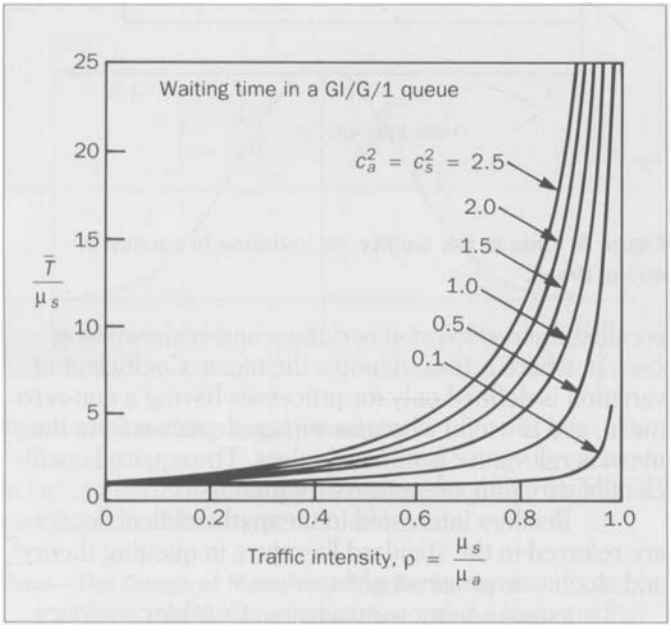

Figure 4. Time in the facility, general independent arrivals.

development of queuing theory. Traffic intensity ranges between zero and one, and is denoted by ρ (rho): $\rho = \mu_s/\mu_a$. It measures how busy the server is; when $\rho = 1$, the server is always busy. Figure 3 depicts average total time in the facility—waiting plus service time—versus traffic intensity.

Now, we relax the assumption that interarrival times are constant, and allow them to vary slightly. If μ_s were 3 minutes, and interarrival times varied between 59 and 60 minutes, we would not expect any waiting, despite the variation, because one customer would always complete well before the arrival of the next. Traffic intensity would be $\rho \approx 3/60 = 0.05$. Now, let the interarrival time decrease as traffic picks up, so the interarrival times vary between 2.5 minutes and 3.5 minutes, with an average interarrival time, μ_a, of 3.3 minutes. Thus, $\rho = 3/3.3 = 0.91 < 1$; but obviously, some customers arrive before the previous one is done, and have to wait. Depending on when the next customer arrives, there may be a wait for both the customer being served and the one already waiting. Thus, a queue begins to form.

As long as traffic intensity is less than one, the queue is stable and the average waiting time is finite, though both queue length and waiting time vary. Clearly, queuing is more severe the closer the traffic intensity is to one, and the greater the variation in interarrival times. Analyzing a single-server facility with general independent arrivals and general service time gives the average interval, $\bar{T}$:[5]

$$\bar{T} \approx \mu_s + k \frac{\mu_s \rho}{1 - \rho} \frac{c_a^2 + c_s^2}{2},$$

where k is a factor in the range [0,1] involving higher order terms in ρ, c_a^2, and c_s^2; $k << 1$ only for relatively low traffic intensity or variation, where:

$$k = \exp\left[\frac{-2(1 - \rho)(1 - c_a^2)}{3\rho(c_a^2 + c_s^2)}\right], c_a^2 \leq 1;$$

and:

$$k = \exp\left[\frac{-(1 - \rho)(c_a^2 - 1)}{c_a^2 + 4c_s^2}\right], c_a^2 \geq 1.$$

$\bar{T}/\mu_s$ versus ρ is shown in Figure 4; successive curves are for successively higher values of $c_a^2 = c_s^2$, the squared coefficient of variation of the arrival and service processes:

$$c_a^2 = \frac{\sigma_a^2}{\mu_a^2}, \quad c_s^2 = \frac{\sigma_s^2}{\mu_s^2},$$

where σ_a^2 and σ_s^2 are the variances of interarrival and service times, respectively. The second term in the equation

for $\bar{T}$ is the waiting time. Notice waiting time is nearly proportional to the average of c_a^2 and c_s^2; we will use this fact extensively throughout this article as we study the relationship between delay and different causes of variation.

Variation in service time has the same effect as variation in interarrival times. To minimize congestion, it is as important to control variation in the service process as in the arrival process.

The abrupt increase in $\bar{T}$ for traffic intensities above 0.8 to 0.9 suggests that, in the presence of variation, 10 to 20 percent excess capacity is needed to prevent congestion. This supports the popular idea that excess capacity is required for JIT.

We have spoken of interval and congestion as though they were interchangeable. We will now formalize this relationship.

Interval and Inventory: Little's Law. We have discussed interval in a queue—service time plus waiting time—but have said little about *how many* customers are waiting. Everyday experience suggests that the longer the queue, the longer the wait, and that if service time is less, the queue moves more quickly. This intuition is formulated as *Little's Law*:

$$\bar{T} = \mu_a \bar{i}$$

where $\bar{i}$ is the average inventory or length of the queue. This relationship involves *averages* of interval, interarrival times, and inventory in *steady-state* conditions, and was shown by J. D. C. Little in 1961 to hold for most queuing networks.[6]

Little's Law tells us to view a complex factory network as a box, and, without knowing anything about what is happening inside it, to determine the average number of units inside the box if we know only the average interarrival time and the average time each customer spends in it.

For the special case of Figure 4, where $c_a^2 = c_s^2 = 1$ (which pertain when arrivals in non-overlapping intervals are independent, and service times are exponentially distributed),

$$\bar{T} = \frac{\mu_s}{1 - \rho}.$$

Equating this with $\bar{T}$ in Little's Law, for this special case we obtain the dependence of average inventory on traffic intensity:

$$\bar{i} = \frac{\rho}{1 - \rho}.$$

Because Little's Law deals with steady-state conditions, the average inventory does not change, so the average time between departures must be the same as the average interarrival time.

A common error in using Little's Law is to assume the time between departures is the *service time*. This is not generally correct; the servers may be idle at times because of low traffic intensity or variation in the arrival process. The same interval and inventory could result if the service time is longer and nearly constant with higher traffic intensity, or shorter and highly erratic with lower traffic intensity. Little's Law will not tell us what happens inside the box. It tells us the average number of customers inside, the average time they spend there, and the average time between arrivals. It also tells us is that, if the average time between units produced, μ_a, is constant—that is, if steady-state conditions prevail—the average interval is proportional to the average inventory. In steady-state conditions, if we know any two of the three variables $\bar{T}$, μ_a, and $\bar{i}$, we can calculate the third.

