OPERATIONAL ANALYSIS OF A JOB SHOP

Stephen C. Graves

Alfred P. Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

February 1984 WP1552-84

OPERATIONAL ANALYSIS OF A JOB SHOP

ABSTRACT

We propose and develop a discrete-time, continuous-flow model with linear control for studying the operation of a job shop that sees a stationary input mix of job types. We are not concerned with issues of detailed scheduling, but rather hope to develop a planning tool for a job shop operation. With the model we are able to characterize the operational behavior of each work center in the job shop for a given control policy. The control rule that we assume sets the production rate at a work center as a fixed proportion of its queue level in each time period. This control rule is consistent with the assignment of a planned lead time to each work center. For such control rules the model gives the steady-state distribution of the production levels at each work center, as well as the distribution of queue lengths. We show how to use the model not only to evaluate a specification of the control rules but also to find a good specification of the control rules that results in acceptable shop behavior.

1. Introduction

The intent of this work is to develop a model-based framework for performing an operational analysis of a complex batch or discrete-part manufacturing operation as typified by a job shop. The focus of the operational analysis is on understanding the interrelationship and interplay of the three key components in a manufacturing operation, namely the available production capacity, the inherent variability and uncertainty of the production requirements, and the level of work-in-process inventory. We are interested in understanding how job flow time, or equivalently the level of work-in-process (WIP) inventory, depends upon production capacity at each work center or production stage. Similarly, we want to understand how job flow time relates to the variability of production requirements that comes from the inherent job mix faced by the manufacturing operation. To do this we present and illustrate a mathematical model that permits such analyses in the context of a job-shop operation.

A job shop is a very flexible production facility that consists of a set of versatile machine centers or work stations, and is capable of producing a wide variety of jobs. The processing requirements for each job consist of an ordered set ot tasks where each task is to be performed on a distinct machine center. These processing requirements, as specified by the tasks, dictate how the job is routed through the machine centers in the job shop. Due to the wide variety of jobs (i.e. routings) processed by the shop, it may not be possible to discern any strong pattern in the flow of work through the shop. In particular, a machine center may receive jobs from several other machine centers; likewise, jobs at the work center may go next to one of several other work centers or may leave the shop if completed. Because of this lack of dominant work flows through the shop, production control is often very difficult in a job shop. Indeed, it is often not possible to have very sophisticated production control because of the complexity and variety of work flows.

Production control is often based on a queue management system. This approach views the job shop as a network of queues where each work center is a server and the jobs waiting there form its queue. For each work center we assign a planned lead time that represents the expected time, both waiting and in process, that a job will spend at that work center. Production control, in its crudest form, just prioritizes the jobs in queue at each work center, typically by means of some measure of the perceived urgency or criticality of the jobs. Job criticality is usually specified as a function of the difference between the need or promised date for the job and the projected completion date of the job based on the planned lead times (i.e. job slack). Jobs with the least slack get highest priority. The projected completion date reflects the processing time and expected queueing time (i.e. the planned lead times) for each remaining task for the job. This prioritization or sequencing of jobs in each queue is myopic since it is virtually impossible to anticipate fully how all jobs will complete their processing through the shop.

A somewhat more sophisticated scheme for queue management is input/output control (Wight, 1970). Here, the intent is to manage the flow of work through the shop so that the size of the queue at each work center remains relatively stable about a predetermined level. Clearly to do this, one needs to control the input rate to each work center to match the output rate. This is relatively straightforward for the work centers at which new jobs enter the shop; namely, new jobs are released to the shop at a rate in accordance with the production rate of the "gateway" work centers. However, it is not at all clear how to maintain input/output control at non-gateway work centers, especially if they receive input from multiple sources. Nor is it all clear how to determine the proper queue levels about which to target the input/output control.

In this paper we present a model that, in a particular way, formalizes Wight's concept of input/output control. In the model we describe a queue management system that relies on planned lead times, as is common in production practice. However, we are not concerned with detailed scheduling issues, and do not use the planned lead times to prioritize jobs; rather we use the planned lead times for planning the operation of the shop, and in particular to prescribe the production rates by work center by time period. This is done in order to achieve some level of input/output control at all work centers. As we will see, the planned lead times are the key decision parameters for implementing this form of input/output control.

Most of the previous research on job shops has been in the area of detailed scheduling, with two major thrusts. One thrust has been research that explores the performance characteristics of various myopic sequencing rules. This research has relied upon simulation studies that compare the performance of a prespecified set of sequencing rules on a particular job shop with a particular job mix; Conway, Maxwell and Miller (1967) give an excellent illustration of this type of work and review some of the earliest studies. A second research thrust has been to determine optimization methods for finding the best way to sequence a given set of jobs through the shop. This research views the scheduling of a job shop as a large combinatorial-optimization problem to which highly specialized solution procedures may be applied. Lageweg, Lenstra and Rinnooy Kan (1978) give a good illustration of this type of research, as well as provide a review of earlier work.

There has not been much work that has tried to step back from the very detailed issues of sequencing to consider the broader issues of planning in a job shop. A noteworthy exception is the work of Jackson (1957, 1963) on queueing networks. This work provided a model for characterizing the flows through a complex job shop. From this model one could get insights into the relevant planning tradeoffs between additional capacity, reduced flow times, and an altered job mix. Other work that has focused on planning issues is that of Jones (1973), Holstein and Berry (1970, 1972), Bertrand (1981), and Bertrand and Wortmann (1981). Jones gives an economic framework for considering the costs of idle resources, of carrying inventory, of missing due dates, and of making extended due date promises; the decision variables in the framework are the level of work-in-process inventory, the tightness of the promised due dates and the sequencing rule. Holstein and Berry explore the development of a work flow matrix to help identify the dominant flows in a job shop and to serve as a guide for smoothing the work flow in the shop. They also show how to use the work flow matrix to make labor assignments and transfers. Bertrand (1981) and Bertrand and Wortmann (1981) develop and apply a model that strives to control the flow time of jobs by controlling the aggregate work load in the shop. They model the behavior of the shop at a very aggregate level and provide a discrete-time analysis of the flow of jobs through the shop. In this respect their model and analysis are similar to that given in this paper.

There has recently been work that focuses on understanding the impact of lot sizing on shop behavior. Karmarkar(1983) proposes a simple queueing model to examine the relationships across lot sizes, manufacturing lead times and resource utilization. Zipkin (1983) also uses queueing models to model a production facility; he then develops an optimization framework that combines queueing considerations with inventory considerations in order to set lot sizes for a multi-item, batch production system. These papers are similar to the current paper in their recognition of the importance of understanding and controlling shop floor time. They focus on the use of lot sizing to control flow time, while the current paper does not consider lot sizing at all. Rather, the current paper uses production rates as the mechanism for control.

