European Journal of Operational Research 40 (1989) 222–242
North-Holland

Theory and Methodology

Real-time scheduling of an automated manufacturing center *

Narayan RAMAN
University of Illinois, College of Commerce and Business Administration, Champaign, IL 61820, USA

Ram V. RACHAMADUGU and F. Brian TALBOT
University of Michigan, School of Business Administration, Ann Arbor, MI 48109-1234, USA

Abstract: This paper investigates the dynamic scheduling of an automated manufacturing workcenter at which jobs are processed in batches, and there is a constant changeover time between batches of different part types. The primary measures of schedule performance are mean flow time and mean tardiness.

The dynamic scheduling problem is treated as a series of static problems which are solved on a rolling-horizon basis. Characteristics of the optimal solutions to the mean flow time and mean tardiness problems are developed, and an implicit enumeration approach to the mean tardiness problem is proposed. These results are used for constructing efficient scheduling procedures for the dynamic problem. We also derive the steady state relationship between workcenter utilization level, batch size and mean flow time for one and two part types. A simulation study extends this relationship to a larger number of part types.

Keywords: Scheduling, tardiness, flexible manufacturing systems

1. Introduction

This study is part of an ongoing effort to develop a real-time scheduling system for the Automated Manufacturing Research Facility (AMRF) at the National Bureau of Standards, Gaithersburg, Maryland. This facility has been established to serve as a realistic test environment for standards metrology research. [The reader is referred to Jones and McLean (1986) and Simpson, Hocken and Albus (1982) for detailed descriptions of the AMRF.] To reach this goal, computerized planning and control systems are being developed to operate the facility; one component of which is real-time scheduling.

The AMRF configuration facilitates the decomposition of the problem of scheduling the entire system into single-machine problems for the individual workstations. Consequently, a two-step research approach has been followed. The first step is to examine each workstation individually. The results of this step are then modified in the second step within an integrated model of the system.

* This research was supported in part by grants from the National Bureau of Standards and the School of Business, University of Michigan.
Received February 1987; revised May 1988

This paper reports on the results of Step One for the Automatic Turning Center (ATS). The ATS manufactures cylindrical parts of several different part types to specific orders. The relatively small sizes of these parts makes it impractical to transport and machine them individually. Therefore, workpieces of the same part type are processed together in a *batch*. The batch size of any part type is determined by the size of the workpiece and the cassette in which the workpieces are transported and stored in the input buffer at the ATS. While changing over from one part type to another, a setup time is incurred at the ATS for changing collets (which hold the workpiece on the machine tool). However, this setup time is independent of the part types involved in the changeover. Orders arrive randomly at the ATS. Because of the batching requirements, a job of a given part type is schedulable only if the number of jobs of that part type is large enough to constitute one or more batches.

Because the ATS manufacturers parts to specific orders, it is appropriate to consider due date based objectives for the scheduling problem. The major criteria addressed in this paper are mean job flow time and mean job tardiness. Although mean job flow time is not a due date related scheduling measure per se, any production system's ability to quote (and maintain) tight due dates is improved by minimizing the average time spent by a job in the system; hence the inclusion of this objective. In addition, to examine the robustness of a given scheduling rule, we also consider the measures of proportion of tardy jobs and standard deviation of tardiness.

In order to facilitate the implementation of any scheduling procedure in real time, we treat the dynamic problem as a series of static problems. A static problem is generated whenever the ATS becomes available, and it involves making a decision on the batch to be processed next based only on those batches which are available in the system at that point in time. We note that if the system were making parts to inventory, it may be more appropriate to include those jobs for which orders have not yet been received, and, thereby, improve machine utilization (albeit, at the expense of inventory holding costs).

The literature on deterministic scheduling of jobs in conventional machine shops is quite rich. Excellent source materials on this subject include Conway, Maxwell and Miller (1967), Baker (1974), Coffman (1976), Rinnooy Kan (1976), Graves (1981), French (1982), Lawler, Lenstra and Rinnooy Kan (1982), Lenstra and Rinnooy Kan (1985), and Blazewicz, Cellary, Slowinski and Weglarz (1986). The static single machine scheduling problem, in particular, has attracted considerable research. Smith (1956) established the optimality of the shortest processing time (SPT) rule for the mean flow time criterion. Major research efforts on the mean tardiness problem include Srinivasan (1971), Shwimer (1972), Rinnooy Kan, Lageweg and Lenstra (1975), Picard and Queyranne (1978), and Potts and Van Wassenhove (1985).