Common Sources and Effects of Variation. We have said that variation leads to queuing, delays, and excess inventory. We now look at common situations in a factory that lead to variation, and estimate their relative importance by comparing the values of c^2 they produce, since waiting time tends to be almost proportional to c^2. This helps in concentrating our efforts to reduce inter-

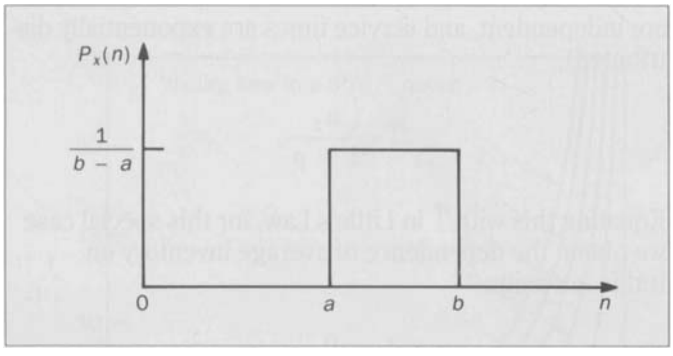

Figure 5. Uniform probability density function.

vals. We will look at five situations:

- Random independent arrivals
- Random arrivals with uniformly distributed factory load
- Intermittent operation
- Yield losses
- Batching.

Random Independent Arrivals—Poisson Process. Some arrival processes can be modeled by assuming that arrivals occurring in non-overlapping intervals are statistically independent, and that the probability of an arrival in a small interval is proportional to the length of the interval. Customers arriving independently (as opposed to batched or according to a predetermined regular pattern) at an on-demand service facility (e.g., telephone switch, toll booth, on-demand warehouse, restaurant) might be modeled this way. This is a *Poisson process*; the times between successive arrivals (first-order interarrival times) are exponentially distributed:

$$p(t) = \begin{cases} \frac{1}{\lambda} e^{-\frac{t}{\lambda}}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

where $p(t)$ is the probability density for the interarrival time, t, and λ (lambda) is a parameter. The mean and variance of interarrival time, and its squared coefficient of variation for the Poisson process, are:

$$\mu_a = \lambda \quad \text{and} \quad \sigma_a^2 = \lambda^2,$$

$$c_a^2 = \frac{\sigma_a^2}{\mu_a^2} = 1.$$

The Poisson process will be the benchmark against which to compare other situations.

Random Arrivals—Uniform Distribution. The *uniform distribution* model reflects some smoothing of demand by production control departments to help level factory load. Through the Sales, Inventory, and Production Planning (SIPP) process, the business strives to harmonize order projections, inventory levels, and production plans, and to produce a load for the factory that is as constant as possible, subject to customer orders. The SIPP process might result in a load, described in units, being established each week, but varying from week to week between a minimum, a, and maximum, b, so it can be modeled by a uniform probability density function (see Figure 5). The statistics of the uniform distribution are:

$$\mu = \frac{b+a}{2}, \quad \text{and} \quad \sigma^2 = \frac{1}{12}(b-a)^2,$$

$$c^2 = \frac{1}{3}\left(\frac{b-a}{b+a}\right)^2 = \frac{1}{3}\eta^2,$$

where η (eta) is the maximum fractional deviation from the mean. Because $0 \leq a < b$ (negative demands are disallowed), the worst case is when $a=0$; then $c^2 = 1/3$. For example, if weekly demand is uniformly distributed between 400 and 600 units, the squared coefficient of variation of the weekly demand would be:

$$c_a^2 = \frac{1}{3}\left(\frac{600-400}{600+400}\right)^2 = \frac{1}{3}\left(\frac{1}{5}\right)^2 = \frac{1}{75} = 0.013$$

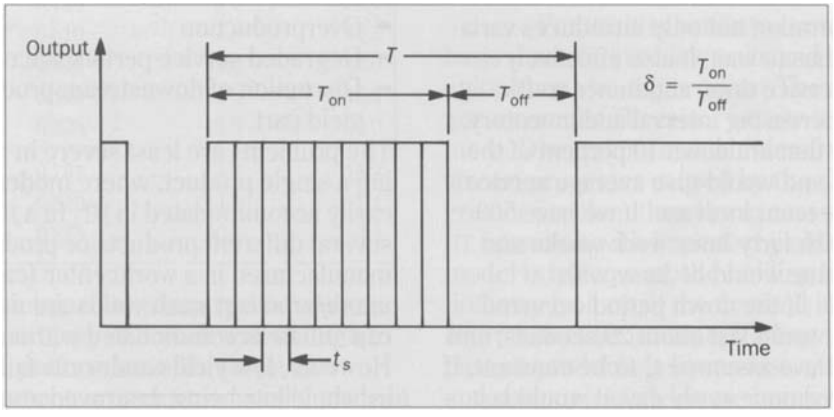

Figure 6. Intermittent operation.

When production control imposes a weekly quantity on the factory, the implication is that, at the starting point of manufacture, the interarrival times for the units in the week's load are the same; any variation in starting intervals is introduced by the shop. If the shop works S hours per week to produce a weekly demand of d units, the interarrival times for the week are $t=S/d$, i.e., t will vary from week to week in inverse proportion to the demand. For small η, then, c^2 for interarrival times can be approximated by c^2 for the demand: $c_a^2 \approx c_d^2$, $\eta << 1$. Where $a=400$ and $b=600$ ($\eta=\pm 20\%$ variation around the mean), then $c_a^2 \approx c_d^2 = 1/75$. Production control will have reduced the variation by a factor of 75, if the underlying demand process was Poisson. Production control has effectively smoothed, or averaged, the underlying demand to get this reduction.

Though plus or minus 20 percent swings in load could reduce congestion compared to Poisson, the required swings in capacity would probably be hard to manage, and capacity utilization would be poor. Experience suggests that factories can tolerate up to about 10 percent overall fluctuation in the week-to-week load (plus or minus 5 percent). More than this affects the factory staffing level, overtime required, and labor efficiency, because the number of employees needed in a given week is proportional to weekly load. A plus or minus 5 percent load fluctuation implies a factory tolerance level of $c_a^2 = 1/1200$, or 1,200 times less than Poisson. If the factory did not add significant variation, operating consistently at this level of variation would result in unusually low congestion—even at high traffic intensities—and factory performance would be considered world-class by any standard.

Production control and the SIPP process often can come close to providing this high level of stability in factory load. However, most of the expected advantage in obtaining short intervals and low inventory often is lost because of variation in the factory itself.

Intermittent Operation. Machines break down, or are taken out of service for setup, maintenance or other purposes. The case where this happens periodically is shown in Figure 6, where δ (delta) is the *duty cycle*, or fraction of running time, $\delta = T_{on}/(T_{on} + T_{off})$, t_s is the machine time per unit, and N is the number of units produced between shutdowns, $N = T_{on}/t_s$. The statistics for this service process are:

$$\mu_s = \frac{t_s}{\delta} \quad \text{and} \quad \sigma_s^2 = t_s(T_{on} - t_s)\left(\frac{1}{\delta} - 1\right)^2$$

$$c_s^2 = (N - 1)(1 - \delta)^2.$$

Intermittent operation not only introduces variation, but because work has to wait, it also effectively increases the average service time, and hence traffic intensity, thus doubly increasing interval and inventory. For example, machines that are down 10 percent of the time would have $\delta = 0.9$, and would give average service time $\mu_s = 1.11 t_s$, an 11 percent increase. If we have 500 units per week to run with forty hour work weeks and $t_s = 4$ minutes, the machine would be busy with $\rho = 33.33/36 = .93$ if $\delta = .9$. If the down period occurred after every unit, $N = 1$, it would last about 29 seconds, and then $c_s^2 = 0$, because we have assumed t_s to be constant. If the down period occurred once every day, it would last 48 minutes, and then $c_s^2 = .99$, about the same as Poisson. If the down period occurred once every two weeks, it would last eight hours, and then $c_s^2 = 9.9$. Evidently, shorter, more frequent outages are preferable; it is better to maintain a machine a few minutes per day, than to save up maintenance items and shut it down for a day every two weeks.