The current paper presents a new planning model for analyzing the operation of a job shop. We will try to argue that this model is a valuable addition to the existing array of planning models. The remainder of the paper is organized into three sections. The next section develops the model. The model represents the job shop as a continuous-flow, discrete-time system with linear control. Section 3 gives an illustrative example that shows how the model might be used to analyze the operations of a job shop. Section 4 gives a discussion of the model and its assumptions, indicates how the model might be generalized, and indicates how the model compares with alternative approaches.

2. Model Development

This section presents the model that underlies the intended operational analysis. First we present the assumptions of the model and develop the basic operational equations that describe the shop behavior. Next we provide the analysis of these equations that allows us to characterize the work flow. Finally we show how to determine the effect on the work flow from marginal changes in the shop parameters.

We base the analyses on a discrete-time model of the job-shop operation. Implicit in the model is an underlying time period that governs the transitions within the model. The model assumes that the movement of jobs from one work center to another, as well as the arrival of new jobs to the shop, can occur only at the start (or equivalently the end) of a time period; that is, a job completed during a time period at a work center moves to its next work center at the start of the next time period. Clearly, we must choose the time period carefully in order for the model to be a meaningful representation of the job shop under consideration. On the one hand, the time period should be short enough so that it would be highly unlikely that one job would move through two successive work centers during one time period. Yet, on the other hand, the time period should be long enough so that each work center is capable of completing a handful of tasks during each time period. (The reason for this statement will be clearer after we present the model.) The time period is clearly dependent upon the shop. In some shops the job movement may be such that a two-hour period is appropriate, whereas in other shops a multi-day period may best correspond to the way jobs move.

The model is a continuous-flow model in which we track work loads rather than jobs. As will be seen, we express the arrivals to the work center and the queue at the work center in terms of the backload (e.g. hours) for the work center rather than as the number of jobs. Similarly the production at the work center during a time period is given as the amount of work performed, not the number of jobs completed. Consequently, individual jobs have no identity in the model. This may be a serious drawback for certain instances; however we contend that for the purposes outlined earlier, concentrating on the aggregate work flow will be adequate.

We model each work center by describing a control rule that determines the amount of work performed by the work center in a time period; this control rule is

$$P_{it} = \alpha_i Q_{it} \tag{1}$$

where P_{it} is the production of work center i in time period t, and Q_{it} is the queue of work or backlog at the start of time period t. The parameter α_i, $0 < \alpha_i \leq 1$, relates the current backlog to the current production. In particular, the model states that production is a fixed portion (α_i) of the queue of work remaining at the start of the period. For instance, if $\alpha_i = .25$ then we say that each time period the work center produces one quarter of its queue; on average, a job would take four time periods ($1/\alpha_i$) to get through the work center.

This model of production at a work center treats the work center as if it had no capacity constraint. The model assumes that the work center is always able to complete the fixed portion α_i of its queue, regardless of the queue size. In some instances this may be a very strong assumption; however, I would raise a few points in support of this model. First, the choice of the parameter α_i is critical. As will be seen, this is a smoothing parameter. We set the parameter α_i such that the resulting time series for production is consistent with available capacity at the work center, i.e. we need set α_i so that we are assured that P_{it} is achievable most of the time. Second, the model asserts that the production rate varies directly with the queue length. This says that when the queue grows, the work center works harder, and vice-versa.. There is evidence, albeit primarily anecdotal, that complex shops behave in this manner, especially when production is both labor and machine-constrained (e.g. Gomersall 1964). As a queue builds at a work center, a manager will direct more resources to the work center to reduce the queue to normal levels. This may entail shifting workers to the heavily-loaded work center, or working overtime, or working more efficiently (e.g. postponing maintenance or other non-productive activities). SImilarly as a queue at a work center drops below its normal level, the manager may divert resources away from the work center. Labor may be shifted to other work centers, and more non-productive activities such as maintenance, training, and trial production will be undertaken.

Although we can view this model of work-center behavior as a descriptive model, we primarily think of it as being prescriptive of how a shop should be run. The model lends itself to cases where production control in the shop is based on planned lead times at each work center. If the planned lead time at a work center is n time periods ($n \geq 1$), then the work center, on average, must process 1/n of its queue each period. But this is what (1) does; the control rule prescibes that exactly 1/n of the queue be processed each time period, where $\alpha_i = 1/n$. Furthermore, we will see that this control rule not only is consistent with the planned lead time, but also acts to stabilize the work flows through the shop. Each work center behaves as a filter that smooths its arrival stream of work before passing the work onto other work centers. Indeed, we will argue that a shop ought to be managed in this manner.

Now to use (1) we need specify the queue level Q_{it}. The first step is to pose the standard balance equation

$$Q_{it} = Q_{i,t-1} - P_{i,t-1} + A_{it} \tag{2}$$

where A_{it} is the amount of work that arrives at work center i at the start of time period t. By using (1) to replace Q_{it} in (2), we obtain

$$P_{it} = (1-\alpha_i)P_{i,t-1} + \alpha_i A_{it}, \tag{3}$$

which is a simple smoothing equation. By repeated substitution, we can then write production as

$$P_{it} = \sum_{s=0}^{\infty} \alpha_i (1-\alpha_i)^s A_{i,t-s} \tag{4}$$

where we assume we have an infinite history of arrivals. Thus we see that the production model given by (1) is effectively a simple smoothing function where the output time series (production) is just a smoothed version of the input time series (arrivals). If we can characterize the arrivals to the work center, then we can characterize the production. For instance, if the elements of the time series $\{A_{it}\}$ are i.i.d. random variables with mean μ and variance σ^2, then we find that

$$E\{P_{it}\} = \mu,$$

and

$$\mathrm{Var}\{P_{it}\} = \frac{\alpha^2\sigma^2}{2\alpha-\alpha^2} .$$

Unfortunately, though, the arrival stream to a work center in a job shop will not consist of i.i.d. random variables. Rather the arrival stream will tend to be highly correlated over time, as will be seen. Consequently, a more complex derivation is needed to characterize the time series $\{P_{it}\}$.