The major assumption in the bulk of this research is that each job is processed in unit batch size. However, the recent works of Zipkin (1986), Karmarkar (1987a) and Karmarkar (1987b) among others show that mean job flow time is dependent on batch sizing decisions. In particular, Naddef and Santos (1984), Santos and Magazine (1985), and Dobson, Karmarkar and Rummel (1987) develop the flow time-batch size relationship for a static single machine problem.

This study differs from the previous work on batch scheduling in single-machine systems in at least one of three aspects. First, we address a dynamic system. However, given our approach of decomposing the dynamic problem into static problems, we have investigated the latter as well. The results obtained for the mean flow time criterion in the static case parallel those reported in independent investigations by Naddef and Santos (1984), Santos and Magazine (1985), and Dobson et al. (1987). Secondly, we address due data based scheduling measures. Finally, and perhaps most importantly, in this study, batch sizing decisions are design-level decisions and are, therefore, not a part of real-time scheduling decisions. Herein lies an important difference between conventional job shops and FMSs. While it is possible in a job shop to delay the batch sizing decision until the scheduling decision needs to be made, it is not technically feasible to do so in many FMSs because of the need to move, locate and store parts in cassettes of fixed sizes and the limited storage space at the machines. Hence, in our study, batch sizes are considered fixed for the purpose of real-time scheduling. However, in so far as the batch size affects the system performance, we do address the impact of batch size on flow time and tardiness. This is done in two ways. First, we develop steady-state relationships between batch size and mean flow time. Next, while discussing various scheduling procedures, we study their performance with respect to variations in batch size.

Table 1
Notation

p	=	Part type index, $p = 1, 2, \ldots, P$.
π_p	=	Processing time for a job of part type p.
N_p	=	Batch size of part type p.
j	=	Index of job within a batch (for a given part type), $j = 1, 2, \ldots, N_p, \forall_p$.
B_p	=	Number of currently schedulable batches of part type p.
b	=	Index of batch p (for a given part type), $b = 1, 2, \ldots, B_p, \forall p$. For clearer presentation, wherever necessary, b will be used with subscript p to indicate its part type.
T	=	The length of the scheduling horizon is $\sum_{p=1}^{P} (\pi_p N_p B_p) + S \sum_{p=1}^{P} B_p$.
t	=	Time period, $t = 1, 2, \ldots, T$.
S	=	Constant changeover (setup) time from one part type to another
δ_p	=	$\begin{cases} 1 & \text{if the collet currently on the machine is of part type } p, \\ 0 & \text{otherwise.} \end{cases}$
d_{pbj}	=	Due date of job j in batch p of part type p.
T_{pbt}	=	Tardiness of batch p of part type p if it is completed at time t, $\sum_{j=1}^{N_p} \max(0, t - d_{pbj})$.
X_{bpt}	=	$\begin{cases} 1 & \text{if batch } b \text{ of part type } p \text{ is completed at time } t, \\ 0 & \text{otherwise.} \end{cases}$
Y_{cq}^{bp}	=	$\begin{cases} 1 & \text{if batch } b \text{ of part type } p \text{ precedes batch } c \text{ of part type } q \text{ in the sequence,} \\ 0 & \text{otherwise.} \end{cases}$

Given our approach of decomposing the dynamic scheduling problem into a series of static problems, we first consider the static case in Section 2. The subsequent sections deal with the dynamic case. Section 3 discusses the impact of batch size on the steady-state performance of the system. Section 4 presents several scheduling procedures—three of which are based on the results obtained in Section 2, and others are modifications (to account for job batching) of procedures developed by other researchers. We also present the results of a simulation study of these procedures. We conclude in Section 5 with a summary evaluation of the analytical and simulation results. The notation used in this paper is given in Table 1.

2. Static scheduling

A static scheduling problem is generated whenever the ATS is available, and there are one or more schedulable job batches ahead of it. It is characterized by (i) formation of batches of fixed size for a given part type, and (ii) a constant setup time between one part type and another. Because jobs can be moved from the ATS only in batches, the completion times of all jobs in a batch equal the completion time of the last job in the batch.