It is common in many factories to run machines intermittently with high excess capacity. For example, a machine capable of 100 parts per hour would be run only 5 hours per week if the weekly demand were 500. If this were all done at once in a 40 hour week, we would have $\delta = 5/40 = 1/8 = 0.125$, and $c_s^2 = 382$—*much* worse than Poisson.

Part shortages and yield losses create intermittent operation, with results similar to those previously described. In a multistage manufacturing line, these effects accumulate going down the line, exacerbating the overall result.

Yield Losses. There is nothing good about low yields. Not only are there obvious economic losses from scrap and rework, but many problems are introduced into the queuing processes when yields are low. These include:

- Increased variation in product flow
- Increased traffic intensity resulting from the loss of capacity
- Overproduction
- Degraded service performance
- Disruption of downstream processes requiring the low yield part.

The problems are least severe in repetitively manufacturing a single product, where moderately good yields are easily accommodated in JIT. In a JIT environment where several different products or product variants are manufactured in a workcenter (called *mixed-model manufacturing*) such yields are more problematic, but can still be accommodated without too much difficulty. However, low yields and some failure mechanisms, such as whole lots being destroyed at once, can disable JIT in mixed-model environments.

The most obvious effect of low yield is output stream gaps, creating downstream interarrival times similar to intermittent operation with δ=yield and $N = 1/(1 - \text{yield})$. To offset the losses, it is common to anticipate them and start additional units to ensure having enough good product to meet demand; but this increases traffic intensity. We will examine how a mixed-model JIT process might compensate for imperfect yields.

A process that causes random failures to occur *independently* to any unit with probability (1 – yield) is called a *Bernoulli process*. Let us assume a mixed-model process with each model having starting lot size n. Let the probability a unit is good be p, and the expected number of good units be E. The statistics of this Bernoulli process are:

$$E = np \quad \text{and} \quad \sigma^2 = np(1 - p)$$

$$c^2 = \frac{1 - p}{np}.$$

Note that yield = p, and increasing starts, n, by a factor of $1/p$ on average results in n good units out—but traffic intensity is also increased by the same factor, unless offset by increased capacity.

If one product is made continuously, and if ade-

Table I. Yield and overproduction

m	p	n	Effective yield (.99+)	Overproduction units	Overproduction %
1	0.900	2	0.500	1.0	100.0
5	0.900	9	0.556	4.0	80.0
20	0.900	28	0.714	5.2	26.0
25	0.900	34	0.735	5.6	22.4
25	0.800	41	0.610	7.8	31.2
25	0.700	50	0.500	10.0	40.0
25	0.600	61	0.410	11.6	46.4
25	0.500	77	0.325	13.5	54.0
50	0.900	64	0.781	7.6	15.2
200	0.900	238	0.840	14.2	7.1
1000	0.900	1145	0.873	30.5	3.1

quate buffer stock is maintained to compensate for the variation in output, this strategy will work well enough to give satisfactory service performance. However, in a mixed-model JIT environment, a product is usually made in small lots, with other models. Each lot must fulfill a just-in-time requirement, ideally with little or no buffer stock.

In JIT, a reliable supply of product is important. If a lot of size $n=m/p$ is started, the probability of at least m good output units is about 50 percent. If we need at least m with 99+ percent probability for good service performance, we will need to start more. But then we will expect to get more good units on average, so we will be overproducing and saving the surplus for the next order. How many more units to start depends on the shape of the probability distribution of yield, and the cost of overproduction. The larger the variance of the yield distribution, the more starts are required to be confident of obtaining at least m good units; larger yield variance is often the result of systematic failures, e.g., resulting from processing errors that destroy many units at once, and that are not Bernoulli.

We can speak of an *effective yield* equal to m/n, for a given probability we will get at least m good. Starting n will result in expected *overproduction* of $np - m$ units, or $(np - m)/m$ percent. This process is sensitive to both yield and lot size. For a given service performance, the overproduction percentage is greater, and effective yield is lower, the lower the average yield and the smaller the lot size. This is illustrated in Table 1 for 0.99+ service performance and a case where individual failures are statistically independent (Bernoulli process).

For JIT, it is better to overproduce on a lot-by-lot basis than start too few units and expedite replacement units when shortages occur. The overproduction does not accumulate because we use it up before making more. Disruptions and additional variation in the downstream processes caused by the shortages may be a greater problem than the excess production required for JIT operation with imperfect yields. If demand for a given model is infrequent, the holding time for the overproduction may be long; if the demand is not repetitive, the overproduction is waste, leading to a direct trade-off for custom products between fast, make-it-all-at-once service, and cost, unless yields are improved.

Batching. It is commonplace in discrete manufacturing to group units into a lot and batch-process them at each operation. If the number of units in a batch is n, we model the batching process as one arrival with interval t, where t is the time between batches, followed immediately by $n - 1$ arrivals with zero interarrival time. If the average time between batches is $\bar{t}$, then:

$$\mu_a = \frac{\bar{t}}{n}$$

$$c_a^2 = n(1 + c_t^2) - 1,$$

where c_t^2 is the squared coefficient of variation of the time between batches. For example, if the 500 units per week were processed in daily batches of 100 units, $c_a^2 = 100 - 1 = 99$, even if $c_t^2 = 0$. This is over 100,000 times the c_a^2 the factory might attain through production control's efforts to limit the factory load variation to plus or minus 5 percent. Batching exacts enormous penalties in variation, interval, and inventory. It is the greatest single cause of long intervals.

Lot sizes often are significantly greater than the current demand for a particular product, causing the excess production to be stored. While a larger-than-needed batch is being produced, other urgently needed products may be waiting for machine time. These factors

complicate production control and material handling procedures, further aggravating the congestion caused by batching.

The usual reasons for batching are to attain reasonable use of machines despite lengthy setup procedures, and to minimize transportation costs. There are several mitigating actions the factory can take when implementing JIT:

- Reduce the need for setups by increasing common tooling and component usage in product design, and by using flexible manufacturing equipment.
- Do more setups to reduce *manufacturing lot sizes*—the amount produced between setups—to the smallest values consistent with available capacity. Limits to this process are discussed below.
- Reduce setup times to permit more setups and smaller manufacturing lot sizes.
- Put the operations close together, matching the closer distance with smaller, more frequent pickups. This also aids communication between work areas, helping them better achieve synchronized operations.
- Establish a *transport lot size*—the amount moved together from operation to operation—considerably smaller than the manufacturing lot size.
- Process and transport together different items used in matched sets.