The arrival stream to a work center is comprised of two types of flows. One flow consists of new jobs entering the shop that have their first processing step (task) at this work center. The second flow consists of jobs in process that have just completed a processing step at another work center and have their next processing step at this work center. We describe the arrival process to each work center from each other work center by

$$(5) \qquad A_{ijt} = \phi_{ij} P_{j,t-1} + \varepsilon_{ijt}$$

where A_{ijt} is the amount of work arriving to work center i from work center j at the start of time period t, ϕ_{ij} is a positive scalar, and ε_{ijt} is a random variable with zero mean. That is, every time unit (e.g. hour) of production at work center j generates ϕ_{ij} times units (hours) of input to work center i, on average. The term ε_{ijt} is an error or noise term that introduces uncertainty into the relationship between production at work center j and inputs to work center i. We assume that for each pair (i,j) the elements in the time series $\{\varepsilon_{ijt}\}$ are i.i.d.

We offer two comments with regard to (5). First, we have made a strong assumption here that we can model the work flow using a Markov property. That is, we assume that the arrivals to work center i from work center j do not depend on how that work got to j. In essence, we assume that each time period work center j processes a relatively stable or representative mix of jobs, so that subsequent inputs to downstream work centers are similarly stable. The validity of this assumption depends upon overall stability of the job mix in the shop, as well as the length of the time period. If the job mix varies drastically (in terms of production requirements by work center) from one week to another, then the assumption may not be very good. Similarly, if the period length is such that at most only a few jobs are completed at the work center each time period, then it may be unlikely that there is a very stable output.

The second comment concerns how uncertainty or noise enters the relationship between production at j and inputs to i. One might argue that the noise should be proportional to the volume of production at work center j; namely, we might expect with greater production volume, we would have greater variability in the input stream to i. In (5) the noise term is independent of the production level. As will be seen, this assumption permits a great deal of tractability in analyzing the model. Clearly the validity of this assumption would have to be examined in the light of actual shop data.

Now the arrival stream to work center i is given by

$$A_{it} = \sum_j A_{ijt} + N_{it} \tag{6}$$

where N_{it} is a random variable that represent the work load from new jobs that enter the shop at time t and have their first processing step at work center i. We assume that for each work center i the elements of the time series $\{N_{it}\}$ are i.i.d. We now substitute (5) into (6) to get

$$A_{it} = \sum_j \phi_{ij} P_{j,t-1} + \varepsilon_{it} \tag{7}$$

where

$$\varepsilon_{it} = N_{it} + \sum_j \varepsilon_{ijt} .$$

Thus ε_{it} represents the part of the arrivals that are not predictable from the production levels of the previous periods, i.e. the new arrivals and the noise in the flows from other work centers. By assumption, the elements of the time series $\{\varepsilon_{it}\}$ for each work center are i.i.d. Note that the expected values of ε_{it} equals the expected amount of new arrivals each time period to work center i.

We are now ready to perform the analysis of the job shop model specified by equations (3) and (7). It will be convenient to rewrite these equations in vector notation as

$$(3') \qquad \underline{P}_t = (\underline{\underline{I}} - \underline{\underline{D}})\underline{P}_{t-1} + \underline{\underline{D}}\,\underline{A}_t$$

$$(7') \qquad \underline{A}_t = \underline{\underline{\Phi}}\,\underline{P}_{t-1} + \underline{\varepsilon}_t$$

where $\underline{P}_t = \{P_{1t},\ldots,P_{nt}\}'$, $\underline{A}_t = \{A_{1t},\ldots,A_{nt}\}'$, $\underline{\varepsilon}_t = \{\varepsilon_{1t},\ldots,\varepsilon_{nt}\}'$ are column vectors of random variables, n is the number of work centers, $\underline{\underline{I}}$ is the identity matrix, $\underline{\underline{D}}$ is a diagonal matrix with $\{\alpha_1,\ldots,\alpha_n\}$ on the diagonal, and $\underline{\underline{\Phi}}$ is an n-by-n matrix with elements ϕ_{ij}. By substituting (7') into (3') we obtain

$$(8) \qquad \underline{P}_t = (\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\underline{\underline{\Phi}})\underline{P}_{t-1} + \underline{\underline{D}}\,\underline{\varepsilon}_t \quad .$$

By repeated substitution we can rewrite (8) as a geometric series

$$(9) \qquad \underline{P}_t = \sum_{s=0}^{\infty} (\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\underline{\underline{\Phi}})^s\, \underline{\underline{D}}\,\underline{\varepsilon}_{t-s}$$

where we assume an infinite history of the system exists. We use (9) to characterize the joint distribution of the production vector $\underline{P}_t$. To do this, we let the noise vector $\underline{\varepsilon}_t$ have mean $\underline{\mu} = \{\mu_1,\ldots,\mu_n\}'$ and a covariance matrix given by $\underline{\underline{\Sigma}} = \{\sigma_{ij}\}$. We note from the definition of ε_{it} that its mean, μ_i, corresponds to the expected amount of new arrivals to work station i, that is $\mu_i = E\{N_{it}\}$.

The expectation of the production vector, call it $\underline{\rho} = \{\rho_1, \ldots, \rho_n\}$, is given by

$$(10) \qquad \underline{\rho} = \sum_{s=0}^{\infty} (\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\underline{\underline{\Phi}})^s \underline{\underline{D}}\underline{\mu}$$

$$= (\underline{\underline{I}} - \underline{\underline{\Phi}})^{-1} \underline{\mu}$$

provided the spectral radius (maximal absolute eigen value) of $\underline{\underline{\Phi}}$ is less than 1. If the spectral radius of $\underline{\underline{\Phi}}$ is greater than or equal to 1, then the above power series diverges and the expectation of the production vector is not defined [see appendix for details]. This condition on $\underline{\underline{\Phi}}$ is the standard requirement on an input/output matrix, namely a unit of work at any work center (input) cannot ultimately result in more than one unit of additional work (output) at that work center. If this condition is violated, then the system does not reach a steady state but "blows up" over time (i.e. infinite queues). Finally, we note that the existence of a steady state does not depend on the smoothing parameters $\{\alpha_i\}$, but is entirely determined by the matrix $\underline{\underline{\Phi}}$. As will be seen, the smoothing parameters just influence the variance of $\underline{P}_t$, and do not affect its mean as might be expected after a little thought.