At any given point in time, the ATS can be in any one of two states, (i) State A, in which there is no collet on the machine, if for instance, the machine was torn down just prior to the period in which the scheduling decisions need to be made, and (ii) State B, in which there is a collet for a specific part type on the machine. Because of the changeover time, the optimal decision would in general depend upon on the current state of the machine.

2.1. *Mean flow time*

The optimal sequence for minimizing mean flow time can be derived from the following result.

Theorem 1. *Given that the ATS is in State* A, *the mean flow time of all jobs is minimized by grouping the batches of the same part type together, and sequencing the part types in the non-decreasing order of their weighted batch processing times (WBPT),* $S/(B_p \cdot N_p) + \pi_p$.

Proof. Refer to Rachamadugu, Raman and Talbot (1986). Dobson, Karmarkar and Rummel (1987), and Santos and Magazine (1985) have shown this result to hold for a related problem.

If the ATS is in State B, the optimal sequence will either be the WBPT sequence itself, or the sequence in which all batches of the part type for which the collet is currently on the machine are scheduled first, and the relative positions of other batches are the same as in the WBPT sequence.

2.2. Mean tardiness

Raman (1988) shows the mean tardiness problem for the batch scheduling case to be NP-complete. While polynomial solution procedures are, therefore, unlikely for this problem, the following results establish some characteristics of an optimal sequence.

Theorem 2. *In a sequence optimal for the problem of minimizing mean tardiness, the batches of jobs for a given part type are sequenced in the non-decreasing order of the due dates of the jobs.*

Proof. Refer to Rachamadugu et al. (1986).

It follows from Theorem 2 that, while seeking the optimal solution, the waiting jobs of a given part type p can be ordered in the Earliest Due Date (EDD) sequence, and batches can be formed from this ordered list by grouping the first N_p jobs, then the next N_p jobs, and so on. The jobs remaining after the last batch has been formed need to be considered only in the next cycle, i.e., when the ATS next completes a batch of jobs. The implied precedence relationship among batches yielded by this procedure is depicted in Figure 1.

Henceforth, unless stated otherwise, it is assumed that jobs of the same part type are batched in the manner stated above.

Theorem 3. *Given that there is no more than one batch of each part type, any two adjacent batches,* x *of part type* p, *and* y *of part type* q, *satisfy the following condition in an optimal sequence if* x *preceeds* y:

$$\alpha_y c_x - \alpha_x c_y + \sum_{j \in \Delta_y} (K - d_{qyj}) - \sum_{i \in \Delta_x} (K - d_{pxi}) \leqslant 0,$$

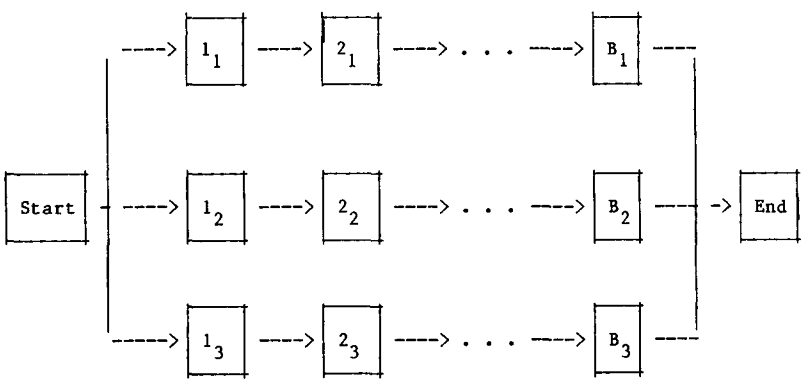

Figure 1. The precedence relationship among batches in an optimal sequence

where

$\alpha_x(\alpha_y)$ = *Number of jobs which are tardy in* $x(y)$ *when it precedes* $y(x)$.
$c_x(c_y)$ = $S+\pi_p(S+\pi_q)$.
K = *Completion time of the batch scheduled in the later position.*
$\Delta_x(\Delta_y)$ = *Set of jobs which are not tardy when* $x_p(y_q)$ *precedes* $y_q(x_p)$ *but which are tardy when the positions of these batches are interchanged.*

Proof. Refer to Rachamadugu et al. (1986).