Many operators consider setups arduous, exacting, and unpleasant, so there is an aversion to doing them; this feeling must be dealt with in any effective setup-reduction program. Where many setups are needed, as in flexible manufacturing cells, the operator's view of the job is that he or she manufactures setups. Thus, the setup process must be made routine, even automated, and *must be reliable*. A great deal of setup time—and scrap—is wasted in proving-in and adjusting setups because they were not right the first time.

Summary. We have looked at several causes of variation, compared their relative magnitudes, and have seen that, by averaging demand, it is possible to reduce the variation from independent random customer arrivals. We have also seen that several common manufacturing operations practices virtually nullify these efforts of production control, and that batching and disruptions from intermittent operations and yield losses are leading causes of variation. We have seen that it is sometimes possible to compensate for imperfect yields by calculated overproduction to attain reliable JIT operation. We also observe that the most important factors related to variation are under the control of factory personnel, not production control or the customers.

Setup Time, Interval, and Manufacturing Lot Size. The real benefits of lower setup time are that batch sizes can be reduced without increasing traffic intensity, and that scrap, rework, and delays from bad setups can be reduced.

It is a JIT cliché that a lot size of 1 is ideal. But how far do we need to go to get the benefits of smaller lots? What setups should we try to improve first? These are the questions addressed in this section. We will not cover specific setup reduction techniques; there is substantial literature and training material on the subject.[7]

How Far Should Lot Size Be Reduced? The ideal manufacturing-lot size is the amount requested by the customer at one time. The ideal transport-lot size through a multistage manufacturing process is one unit. We will explore next the practical issues of manufacturing lot sizing and setup times.

How Far Can Lot Size Be Reduced? Machine setup can be divided into *internal* and *external* phases. The internal setup refers to when the machine must be shut down; external setup includes activities that can be done off-line while the machine is running either the previous or the new part. In the context of this article, reducing setup times means reducing internal setup times. We will find the optimum setup time and lot size combination to minimize interval at a workcenter.

Setups use machine capacity to the extent of the internal setup time. External setups add cost, but do not directly affect capacity. Internal setup time is part of the average service time of the facility; thus, increasing the

number of setups per week increases average service time for the week's demand, thereby increasing traffic intensity. Because traffic intensity must be kept somewhat below 1 to prevent congestion, the number of setups we can perform is limited unless time per setup is reduced.

Considering only the effective time, T, that a machine is available for performing setups and running product during a work week, it is required that:

$$\rho = \frac{T_r + T_s}{T} < 1$$

where T_r is the total run time for the week's demand, and T_s is the total time spent doing setups during the week. For a fixed weekly demand, neither T nor T_r changes as either the number of setups or the time per setup changes. However, $T_s = S\tau_s$, where S is the number of setups per week, and τ_s is the internal time per setup, if we define a *lot* to be the amount produced between setups, τ_r to be the run time per lot, and μ_a to be the average time between lots, then the above expression reduces to the definition of ρ given earlier:

$$\rho = \frac{T_r + T_s}{T} = \frac{S\tau_r + S\tau_s}{S\mu_a} = \frac{\tau_r + \tau_s}{\mu_a} = \frac{\mu_s}{\mu_a}$$

The internal time per setup must satisfy the inequality:

$$\tau_s < \frac{T}{S} - \frac{T_r}{S} = \mu_a - \tau_r,$$

for the optimum target lot size to be feasible. This is equivalent to requiring $\mu_s < \mu_a$.

As we saw in Figure 4, we must stay well below $\rho = 1$ to avoid congestion when there is variation, so we would like to develop a useful guideline for how much setup time, or lot sizes, or both, should be changed, to minimize average interval of a lot through a workcenter, $\bar{T}$. To study this question, we first introduce two new dimensionless parameters, γ (gamma), and ρ':

$$\gamma = \frac{\tau_r}{\tau_s} \quad \text{and} \quad \rho' = \frac{\tau_r}{\mu_a} = \frac{T_r}{T}.$$

Note that for a fixed setup time, γ is proportional to τ_r, the run time per lot. Hence, γ is proportional to lot size, because it is assumed that the run time per unit is the same for all units in a lot of identical units. The parameter ρ' is called the *utilization* and is the fraction of available time spent running. Utilization is also the traffic intensity we would have if setup time were zero ($\tau_s = 0$).

The average interval of a lot through the workcenter, $\bar{T}$, is shown in Figure 4. The equation for Figure 4 for the special case of Poisson arrival and exponential service processes, $c_a^2 = c_s^2 = 1$, is:

$$\bar{T} = \frac{\mu_s}{1 - \rho}.$$

Because

$$\mu_s = \tau_r + \tau_s = \tau_r\left(1 + \frac{\tau_s}{\tau_r}\right) = \tau_r\left(1 + \frac{1}{\gamma}\right), \text{and}$$

$$\rho = \frac{\mu_s}{\mu_a},$$

we can substitute for ρ in the above equation for $\bar{T}$,

$$\rho = \rho'\left(1 + \frac{1}{\gamma}\right)$$

to obtain

$$\frac{\bar{T}}{\tau_s} = \frac{1 + \gamma}{1 - \rho' - \frac{\rho'}{\gamma}}.$$

This is graphed in Figure 7.

Notice that for a given utilization and setup time,

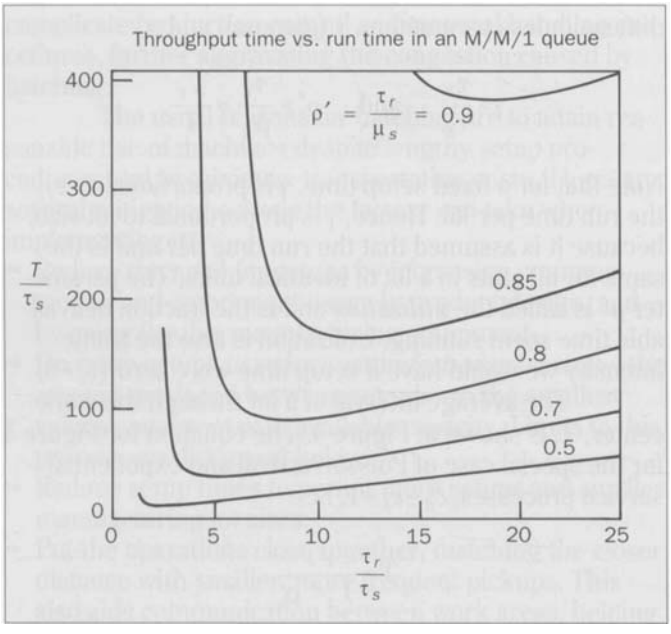

Figure 7. Effect of setup time and lot size on time in the facility.

there is a value of γ—and therefore lot size—that gives minimum interval. This optimum value of gamma, γ_o, occurs when:

$$\frac{1}{\gamma_o} = \frac{1}{\sqrt{\rho'}} - 1$$

Values of γ less than γ_o result in greater intervals from increased traffic intensity caused by too many setups. Greater intervals at values greater than γ_o are caused by the extra delay of a lot size that is too large. At $\gamma=\gamma_o$, the lot size is more than twice the lot size that would make $\rho=1$, providing some margin for variation. To optimize, we determine the changes in setup times and lot sizes required to make $\gamma=\gamma_o$. Any deviation from this optimum should be on the side of a too-large lot size ($\gamma \geq \gamma_o$), because congestion increases abruptly for values of γ less than γ_o. Using this criterion as our guideline, we find the ratio of setup time to run time per lot must satisfy the following inequality:

$$\frac{\tau_s}{\tau_r} \leq \frac{1}{\sqrt{\rho'}} - 1.$$

The optimum relationship between setup and run time occurs at the point of equality. The optimum total time per week spent setting up, T_{so}, is:

$$T_{so} = T_r \left(\left(\frac{T}{T_r} \right)^{1/2} - 1 \right) = \sqrt{TT_r} - T_r.$$