We find from (9) the covariance matrix of $\underline{P}_t$, call it $\underline{\underline{S}} = \{s_{ij}\}$, to be

$$(11) \qquad \underline{\underline{S}} = \mathrm{Var}(\underline{P}_t) = \sum_{s=0}^{\infty} \underline{\underline{B}}^s \underline{\underline{Z}}_o \underline{\underline{B}}'^s$$

where

$$(11a) \qquad \underline{\underline{B}} = \underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\underline{\underline{\Phi}}$$

and

$$(11b) \qquad \underline{\underline{Z}}_o = \underline{\underline{D}} \underline{\underline{\Sigma}} \underline{\underline{D}}$$

We show in the appendix that the power series again converges provided that the spectral radius of $\underline{\underline{\hat{C}}}$ is less than one. Now we can simplify (11) if $\underline{\underline{B}}$ has a set of distinct eigen values; if this is true then we can diagonalize $\underline{\underline{B}}$ so that

$$(12) \qquad \underline{\underline{B}} = \underline{\underline{P}}\,\underline{\underline{\Lambda}}\,\underline{\underline{P}}^{-1}$$

where $\underline{\underline{\Lambda}}$ is a diagonal matrix with the eigen values of $\underline{\underline{B}}$, $\{\lambda_1,\ldots,\lambda_n\}$, on the diagonal and $\underline{\underline{P}}$ is the corresponding matrix of eigen vectors for $\underline{\underline{B}}$. By substituting (12) into (11) we can reexpress $\underline{\underline{S}}$ as

$$(13) \qquad \underline{\underline{S}} = \underline{\underline{P}}\,\underline{\underline{C}}\,\underline{\underline{P}}'$$

where $\underline{\underline{C}} = \{c_{ij}\}$ is such that

$$(14) \qquad c_{ij} = \hat{c}_{ij}/(1-\lambda_i\lambda_j)$$

where

$$(15) \qquad \underline{\underline{\hat{C}}} = \{\hat{c}_{ij}\} = \underline{\underline{P}}^{-1}\,\underline{\underline{Z}}_0\,\underline{\underline{P}}'^{-1} \quad .$$

Hence, once we diagonalize $\underline{\underline{B}}$ as in (12), then we can immediately find the covariance matrix $\underline{\underline{S}}$ from (13) - (15).

An alternate approach to evaluate $\underline{\underline{S}}$ is to approximate the infinite series in (11) by a finite series. To do this, first define $\underline{\underline{S}}$ as the sum of the first n terms, i.e.

$$\underline{\underline{S}}_n = \sum_{s=0}^{n-1} \underline{\underline{B}}^s\,\underline{\underline{Z}}_0\,\underline{\underline{B}}'^s \quad .$$

Then we see that

$$(16) \qquad \underline{\underline{S}}_{2n} = \underline{\underline{B}}^n\,\underline{\underline{S}}_n\,\underline{\underline{B}}'^n + \underline{\underline{S}}_n .$$

By repeated application of (16) we quickly obtain a very good estimate of $\underline{\underline{S}}$; for instance, six applications gives the sum of the first 64 terms in the series.

In addition to $\underline{P}_t$ we will find it useful to characterize the distribution of the queue levels at each work center. From (1) we see immediately that

$$\underline{Q}_t = \underline{\underline{D}}^{-1} \underline{P}_t \tag{17}$$

so that

$$E(\underline{Q}_t) = \underline{\underline{D}}^{-1} \underline{\rho} \tag{18}$$

and

$$\operatorname{Var}(\underline{Q}_t) = \underline{\underline{D}}^{-1} \underline{\underline{S}} \, \underline{\underline{D}}^{-1} \tag{19}$$

where $\underline{\rho}$ and $\underline{\underline{S}}$ are given by (10) and (11). But we may also desire to describe the make-up of the queue in order to measure the waiting time at each work center. To do this we define Q^m_{it} to be the amount of queue at work center i at time period t that has been in queue for at least m periods. Assuming that we process the oldest part of the queue first, i.e. FIFO, then we define for $m \geq 1$

$$Q^m_{it} = Q^{m-1}_{i,\, t-1} - P_{i,t-1} \; ; \tag{20}$$

that is, the queue at time t that is age m or older, is just the queue at time t-1 that is age m-1 or older minus production in time period t-1. For m = 0 we have $Q^0_{it} = Q_{it}$ as given by (1) and (2). We see from (20) that we permit Q_{it}^{m} to take on negative values; this denotes not only that none of the current queue has been there for m periods, but also that the work center has processed an amount of work, equal to $-Q_{it}^{m}$, of more recent arrivals. In matrix notation we rewrite (20) and simplify to find for $m \geq 1$

$$(21) \qquad \underline{Q}_t^m = \underline{Q}_{t-1}^{m-1} - \underline{P}_{t-1}$$

$$= \underline{Q}_{t-m}^{0} - \sum_{s=1}^{m} \underline{P}_{t-s}$$

where $\underline{Q}_t^0 = \underline{Q}_t$ is given by (17). Thus, by substituting (17) into (21) we obtain

$$(21') \qquad \underline{Q}_t^m = \underline{\underline{D}}^{-1}\underline{P}_{t-m} - \sum_{s=1}^{m} \underline{P}_{t-s}$$

so that the queue is expressed entirely in terms of the production random vectors. From (9) it is clear that we can rewrite (21') as an infinite series in the i.i.d. random vectors $\underline{\varepsilon}_t$. From this representation and after a certain amount of algebra, we can find that

$$(22) \qquad E(\underline{Q}_t^m) = (\underline{\underline{D}}^{-1} - m\underline{\underline{I}})(\underline{\underline{I}} - \underline{\underline{C}})^{-1}\underline{\mu}$$

and

$$(23) \qquad Var(\underline{Q}_t^m) = \sum_{j=1}^{m-1} (\underline{\underline{I}} + \ldots + \underline{\underline{B}}^{j-1})\, \underline{\underline{Z}}_0 (\underline{\underline{I}} + \ldots + \underline{\underline{B}}^{j-1})'$$

$$+ (\underline{\underline{D}}^{-1} - \underline{\underline{I}} - \underline{\underline{B}} - \ldots - \underline{\underline{B}}^{m-1})\, \underline{\underline{S}} (\underline{\underline{D}}^{-1} - \underline{\underline{I}} - \underline{\underline{B}} - \ldots - \underline{\underline{B}}^{m-1})'$$

where $\underline{\underline{S}}$, $\underline{\underline{B}}$ and $\underline{\underline{Z}}_0$ are defined in (11), (11a) and (11b). Knowledge of the distribution of $\underline{Q}_t^m$ will permit us to get some notion of how long work waits in queue at each work center, as we will see in the next section.