Because of its restrictive assumptions, this theorem has limited applicability as an optimality criterion. However, it forms the basis for a strong heuristic which is discussed in Section 4.

The mean tardiness problem can be modeled as an integer program as shown below:

$$\text{Minimize} \quad \sum_{p=1}^{P} \sum_{b=1}^{B_p} \sum_{t=1}^{T} T_{bpt} X_{bpt}$$

$$\text{Subject to} \quad \sum_{t=1}^{T} tX_{cqt} - \sum_{t=1}^{T} tX_{bpt} + M\left(1 - Y_{cq}^{bp}\right) - N_q\pi_q - S' \geqslant 0,$$
$$1 \leqslant t \leqslant T, \quad p,\ q = 1,\ldots,P, \quad 1 \leqslant b \leqslant B_p, \quad 1 \leqslant c \leqslant B_q, \tag{1}$$

$$\sum_{t=1}^{T} tX_{bpt} - \sum_{t=1}^{T} tX_{cqt} + MY_{cq}^{bp} - N_p\pi_p - S' \geqslant 0,$$
$$1 \leqslant t \leqslant T, \quad p,\ q = 1,\ldots,P, \quad 1 \leqslant b \leqslant B_p, \quad 1 \leqslant c \leqslant B_q, \tag{2}$$

$$X_{bpt},\ X_{cqt},\ Y_{cq}^{bp} \in \{0, 1\} \quad \forall p,\ q,\ b,\ c,\ t, \tag{3}$$

where

$$S' = \begin{cases} S & \text{if } p \neq q, \\ 0 & \text{otherwise.} \end{cases}$$

Constraints (1) and (2) represent the disjunctive relationships between batches in the precedence network depicted in Figure 1. Constraint 3 specifies the binary nature of the variables.

In the above formulation, T denotes the length of the scheduling horizon. Clearly, $T \leqslant \sum_{p=1}^{P}\pi_p N_p B_p + S\sum_{p=1}^{P} B_p$. However, the following result provides a stronger upper bound on T, thereby reducing the number of variables in the suggested formulation.

Theorem 4. *If the schedulable batches belong to P different part types, the maximum number of setup changes* NS_P *is given by*

$$\mathrm{NS}_P = \min\left\{ \sum_{j=1}^{P} B_j,\ 2\sum_{j=1}^{P-1} B_j + 1 \right\},$$

where $B_0 = 0$ and the part types are numbered such that $B_1 \leqslant B_2 \leqslant B_3 \leqslant \cdots \leqslant B_P$.

Proof. Refer to Raman (1988).

T can be limited to $\sum_{p=1}^{P}\pi_p N_p B_p + S\,\mathrm{NS}_P$. [Note that $\mathrm{NS}_P \leqslant \sum_{p=1}^{P} B_p$.] In spite of this reduction in the problem size, this formulation is likely to yield a large number of constraints and variables even for moderately sized problems. The proposed implicit enumeration solution approach, however, obviates the need for generating or testing constraints (1) and (2) explicitly, and efficiently exploits the characteristics of the structure depicted in Figure 1. The solution procedure uses depth-first search and builds the

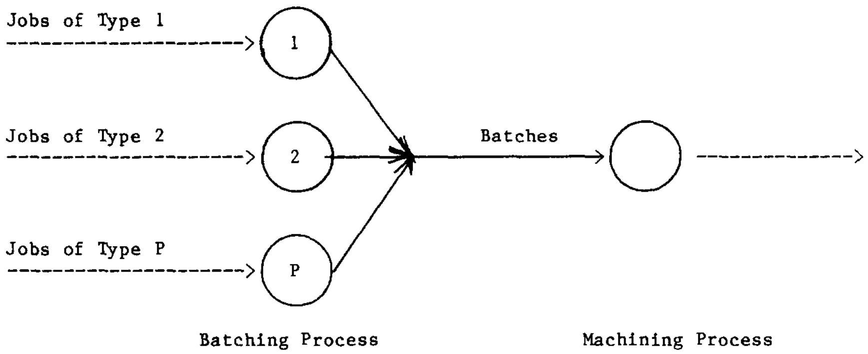

Figure 2. Representation of the dynamic system

schedule forwards in time. A node at level L in the enumeration tree corresponds to a partial sequence of L batches. Any given node n in the tree has an associated array A_n which contains the indexes of batches which are schedulable at the next level. The precedence relationships lead to a significant reduction in the computer storage requirements since the cardinality of A_n is limited to the total number of the part types of the unscheduled batches.