We wish to find a new setup time, τ_{so}, and a new number of setups per week, S_o, such that $T_{so} = S_o\ \tau_{so}$. This is done by changing either the number of setups per week (i.e., the manufacturing lot size), or the time per setup, or both. Any combination of lot size and setup time such that $S\tau_s = T_{so}$ will result in traffic intensity $\rho = \sqrt{\rho'} = \sqrt{T_r/T}$; under this condition, the interval, T, is proportional to setup time, τ_s.

Suppose a shop produces a variety of parts for an average of 500 units per week, and that the average usage is about 25 units. Our manufacturing lot size target is thus decided as 25 units, implying an average of 20 lots per week. Suppose the total run time of all parts processed by a particular machine averages 31 hours a week, and that the machine is available 36 hours a week for runs and setups. To find the maximum allowable setup time, we first determine the utilization of the machine to be $\rho' = 31/36 = 0.86$, giving $\gamma_o = 12.8$. The maximum allowable average internal setup time for each of the 20 lots is then determined to be:

$$\tau_s \leq \frac{31}{20}\left(\frac{1}{\sqrt{0.86}} - 1\right) = 0.1214 \text{hours},$$

or just over 7 minutes. With this value for τ_s, traffic intensity would be $\rho = \sqrt{.86} = .93$, and the average interval for a lot would be $\bar{T} = 23$ hours, or just under three working days, assuming that $c_a^2 = c_s^2 = 1$ for the *lots*; the average inventory of lots in process is $\bar{i} = \gamma_o = 12.8$ under this assumption. If the actual c^2's are less, the interval would be less—as little as 1.67 hours (the lot service time) if there were no variation in lot interarrival or service times. If the service times were always the same, but the lot arrival process was Poisson, the interval would be 12.7 hours. If there were no setup time, and no variation, the interval would be reduced to the service time for one unit, or 3 minutes 43 seconds, and 1.55 hours would be required to accumulate 25 units. The setup time and variation in this example have increased the interval at this workstation by a factor of 371 over single unit production, mostly attributable to variation.

What Setups To Reduce First? Generally, a setup reduction program first addresses machines having the greatest traffic intensity at the target manufacturing lot size. Certainly all machines where traffic intensity would be greater than 1 at the target lot size must be addressed.

Buffers

A *buffer* is a stock of inventory used to match unequal product flows. All inventory in the system is either at a work station, in transit, or in a buffer. Because most inventory is usually in the buffers, JIT might be viewed as the practice of managing buffers. Buffers can range from a single unit between two adjacent workstations, to an entire warehouse. The size of a buffer is determined more by the synchronization mismatch between production and usage than by the level of production or usage.

To estimate the average buffer inventory needed for good service performance, we first estimate the variance of buffer inventory, σ_i^2, and then use the criterion

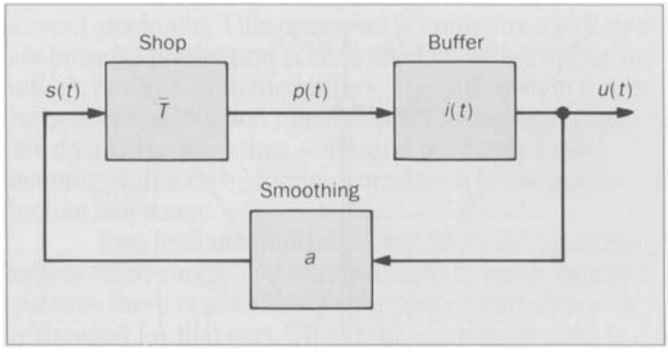

Figure 8. Production System (without forecasts).

that average inventory should be $3\sigma_i$, so less than 1 percent of the time would there be no stock when a demand arrives (i.e., a *stockout*). Inventory would then be expected to fluctuate between zero and $6\sigma_i$. If a lower probability of stockout were required, the average would be set to a higher multiple than $3\sigma_i$. More inventory than this is unnecessary.

We will discuss buffer behavior, the need to control buffers, how this control might be obtained, and proper buffer size. For simplicity, the theory will be explained in terms of stationary processes; however, it can be extended to non-stationary processes.

Dynamics of Buffers. Buffers are controlled by controlling their input and output. In Figure 8, $p(t)$ is the *production* stream, in units per unit of time; $u(t)$ is the *usage* stream; and $i(t)$ is the inventory, in units, in the buffer. These are assumed to fluctuate over time. The difference between the rate product flows into the buffer and the rate at which it leaves the buffer is the rate of change of buffer inventory, in units per unit time. We call this difference the *error* (or mismatch), and denote it by $e(t)$. Stated mathematically, this material balance condition is:

$$\frac{di}{dt} = p(t) - u(t) = e(t).$$

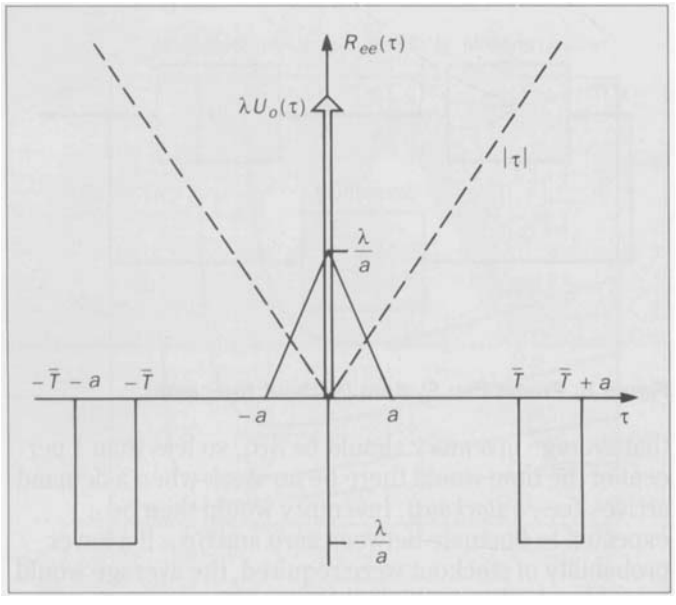

Figure 9. Autocorrelation of $e(t)=p(t) - u(t)$ in Figure 8, $h(t)$ =constant delay, $u(t)$ =Poisson process.