If we now assume that the noise vector $\underline{\varepsilon}_t$ has an i.i.d. normal distribution with mean $\underline{\mu}$ and covariance matrix $\underline{\underline{\Sigma}}$, then we have that $\underline{P}_t$ is also normally distributed with mean $\underline{\rho}$ and covariance matrix $\underline{\underline{S}}$ given by (12)-(15). (Similarly we see that $\underline{Q}_t^m$ is normally distributed with mean and variance given by (22) and (23) for $m \geq 1$, and by (18) and (19) for $m = 0$.) We can use this information to assess the performance of the job shop. In particular we are interested in assessing whether the work flow that results from the choice of the parameters $\{\alpha_i\}$ is consistent with the available capacity at each work center. In general, specification of α_i corresponds to setting a planned lead time for work center i equal to $1/\alpha_i$. On the one hand, we desire for these parameters to be set large so that the lead times are as small as possible and the work-in process inventory is minimal. On the other hand, we also want the production requirements at each work center to be as smooth as possible in order to utilize available resources efficiently. But this suggests setting the smoothing parameters at small values. Hence we intend to use the above model as a guide to looking primarily at the tradeoff of smoother flow and better resource utilization versus shorter lead times and lower WIP inventory. Furthermore, we will use the model to discern the benefits from reducing the uncertainty or noise in the work flow. Since the tradeoff between resource utilization and inventory is largely a consequence of the uncertainty in the work flow, one must be able to assess explicitly the ramifications of the various sources of this uncertainty.

Since we are concerned with the consequences that arise from changes in the system parameters, it is of value to compute the appropriate derivatives. We let $\partial \underline{\underline{S}}/\partial\alpha_k$ denote the matrix whose (i,j) element is $\partial s_{ij}/\partial\alpha_k$. Then from (11) we find that

$$\frac{\partial \underline{\underline{S}}}{\partial \alpha_k} = \sum_{s=0}^{\infty} \underline{\underline{B}}^s \underline{\underline{Z}}_k \underline{\underline{B}}'^s \tag{24}$$

where $\underline{\underline{B}}$ is given in (11a) and

$$\underline{\underline{Z}}_k = \underline{\underline{E}}_{kk}[(\underline{\underline{\Phi}}-\underline{\underline{I}})\underline{\underline{S}}\underline{\underline{B}}' + \underline{\underline{\Sigma}}\underline{\underline{D}}] + [\underline{\underline{B}}\underline{\underline{S}}(\underline{\underline{\Phi}}'-\underline{\underline{I}}) + \underline{\underline{D}}\underline{\underline{\Sigma}}]\underline{\underline{E}}_{kk}$$

where $\underline{\underline{E}}_{ij}$ is a matrix of all zeroes except for a one in element (i,j). We note that the infinite series in (24) is the same as that in (11), except that $\underline{\underline{Z}}_o$ is replaced by $\underline{\underline{Z}}_k$. Hence we can find $\partial \underline{\underline{S}}/\partial\alpha_k$ by the same manner as we find $\underline{\underline{S}}$, but with $\underline{\underline{Z}}_o$ replaced by $\underline{\underline{Z}}_k$. From this observation it is easy to see that once we have found $\underline{\underline{S}}$ [for instance, by performing the diagonalization in (12)], we immediately can obtain $\partial \underline{\underline{S}}/\partial\alpha_k$ for any k.

We may also have interest in changes to the covariance matrix $\underline{\underline{S}}$ with respect to changes in an element ϕ_{ij} in the input/output matrix $\underline{\underline{\Phi}}$ and to changes in an element σ_{ij} in the covariance matrix for the noise process, $\underline{\underline{\Sigma}}$.[1] If we let $\partial \underline{\underline{S}}/\partial\phi_{ij}$ and $\partial \underline{\underline{S}}/\partial\sigma_{ij}$ denote these derivatives, then we find that

$$\frac{\partial \underline{\underline{S}}}{\partial \phi_{ij}} = \sum_{s=0}^{\infty} \underline{\underline{B}}^s \underline{\underline{Y}}_{ij} \underline{\underline{B}}'^s$$

and

$$\frac{\partial \underline{\underline{S}}}{\partial \sigma_{ij}} = \sum_{s=0}^{\infty} \underline{\underline{B}}^s \underline{\underline{X}}_{ij} \underline{\underline{B}}'^s$$

[1] Since the covariance matrix $\underline{\underline{\Sigma}}$ is necessarily symmetric we define the derivative of $\underline{\underline{S}}$ to be in terms of changes in both ϕ_{ij} and ϕ_{ji} for $i \neq j$.

where

$$\mathbf{Y}_{ij} = \mathbf{D}\,\mathbf{E}_{ij}\,\mathbf{S}\,\mathbf{B}' + \mathbf{B}\,\mathbf{S}\,\mathbf{E}_{ji}\,\mathbf{D}$$

and

$$\mathbf{X}_{ij} = \begin{cases} \mathbf{D}(\mathbf{E}_{ij} + \mathbf{E}_{ji})\mathbf{D} & \text{for } i \neq j \\ \mathbf{D}\,\mathbf{E}_{ii}\,\mathbf{D} & \text{for } i = j\ . \end{cases}$$

Hence these derivatives also have the same form as $\mathbf{S}$ and can be computed immediately once $\mathbf{S}$ is determined from (13) - (15).

3. Example

We illustrate the model here with a small example. The work flow matrix is derived from a work cell at a factory that produces grinding machines. The work cell fabricates several families of spindles that are components in the assembly of the grinding machines. The work cell consists of ten types of machining stations, and the work flow is described by the matrix $\underline{\underline{\Phi}}'$ given in Table 1. We note that the work flow matrix is quite sparse, indicating that there are dominant flows through the work cell. In particular, all work enters the shop at the first work center (lathe) and then moves serially through the shop but with some recycling. We remind the reader that the work flow matrix is not a probability matrix but rather its elements indicate the expected amount of work generated at a subsequent station by a fixed amount of work at the current station. For example, one hour of work at station 8 generates, on average, 3.43 hours of work at station 9.

In addition to the matrix $\underline{\underline{\Phi}}$ we need to specify the time period for the model, the vector of expected inputs $\underline{\mu}$ and the covariance matrix $\underline{\underline{\Sigma}}$. For this example we set the nominal time period to be two hours. Work enters the shop only at station 1 and we assume that the expected input is four hours of new work every time period (i.e., $\mu_1 = 4$, $\mu_i = 0$ $i = 2, \ldots 10$). We note that there are three identical lathes at work center 1 so an average input of four hours of work each two hours is not obviously infeasible; all other work stations have a single machine. We assume the noise process $\{\varepsilon_{it}\}$ is normally distributed with its covariance matrix $\underline{\underline{\Sigma}}$ being a diagonal matrix as specified in Table 1. We note that most of the uncertainty is introduced at the work center 1, presumably by the stream of new arrivals; however, the arrival streams to the other work centers also are subject to a noise process, but with smaller variances.