The buffer inventory is a function whose rate of change is $e(t)$. Negative values of $i(t)$ are interpreted as a back-scheduled quantity. For the average buffer inventory to be finite, the average of $e(t)$ must be zero; that is, $E\{e(t)\}=\mu_e=0$ ($E\{\}$ means *expected value*, or average). Because $e=p - u$, we conclude that a necessary condition for average buffer inventory to be finite is:

$$E\{p(t)\}=E\{u(t)\}.$$

For the average inventory to be finite, then, the average of the buffer input stream must equal the average of the output stream. However, because we are interested in controlling inventory fluctuations as well as average value, average control alone is inadequate. It can be shown that, to prevent infinite fluctuations in buffer inventory, a stronger condition is necessary:

$$\int_{-\infty}^{\infty} R_{ee}(\tau)\ d\tau=0,$$

where $R_{ee}(\tau)$ is the *autocorrelation function* of $e(t)$, defined as the expected value of the product $e(t+\tau)\,e(t)$. The condition $E\{p(t)\}=E\{u(t)\}$ is necessary but not sufficient for $R_{ee}(\tau)$ to integrate to zero, so this second condition includes the first, but is stronger.

Autocorrelation deals with *timing differences* between the buffer input and output streams. It measures the correlation of the value of $e(t)$ at any time, t, with its value at a different time, $t+\tau$. $R_{ee}(0)$ is the expected value of e^2, called the *mean square value*, or *average power* of $e(t)$ (from electrical engineering applications of the theory), and is always positive. Also, the maximum value of $R_{ee}(\tau)$ always occurs at $\tau=0$. With these general properties of autocorrelation functions in mind, we infer from this second condition that, since $R_{ee}(\tau)$ must integrate to zero, for every fluctuation in $e(t)$ (up or down), there must occur at other time(s) other, not entirely random, *offsetting* fluctuations in the opposite direction—negatively correlated with the initial fluctuation—to prevent random fluctuations from accumulating indefinitely.

Figure 9 shows the $R_{ee}(\tau)$ for the system of Figure 8, where demand is a Poisson process, the manufacturing interval is a constant delay, and factory starts are the average demand over the previous a time units. The negative correlations beginning at time $\tau=\bar{T}$ represent the reactions of the system, changing the buffer replenishment rate starting at time $\bar{T}$ after a demand fluctuation—withdrawal and replenishment being negatively correlated, because they affect inventory oppositely.

When $R_{ee}(\tau)$ integrates to zero, the variance of buffer inventory is:

$$\sigma_i^2=-\int_{-\infty}^{\infty} |\tau|\ R_{ee}(\tau)\ d\tau.$$

Note that autocorrelation functions of real processes are

always even, i.e., $R_{ee}(\tau) = R_{ee}(-\tau)$. To use an analogy from mechanics, if $R_{ee}(\tau)$ were seen as two mass distributions, one on the $+\tau$ axis and the other symmetrically disposed on the $-\tau$ axis, then σ_I^2 would be the distance between the centers of gravity of these two mass distributions (see Figure 9). The wider $R_{ee}(\tau)$ is, the larger is σ_I^2, i.e., the longer the times between a fluctuation in e and its negatively correlated offsetting fluctuation(s), the larger is σ_I^2, and the greater the inventory required. This has very important implications on the design of controls for integrated production and distribution systems, particularly with respect to needed temporal coupling between distribution center activity and factory activity.

To minimize inventory, the production system must operate to minimize the width of $R_{ee}(\tau)$. Minimizing inventory is desirable, but is not the control system's only design objective. To maximize utilization and factory efficiency, the control system and its buffer are also designed to control fluctuations in factory load. To achieve both objectives, it is necessary to minimize both the manufacturing interval and its variance. Systems with these characteristics have much in common with servomechanism control systems.[8]

In summary, the response of a successful JIT production and distribution system to a random demand is not random, but is closely correlated in time to even small fluctuations in demand. Dominant time constants in the response are often the average manufacturing interval, and the standard deviation of manufacturing interval. Just as in driving a car one makes frequent, small, *immediate* steering corrections to stay on track, so it is with a JIT system—which also must produce a smooth ride while following an often tortuous course.

Factory Buffers. Small buffers—between two adjacent work stations, or between feeder shops in the factory—usually are controlled in JIT by limiting their contents to a predetermined maximum value, and briefly interrupting production into the buffer, $p(t)$, until the usage has brought the buffer inventory below the limit. Conversely, production rates are increased as needed to prevent stockouts. This operation is known as a *pull system* because production is controlled by withdrawing, or *pulling*, product from the buffers. The pull system forces the average production rate to equal the average usage rate on a fairly short time scale, and positively limits inventory variance by forcing correlation between production and usage.

Two fundamentally different kinds of in-process buffers exist: single- and multiproduct. In single-product systems, there is a buffer for every part, controlled only by demand for that part. This is the simplest system to operate. But it is difficult to change (because of the number of parts usually involved) if the level of production changes. Of the two systems, the single-product system usually has the higher total inventory, because an inventory of each part is always maintained.

Multiproduct buffering requires one buffer per operation or workcenter. The buffer contains, in a specified order, a mix of the products produced in the sequence it is expected they will be used. Only parts needed in the next few minutes or hours are stored in the buffer, not all parts produced by the shop. Multiproduct buffering requires an information system to send a common production sequence to all feeders and the final-assembly shop. It requires greater shop discipline in sticking to the sequence—through better quality control, more reliable machines, and more disciplined management. It also requires fewer raw material shortages. Multiproduct buffers adapt themselves to changes in product mix, and naturally support mixed-model manufacturing in a flexible manufacturing cell.

Single-product buffering is appropriate when the number of different parts is not too large, when they are not used together in sets, or when shop conditions will not yet support multiproduct buffering. With a highly volatile product mix, or when several parts are used together in sets, multiproduct buffering and production control keep the parts together through the cell to their point of use in the required order, and with minimal inventory.

Kanbans. Visual sensing of buffer limits is com-

mon. That is, the buffers are located near (or at) the workstations that feed them. The operator watches the pile to see that it does not get too high or too low, thus matching his or her own operation rate to the next operation.

This approach was developed by Taiichi Ohno and Shigeo Shingo of Toyota Motor Company in the 1970s.[9] The visual signals are called *kanban* in Japanese. When dealing with feeder shops where the inventory is not all visible, kanban cards are surrogates for storage locations in the buffer. A kanban card will typically be imprinted with the part number, lot size, reorder point, producing location, using location, and a kanban serial number. The buffer itself is distributed throughout the production system, because the cards remain attached to the work in process until it is finished and used for its intended purpose. An unattached kanban card is equivalent to an empty slot in a buffer. It is interpreted as a manufacturing order to build a stated quantity of the item, and then take it to the indicated location.

There are several variations of kanban systems.[10] A popular one leaves the card attached to a material handling fixture, e.g., a cart or tub. When the finished parts are used, the empty carts are returned to the feeder shop buffer and are exchanged for a full cart. The empty cart in a feeder shop buffer is an order to produce more, as indicated by the card attached to the cart.

Pull System Rules. The rules for operating a pull system are few and simple:

1. Finished product is kept within visual contact of the producer whenever possible.
2. The producer owns the finished product until it is claimed by the customer, i.e., the next operation.
3. The customer gets what is wanted when it is wanted, carrying light loads and making frequent trips.
4. The producer is responsible for packaging product in the agreed way.
5. The customer is responsible for transportation, and for recording the transfers when necessary.
6. The producer must never let the customer run out of product, nor allow the inventory to grow beyond established limits.
7. The aim of the entire process is to run at a steady rate, all operations together.

Sizing Small Buffers. Appropriate sizes for small buffers are usually determined by observation; for new facilities, simulation modeling is sometimes used. When installing JIT, an initial buffer size is selected, based both on experience with and estimates of the expected duration of asynchronous disruptions between buffer production and usage streams, and on the amount of product on hand when the pull system starts. To avoid starving the next operation, the buffer must contain enough to supply it for the duration of "normal" disruptions experienced by the system; be reasonable, and do not overdesign the buffer sizes for catastrophes.

To see if the buffer size is correct, a chart of buffer contents is maintained next to it. The person making a withdrawal marks on the chart the number of containers of product that remain in the buffer. Over a few days or weeks, this record shows the range of fluctuations in the buffer. If the minimum buffer contents are greater than zero, the buffer size can usually be reduced by that minimum amount, so the future expected minimum is (occasionally) zero. The chart is also helpful in spotting opportunities to improve and reduce variation, if proximate causes of the peaks and dips shown are also noted on the chart for regular review by the shop quality team.

Buffer Placement. Factory output is limited by the output of the bottleneck operation in each product line. The bottleneck is the most heavily utilized operation. Other operations can catch up to some extent if interrupted, but the bottleneck is limited. It is critical that adequate inventory be kept in front of the bottlenecks so they are not starved by normal upstream disruptions. In some businesses, the bottleneck is in responding promptly to customer demand; thus, adequate buffer inventory in

the bottleneck distribution system is required. In other cases, no finished goods stock is required because everything is made to order, and the bottleneck is in the factory. "Bottleneck" is not pejorative; the term connotes an operation whose smooth, continuous operation is critical to profitable operations. It is easiest to run the factory when there is a clearly identified bottleneck, which is then exploited to its maximum potential.[11,12] As a rule, 1/3 to 1/2 the work in process in a well-run factory should be in front of the bottleneck.

Factory Loading and Flexibility

Good resource utilization is essential to successful manufacturing operations. High utilization of resources means achieving low variation in their application, from hour to hour during the work day. This is particularly critical with regard to the factory labor force. High utilization is obtained by smoothing the workload and building flexibility into the resources.

Flexibility means that resources have multiple applications. A measure of flexibility is the number of independent demands per week the factory can produce independently. We now examine the question of obtaining a level factory load when implementing JIT. This brings us to closure on the execution problem, as we see again that following the JIT paradigm leads us in the right direction.

Transforming a variable market demand into a smooth factory load unavoidably requires smoothing time and inventory holding time. The alternative is accepting a higher level of variation in factory load, and any implied inefficiencies. The task is to find the best compromise between inherent fluctuations of the market and the order interval the market will allow, on one hand, and the factory's flexibility and the amount of finished inventory held, on the other. In any case, minimizing manufacturing interval and maximizing flexibility are beneficial.

Stabilizing the Workload. Earlier we discussed that stabilizing factory load to about plus or minus 5 percent is desirable from the viewpoint of best utilizing the workforce. We also saw that this implied a coefficient of variation in the arrival process at the start of manufacture of $c_a^2 \approx 1/1200$. Here we examine alternatives for effective workload smoothing, and the rationale for the flexible factory.

If demand is a Poisson process with average rate λ; if the manufacturing interval is invariably $\bar{T}$; and if factory load is obtained by averaging the demand over the most recent time period a, as in Figure 8, then it can be shown, using the autocorrelation method discussed above, that the smoothed process will have:

$$\sigma_i^2 = 2\lambda\bar{T} + \frac{2}{3}\lambda a,$$

$$\overline{i_{WIP}} = \lambda\bar{T}, \quad \text{and}$$

$$c_p^2 = \frac{1}{\lambda a}.$$

Here, σ_i^2 is the variance of finished goods inventory, and c_p^2 is the squared coefficient of variation of the factory output rate. Note that $\bar{T}$ affects the required inventory, but does not affect smoothness of factory output.

Combining c_p^2 with our plus or minus 5 percent load fluctuation criterion, and assuming that $c_a^2 \approx c_p^2$ as before, we find that:

$$\lambda a \geq 1200.$$

Adequate factory load smoothness can be had in high-volume situations with short averaging times; low-volume situations require much longer smoothing intervals, a. For example, $\lambda = 1{,}200$ independent demands per week would require one week averaging time; 80 independent demands per week would require 15-weeks averaging time. For example, if a customer typically buys 100 units at a time, one independent demand would be 100 units. As the size of a typical independent demand increases, the time scale of the fluctuations lengthens.

Extreme examples of this are found in the heavy construction industry, where the time between independent demands ranges from months to years. We are most concerned here with independent demands that need a few hours' or days' capacity to fill.

If the required smoothing interval for good factory performance is much longer than the factory's desired customer-ordering interval, unless the demand forecast is good over horizon $\bar{T} + a$, great difficulty ensues trying to obtain simultaneously an adequately smooth factory load, low inventory, and high customer-service performance. Thus, it is desirable to make $\bar{T}$ as small as possible.

For best performance, the factory's combined order entry-plus-manufacturing-plus-distribution interval must be less than the necessary customer order interval. If this is not attained, additional finished goods inventory and reliance on forecasts for production control are necessary. Forecast accuracy is an issue wherever forecasts are used; bad forecasts are worse than none at all, leading to missed orders and excess production and inventory Generally, if the variance of forecast error at time $\bar{T} + a$ before due dates is greater than λ^2, it is better to use the demand history over horizon a, and ignore the forecast when determining factory starts.

Though smoothing the load before starting manufacture is a form of batching, there are some important differences. Batching the product on the shop floor increases physical inventory, investment, manufacturing interval, and ordering interval. Smoothing the load might lengthen the ordering interval, but only about half as much as batching. A batch not only suffers the delay of manufacturing interval, but it also must be accumulated before manufacture or held afterward for sale. A smoothed factory load presumably could be manufactured in a flexible factory in less time, with less investment.

To avoid batching, it is desirable to make factory flexibility at least as high as the number of independent demands per week, λ, of the market served by the factory. If factory resources are specialized to apply only to production of certain products, the formula $\lambda a \geq 1200$ applies separately to each limited product line in the factory. However, if the resources can be flexibly adapted to more products, the value of λ involved is the sum of the λs of all products to which the resources can be applied, and the common value of a can then be correspondingly less. This is the great motivation for flexible manufacturing.

Another way the effective value of λ can be increased, and a reduced, is to base production planning on eventual consumption of the product, to the extent possible, rather than on aggregations at distribution centers. Distribution centers commonly place aggregate demands on factories, and the factories are oblivious to the demands. This is a form of batching—creating dependent demands. Direct, rapid feedback of demands to factories could serve to reduce simultaneously load variation on the factory and required distribution center inventory—by making the factory more responsive to distribution activity.