Given $\underline{\Phi}$ and $\underline{\mu}$ we can compute the expected work load for each work center by (10). We report this in Table 1. We see that work centers 1, 6, 9 and 10 are the most heavily utilized centers. For a nominal time period of two hours, the utilization at work center 1 is 83% since it consists of three lathes. Work center 6 also has a utilization of 83%, and work center 9 has a utilization of 90%, while work center 10 has a utilization of 110%. Indeed this analysis indicates that work center 10 need do, on average, 2.19 hours of work every two hours. This seems impossible given that there is only one precision grinding machine available at this work center. Yet this is what is required to meet the production requirements. Although the model cannot prescribe how to meet this seemingly impossible requirement, it does assist in identifying the necessary resource requirements. In particular, one might expect that this work center will work a ten-hour day while all of the other work centers work eight-hour days; hence, the effective time period for work center 10 might actually be 2.5 hours rather than two hours. The model should help to assess whether or not ten hours per day is sufficient to cover the variability in these production requirements.

We are now ready to consider several different scenarios for managing the flow of work through the shop. We specify a scenario by setting the smoothing parameters α_i or equivalently setting the planned lead times $n_i = 1/\alpha_i$ for each work station. For the first case we set the planned lead time for each work station to be one period ($n_i = 1$); that is, at each work station all work that arrives by the start of a time period is to be processed by the end of that time period. Table 2 gives the characterization of the shop behavior for this case. For each work center we report the expected work load and its standard deviation, the expected queue at the start of a period, and the expected backlog at the start of a period. We define the backlog at work center i to be the amount of the queue that has been in queue at least n_i periods; but this is just the positive part of Q_{it}^m for $m = n_i$, given by (20). From Table 2, we see that the work-in-process levels are quite low and that there is never any backlog since each work center clears its queue each time period. However, the production requirements for each work center are highly variable. For instance, for work center 1 the production requirement per time period has a normal distribution with mean 5.01 hours and a standard deviation of 2.02 hours. Hence, with probability .31 the production requirements for a time period exceed the nominal production capacity of 6.00 hours, in which case overtime would be worked or additional resources would be directed to this work center. Similarly, we see that the other bottleneck work centers have highly variable production requirements that will tend to result in inefficient production and high costs due to their lack of smoothness.

For the second case in Table 3 we attempt to smooth the production requirements at the heavily loaded work centers by imposing a planned queue. We plan a lead time of four periods at work center 1, two periods at work center 6, and three periods each at work centers 9 and 10. This results in longer queues at these work centers as well as larger backlogs. For instance, at work center 9, increasing the planned lead time from one to three time periods triples the size of the queue. The expected backlog at work center 9 increases from zero to .13 hours; that is, on average the queue will contain .13 hours of work that has been in queue for three or more time periods. Since the average production per period at this work center is 1.89 hours, this suggests that roughly 7% of the work takes longer than the planned lead time of three time periods. (Due to the symmetry of the normal distribution, a comparable amount of work, i.e., 7%, takes less than the planned lead time of three time periods.) But the additional queues do result in significant production smoothing as reflected in the smaller standard deviations for production at the bottleneck work centers.

The third case reported in Table 4 is a continuation of the second case in attempting to smooth the work flows. Again we have added queues at the heavily loaded work centers to make their production requirements less variable. But we begin to see here the effects of the decreasing marginal benefits from additional smoothing. For instance, for work center 6 increasing the planned lead time from one to two periods reduces its production standard deviation by 56%, while increasing the planned lead time from two to three periods gives only a 25% reduction in the production standard deviation. (This is not an entirely fair comparison since the reduction in the variability of the production requirements is not only a consequence of the increased lead time at the work center, but also results from the smoother arrival stream to that work center from the other work centers).

The purpose of the fourth case (Table 5) is to show that we can smooth production not only by placing a queue as a buffer at a work center, but also by smoothing the arrival stream seen by the work center. We note from the Φ matrix that work center 9 only receives work from work center 8. In the previous two cases we try to smooth the work load at 9 by imposing a queue there; alternatively we could smooth the arrivals to work center 9 by smoothing the production at 8. This is apparent not only from the above reasoning, but also from computing the derivative of the variance of production at work center 9 taken with respect to the smoothing parameters for work center 8 [equation (24)]. In Table 5 we have increased the planned lead time at work center 8 from one to two time periods; this results in a smoother arrival stream to work center 9 that allows us to reduce its planned lead time from five to four time periods with no increase in its production variability. Hence, we can begin to see how the control at one work center impacts the work flow at another work center.

This example, as described by the four cases, illustrates a type of analysis that one would do with the model. The analysis, as presented, allows one to explore for a given shop configuration the tradeoff between short flow times and low work-in-process inventory versus smooth production and efficient allocation of production resources. As we have seen, attempting to smooth production results in longer queues, and longer and more variable flow times. Similarly, attempts to prune work-in-process or to speed up the work flow will lead to more variable production requirements, if we assume all else is unchanged. We have not prescribed a formal mechanism for doing this exploration, although we have found reference to the derivative matrices $\{\partial \underline{\underline{S}} / \partial \alpha_k\}$ to be most helpful in guiding the search.

The model framework should also be helpful in doing other types of analyses. In particular, we could examine a variety of "what if" questions: What if we had more/less capacity at various work centers? What if the job mix or flow structure changes? What if we had better control over the input stream to the shop so that the arrivals were less uncertain? What if through improved scheduling we could reduce the variability in the work flows between work centers? Indeed, we expect that the model can be a valuable planning tool for designing and assessing control strategies under a variety of environmental conditions.

TABLE 1: Work Flow Matrix for Example

WORK CENTER (from \ to)		1	2	3	4	5	6	7	8	9	10
1 (lathe)			.15	.04	.01	.03	.24		.01		
2 (copy lathe)				.01	.41	.37					
3 (drill press)		.11				1.36					
4 (milling)				.71							
5 (rough grinder)		.68					.15	.10			
6 (internal grinder)				.06					.22		
7 (thread cutting)									1.00		
8 (hole abrading)										3.43	
9 (precision grinder)				.07			.13				1.16
10 (ultra-precision grinder)											
Input	μ_i	4.00									
Covariance matrix	Σ_{ii}	4.00	.01	.01	.01	.04	.04	.00	.01	.04	.04
Work load	ρ_i	5.01	.75	.69	.36	1.37	1.65	.14	.55	1.89	2.19

Notes: (i) 2 hour time period

(ii) 3 machines at work center 1

(iii) work flow has been simplified by ignoring heat treatment operations that occur outside the work cell.