Smoothing the Mix. Nowhere is achieving good resource utilization more difficult than in a multiproduct environment with a volatile product mix. Because all but the most flexible factories are limited in the rate they can efficiently adapt to changes in mix, mix stability complements flexibility in the factory. Though total workload stability depends on the vagaries of the marketplace, stabilizing the mix for short periods can be beneficial, and is more controllable by the factory.

Where production plans can be made ahead of time—from a few days to a few weeks—it is helpful to the factory to freeze the mix and operate the factory at a constant mix for as long as possible, before changing to another mix. In stabilizing the mix, the aim is to ensure that every day is an average day, every hour an average hour—for a while—before shifting to a new "average."

The factory typically has two opportunities, on

different time scales, to stabilize the mix: week-to-week, and within a week. If order intervals are a few weeks, then the factory, at the time of order acceptance, can promise delivery in a specific week, based on known capacity and previous commitments, and, consistent with customer needs, attempt to level the week-to-week load as much as possible.

It is also necessary to smooth the load within a week for best factory performance. Once the weekly production mix is decided, further smoothing is possible by mixing each week's work in a sequence so each work-center in the factory has nearly equal amounts of work to do in each few-hour period over the week. Executing this production sequence at a steady rate can then produce low-variation operation—if the factory has achieved good yield control, machine reliability, and raw material availability—and has the flexibility to follow such a sequence. The more flexible the factory, the more steady the load on each resource, leaving open the possibility of greater output because of a better match between flexibility and mix. The use of synchronized production-sequence lists between final-assembly and feeder shops is an invaluable help in synchronizing operations; use of such lists can also simplify production control, and lends itself to automating the production control process.

Like everything else in JIT, smoothing the mix also is easiest when manufacturing intervals are short.

Influence of Long and Variable Intervals. Factory managers go to great lengths to stabilize workload. To understand why it is often difficult to install JIT in the face of the pressures for workload stability, we will conclude by looking at the natural behavior of factory load and inventory where there is variation.

If the manufacturing interval, instead of being constant as assumed above, is normally distributed with mean $\bar{T}$, standard deviation σ_T, and coefficient of variation $c_T = \sigma_T/\bar{T}$, then, with a Poisson demand process and load smoothing time a as before, it can be demonstrated that:

$$\sigma_i^2 = 2\lambda\bar{T}\left(1+\frac{c_T}{\sqrt{\pi}}\right)+\frac{2}{3}\lambda a,$$

$$\overline{i_{WIP}} = \lambda\bar{T}, \quad \text{and}$$

$$c_p^2 = \frac{1}{\lambda(a + 2\sqrt{\pi}\,\sigma_T)}.$$

Notice that the effect of variation in manufacturing interval is like additional smoothing time in reducing variation in factory load.

This is a case where variation in interval appears to have a beneficial effect. The above equations show a perverse incentive factory managers have to operate with high levels of variation in manufacturing interval, and correspondingly high intervals and inventory—they can stabilize workload by introducing variation into the process. (In practice, we see that c_T is relatively independent of $\bar{T}$, i.e., that σ_T is nearly proportional to $\bar{T}$, so when $\bar{T}$ is increased, σ_T is also increased by about the same percentage.)

Because most factory managers, given a choice between a stable workload and long and variable intervals, will usually choose to stabilize the workload, it is not difficult to see how this happens in practice. Large factory inventories are held and work is continually expedited to focus always on the most urgent items. Coupled with the effects of batching, yields, machine breakdowns, part shortages, and asynchronous scheduling, the large inventories and expediting result in long, variable intervals—values of c_T between 0.2 and 0.5 and $\bar{T}$ between 3 and 8 weeks are commonplace—but smoother factory workload. Everyone is always busy, but an atmosphere of perpetual crisis pervades the factory.

One of the most difficult changes to be made in JIT is to escape this cycle. Stability must not be achieved by maintaining or increasing interval and its variation,

but in spite of their reduction. To do this, *simultaneous* improvements must take place on several fronts; addressing the list *serially* is ineffective. The major flow problems must be resolved and the factory loading process revised to achieve success with JIT. Often, factory organization and management structure and style need to change, and everyone needs to become involved in the change process. The JIT paradigm, focusing on reducing interval and its variation, is an effective guide to arranging priorities among proposed improvement projects, hastening the day when world-class performance is realized.

Conclusion

The JIT paradigm is an exceptionally powerful tool to produce a breakthrough in manufacturing performance. Based on a different philosophy of manufacturing operations, rather than on expensive technology, the method requires little initial capital investment. It requires an understanding of the underlying principles, a vision of the possibilities, and *commitment.* Diligent practice of quality control principles is integral to implementing JIT.[13,14] Teamwork and unity of purpose among the factory workforce is essential.[15] On the road to continuous improvement, the factory will likely want to invest in new floor layouts, information systems, and automation. In so doing, management is guided by the JIT paradigm into making decisions about what and when to automate and how to rearrange the layout to obtain maximum return on these investments.

Acknowledgments

I would like to acknowledge the people of the many AT&T manufacturing locations with whom it has been my privilege to associate. The insights they provided over the years we have worked together are the inspiration for this article. I would also like to thank Dan Krupka, John Svitak, and Ward Whitt for reviewing the manuscript, and for their helpful comments.

References

1. George Stalk, Jr. and Thomas M. Hout, *Competing Against Time—How Time-Based Competition is Reshaping Global Markets*, Macmillan–Free Press, New York, 1990.
2. R. B. Cooper, *Introduction to Queuing Theory*, North Holland Publishing Company, New York, 1981.
3. W. B. Davenport, Jr. and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, 1958
4. Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1984.
5. W. Kraemer and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queuing System GI/G/1," *Congressbook*, Eighth International Teletraffic Congress, Melbourne, Australia, 1976, pp. 235-1/8.
6. John D. C. Little, "A Proof for the Queuing Formula: $L=\lambda W$," *Operations Research*, Vol. 9, No. 3, 1961, pp. 383-387.
7. Shigeo Shingo, *A Revolution in Manufacturing: The SMED System*, Productivity Press, Stamford, Connecticut, 1985.
8. Norbert Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT Press, Cambridge, Massachusetts, 1949.
9. Shigeo Shingo, *Study of Toyota Production System from Industrial Engineering Viewpoint*, Japan Management Association, Tokyo, 1981.
10. Yasuhiro Monden, *Toyota Production System*, Institute of Industrial Engineers, Industrial Engineering and Management Press, Norcross, Georgia, 1983.
11. Eliyahu M. Goldratt and Jeff Cox, *The Goal: A Process of Ongoing Improvement*, 2nd edition, North River Press, Croton-on-Hudson, New York, 1986.
12. Eliyahu M. Goldratt and Robert E. Fox, *The Race*, North River Press, Croton-on-Hudson, New York, 1986.
13. *Statistical Quality Control Handbook*, First edition, 1977 printing, Western Electric Co., Inc., New York, 1956.
14. Kaoru Ishikawa, *Guide to Quality Control*, Asian Productivity Organization, Tokyo, 1982.
15. *The Canon Production System: Creative Involvement of the Total Workforce*, Japan Management Association, Productivity Press, Stamford, Connecticut, 1987.

(Manuscript received April 30, 1990)