TABLE 2: CASE A

	planned lead time n	expected production $E(P)$	standard deviation $\sigma(P)$	expected queue $E(Q)$	expected backlog $E(Q^n)$
Work Center					
1	1	5.01	2.02	5.01	0
2	1	.75	.32	.75	0
3	1	.69	.19	.69	0
4	1	.36	.17	.36	0
5	1	1.37	.39	1.37	0
6	1	1.65	.54	1.65	0
7	1	.14	.04	.14	0
8	1	.55	.17	.55	0
9	1	1.89	.61	1.89	0
10	1	2.19	.74	2.19	0

TABLE 3: CASE B

	planned lead time n	expected production E(P)	standard deviation σ(P)	expected queue E(Q)	expected backlog $E(Q^n)$
Work Center					
1	4	5.01	.80	20.04	.71
2	1	.75	.16	.75	0
3	1	.69	.15	.69	0
4	1	.36	.12	.36	0
5	1	1.37	.32	1.37	0
6	2	1.65	.24	3.31	.05
7	1	.14	.03	.14	0
8	1	.55	.13	.55	0
9	3	1.89	.29	5.68	.13
10	3	2.19	.31	6.58	.10

TABLE 4: CASE C

	planned lead time n	expected production E(P)	standard deviation $\sigma(P)$	expected queue E(Q)	expected backlog $E(Q^n)$
Work Center					
1	8	5.01	.55	40.07	1.05
2	1	.75	.13	.75	0
3	1	.69	.14	.69	0
4	1	.36	.11	.36	0
5	2	1.37	.20	2.74	.06
6	3	1.65	.18	4.97	.07
7	1	.14	.02	.14	0
8	1	.55	.11	.55	0
9	5	1.89	.22	9.45	.18
10	5	2.19	.23	10.96	.13

TABLE 5: CASE D

	planned lead time n	expected production E(P)	standard deviation σ(P)	expected queue E(Q)	expected backlog $E(Q^n)$
Work Center					
1	8	5.01	.55	40.07	1.05
2	1	.75	.13	.75	0
3	1	.69	.14	.69	0
4	1	.36	.11	.36	0
5	2	1.37	.20	2.74	.06
6	3	1.65	.18	4.97	.07
7	1	.14	.02	.14	0
8	2	.55	.08	1.10	.02
9	4	1.89	.22	7.56	.12
10	5	2.19	.23	10.96	.13

4. Discussion

In this section we review and discuss the key assumptions of the proposed model. We also indicate in what directions we may extend the basic model, as well as suggest issues in need of further investigation.

The most controversial assumption is likely to be the control rule given by (1). In particular, we assume that there are no constraints on setting the production levels in each time period. We suggest that one set the control parameters (i.e. planned lead times) so that the production requirements rarely exceed the normal range of production capacity. Yet we do not explicitly constrain the production requirements to do so, but argue that when the requirements exceed the normal capacity, we can still satisfy the requirements (at a cost) by redeploying resources. In some situations one might not be able to do this; in these instances, we might have a rigid constraint so that we need restate the control rule (1) as

$$P_{it} = \min\{\alpha_i Q_{it}, \overline{P}_{it}\}$$

where $\overline{P}_{it}$ is the production capacity at work center i in time period t. Although we have not tested this control rule in the context of a network of queues, Cruickshanks et al. (1984) have studied an analogous rule in a simpler setting consisting of one production stage. They find that the study of the unconstrainted control rule [i.e. (1)] provides a reasonable prediction of the behavior of the constrained control rule. We need to investigate, presumably by a simulation study, whether this observation holds in the more complex setting of a network of queues.

A second critical assumption is the Markov assumption made in (5). We assume that it is possible to model the work flows between work centers in a Markov fashion so that the history of a work flow is not necessary.

In general it is hard to imagine how this assumption might be overcome without resorting to a much more complex model structure. However, one might relax the assumption that all jobs are of the same type and are modelable by a single $\underline{\underline{\Phi}}$ matrix. If there are a few distinct types of jobs with different routings and production requirements, then one might identify a work flow matrix ($\underline{\underline{\Phi}}_k$) for each job type k so that its work flow could be modeled separately. Each work center would have a queue of work for each part type,and we would need a control rule that set the production level as a function of the multiple queues; for instance, we might restate (1) as

$$P_{ikt} = \alpha_{ik} Q_{ikt}$$

and

$$P_{it} = \sum_k P_{ikt}$$

where Q_{ikt} is the queue of work for jobs of type k at work center i, α_{ik} is the corresponding smoothing parameter, and P_{ikt} is production of jobs of type k at work center i. This extension would more faithfully model the work flows when it is possible to identify distinct types of jobs.

We have developed the model in the context of a job shop in which work "pushes" its way through the system. Each work center has a queue of work from which it sets its production level; the work center then pushes its queue of work to the queues of downstream work centers, as specified by the work flow matrix $\underline{\underline{\Phi}}$. In contrast to this we could conceive of a shop in which work "pulls" its way through the shop. After each work center is an inventory of work that has completed processing at that work center; production is triggered by demand on this inventory, which creates a backlog to be replenished. In other words, production at the work center acts to fill the backlog by replenishing the inventory. Furthermore, production at the work center will pull inputs from the inventories of upstream work centers, as specified by a work flow matrix. Such a pull system is a mirror image of the job shop (push system) that we have used to develop the model. We can apply the model directly to this system by equating the queue (push) to the backlog (pull), and by defining the $\underline{\underline{\Phi}}$ matrix to reflect how inputs are pulled into each work center.

Finally, the model may also be valuable for supporting the application of a Materials Requirements Planning (MRP) system [Orlicky (1975)] in a multi-stage or multi-plant production environment. The fundamental construct of an MRP system is the notion of a planned lead time. Associated with each production activity or stage is a lead time that forms the basis for production planning and material procurement. These lead times are the primary control mechanisms for deciding when to order raw materials, when to initiate part production and when to schedule subassemblies and final assemblies in order to satisfy a given set of demand requirements. Yet in the MRP literature I know of no theory on how to set these lead times. What one often hears is that the planned lead times should be set based on experience and observation; for instance, if we observe that the actual lead times for a production activity often exceed the planned lead time, then we need increase the planned lead time. But it is not at all clear how much to increase the planned lead time or even if this is the proper response. Indeed, one can argue that planned lead times beyond a point are just self-fulfilling prophecies; if I plan on an activity to take, say, ten weeks, then I will load the activity with work ten weeks before it is due and not surprisingly, it will take ten weeks (or more if something goes wrong) before the work passes through the activity. What one needs is a normative model that could help to assess the proper lead times for a given production system. It would seem that the model proposed in this paper could be directly extended to the MRP environment and would be of value in setting these planned lead times.

APPENDIX

We show here that the power series in (10) and in (11) will converge if and only if the spectral radius (maximal absolute eigen value) of $\underline{\underline{\Phi}}$ is less than one.

In order for the power series in (10) to converge we need

(A1) $$(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}})^s \longrightarrow \underline{\underline{0}}$$

as s goes to infinity where $\underline{\underline{0}}$ is the matrix of zeroes. But this is equivalent to requiring that the spectral radius of $(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}})$ be less than one (Noble 1969). We will show that this will be true if and only if the spectral radius of $\underline{\underline{\Phi}}$ is less than one. To do this we will use results from the Frobenius theory of positive matrices (e.g. Karlin and Taylor, 1975, pp. 542-551).

Let $\rho(\underline{\underline{A}})$ denote the spectral radius of matrix $\underline{\underline{A}}$. Assume that $\rho(\underline{\underline{\Phi}}) < 1$. Suppose that $\rho(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}}) \geq 1$ and let λ_0 and $\underline{x}_0$ be the maximal absolute eigen value and corresponding eigen vector for $\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}}$. That is

$$(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}})\,\underline{x}_0 = \lambda_0\,\underline{x}_0 \quad .$$

But this can be rewritten as

$$\underline{\underline{\Phi}}\,\underline{x}_0 = \underline{x}_0 + (\lambda_0 - 1)\,\underline{\underline{D}}^{-1}\,\underline{x}_0 \quad .$$

Thus if $\lambda_0 \geq 1$ we have that

$$\underline{\underline{\Phi}}\,\underline{x}_0 \geq \underline{x}_0$$

since $\underline{\underline{D}}^{-1}$ is a positive matrix. But this contradicts the assumption that $\rho(\underline{\underline{\Phi}}) < 1$. Hence, if $\rho(\underline{\underline{\Phi}}) < 1$ we must have

$$\rho(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}}) < 1 \quad .$$

Assume that $\rho(\underline{\underline{\Phi}}) \geq 1$ and let λ_0 and $\underline{x}_0$ be the maximal absolute eigen value and corresponding eigen vector for $\underline{\underline{\Phi}}$. Consider

$$(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}})\underline{x}_0 = (\underline{\underline{I}} - \underline{\underline{D}})\underline{x}_0 + \lambda_0\underline{\underline{D}}\,\underline{x}_0$$

$$= \underline{x}_0 + (\lambda_0 - 1)\underline{\underline{D}}\,\underline{x}_0 \quad .$$

Thus if $\lambda_0 \geq 1$ we have that

$$(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}})\underline{x}_0 \geq \underline{x}_0$$

since $\underline{\underline{D}}$ is a positive matrix. But this implies that $\rho(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}}) \geq 1$. Hence, if $\rho(\underline{\underline{\Phi}}) \geq 1$, then we must have that

$$\rho(\underline{\underline{I}} - \underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\Phi}}) \geq 1 \quad .$$

This completes the proof showing that (10) converges iff $\rho(\underline{\underline{\Phi}}) <$.

We now argue that (11) converges iff $\rho(\underline{\underline{\Phi}}) < 1$. First, if (A1) is not true, then it is easy to see that the series in (11) cannot converge. Second, if (A1) is true, then we can show that the series is absolutely convergent (and thus convergent) by using (10) and (A1) to bound the corresponding series of absolute values term by term.

REFERENCES

BERTRAND, J.W.M. AND J.C. WORTMANN, Production Control and Information Systems for Component Manufacturing Shops, Elsevier Science Publishing Co., Amsterdam (1981).

BERTRAND, J.W.M., "The Effect of Workload Control on Order Flow Times", in Operational Research '81, J. P. Brans (editor), North Holland Publishing Co., (1981).

CRUICKSHANKS, A.B., R.D. DRESCHER AND S.C. GRAVES, "A Study of Production Smoothing in a Job Shop Environment", Management Science, Vol. 30, No. 3, (March 1984), pp. 368-380.

CONWAY, R.W., W.L. MAXWELL AND L.W. MILLER, Theory of Scheduling, Addison-Wesley Publishing Co., Reading, MA, (1967).

GOMERSALL, E.R., "The Backlog Syndrome", Harvard Business Review, Vol. 92, No. 5 (1964), 105-115.

HOLSTEIN, W.K. AND W.L. BERRY, "Work Flow Structure: An Analysis for Planning and Control", Management Science, Vol. 16, No. 6 (February 1970), B-324-B-336.

HOLSTEIN, W.K. and W.L. BERRY, "The Labor Assignment Decision: An Application of Work Flow Structure Information", Management Science, Vol. 18, No. 7 (March 1972), 90-400.

JACKSON, J.R., "Networks of Waiting Lines", Operations Research, Vol. 5 (1957), 518-521.

JACKSON, J.R., "Jobshop-Like Queueing Systems", Management Science, Vol. 10 (1963), 131-142.

REFERENCES (Contd.)

JONES, C.H., "An Economic Evaluation of Job Shop Dispatching Rules", Management Science, Vol. 20 (1973), 293-307.

KARLIN, S. AND H.M. TAYLOR, A First Course in Stochastic Processes, second edition, Academic Press, New York, (1975).

KARMARKAR, U. S., 1983, "Lot Size, Manufacturing Lead Times and Utilization," Graduate School of Management Working Paper (No. QM8312), University of Rochester.

NOBLE, B., Applied Linear Algebra, Prentice Hall, Inc., Englewood Cliffs, New Jersey, (1969).

ORLICKY, J., Material Requirements Planning, McGraw-Hill, New York, (1975).

WIGHT, O., "Input/Output Control - A Real Handle on Lead Time", Production & Inventory Management, 3rd Qrt., (1970).

ZIPKIN, P. H., 1983, "Models for the Design and Control of Stochastic, Multi-Item Batch Production Systems," Graduate School Business Working Paper (No. 496A), Columbia University.