Starting from the unique node at level 0, the procedure implements the Modified Myopic rule (to be discussed in Section 4.1) at each node to determine the relative priorities of batches potentially schedulable at the next level. These batches are maintained in the non-increasing order of their priorities in array A_n. Augmentation at node n at level L requires selecting the batch with the highest priority (as determined a priori by the Modified Myopic rule) among those which have not already been branched from.

The lower bound at node n is given by

$$\text{LB}(n) = T(\sigma) + \max(0,\ L(\sigma')),$$

where $T(\sigma)$ is the tardiness already incurred by batches scheduled according to the partial sequence σ, and $L(\sigma')$ is the total lateness incurred by scheduling the remaining batches in the WBPT sequence. Since the WBPT sequence minimizes mean flow time, it minimizes mean lateness as well, and also since lateness is an underestimate of tardiness for any feasible sequence, LB(n) is a valid lower bound.

Since the breadth of the tree is determined by the cardinality of A_n associated with each node n at the preceding level, computational requirements increase exponentially with an increase in the number of schedulable part types. However, unless all jobs have substantial slack, the effectiveness of the lower bounding procedure, which is dependent upon the number of batches with late jobs, generally improves with an increase in the number of part types. In general, this procedure is more useful at the middle and the lower parts of the tree. At very low levels in the tree, however, the computational overhead associated with the repetitive calculation of LB(n) precludes its usage.

The results of the computational studies with this enumeration procedure for the dynamic problem are discussed in Section 4.

3. Steady state performance of the dynamic system

In order to better understand the dynamic problem, it is meaningful to first consider the impact of batching on the steady-state system performance. In this section, we develop relationships between batch size, ATS utilization and mean job flow time.

The system can be represented as a serial combination of the batching and machining processes as shown in Figure 2. Clearly, the flow time of a given job is the sum of its batching time BT, the waiting time of the batch BWT, and the batch processing time BPT (which includes possible setup time). The mean job flow time can then be expressed as the sum of the mean BT, mean BWT and mean BPT, or

$$\text{MFT} = \text{MBT} + \text{MBWT} + \text{MBPT}. \tag{4}$$

In order to facilitate our analysis, we make some simplifying assumptions. We assume that job arrival follows a Poisson process, and therefore, the interarrival times follow an exponential distribution. Furthermore, the mean interarrival time $1/\lambda$ is the same across all part types. All part types have the same batch size. Finally, we assume that jobs (and batches) are dispatched on a First-Come–First-Serve (FCFS) basis.

While these assumptions are made primarily to provide mathematical tractability, they help in providing useful insights which we exploit in Section 4 to explain some of the simulation results. Two of these assumptions, namely the equality of the mean interarrival times and the batch sizes across all part types can be relaxed. Doing so, however, leads to complicated expressions without adding much to our understanding.

3.1. Single part type

We begin our analysis by considering the case in which only one part type is produced. The steady-state mean batch processing time MBPT in this case is a constant and it equals πN since no setup changes are required.

Because job interarrival times are exponentially distributed with mean $1/\lambda$, the time between the formation of two successive batches follows an Erlang distribution with parameters N and λ, and it has a coefficient of variation $C_Z = 1/\sqrt{N}$. It can be easily seen that the mean batching time

$$\text{MBT} = \frac{(N-1)}{2\lambda} = \frac{(N-1)\pi}{2\rho}, \tag{5}$$

where $\rho = \lambda\pi$ is the ATS utilization.

The machining process can, therefore, be modeled as an $E_N/D/1$ system. Using the approximation formula suggested by Kraemer and Langenbach-Belz (1976), the mean batch waiting time can be expressed as

$$\text{MBWT} = \text{MBPT}\frac{\rho}{2(1-\rho)} C_Z^2 \exp\left\{\frac{-2(1-\rho)}{3\rho} \cdot \frac{\left(1-C_Z^2\right)^2}{C_Z^2}\right\}.$$

Substituting the values of MBPT and C_Z^2, we obtain

$$\text{MBWT} = \frac{\pi\rho}{2(1-\rho)} \exp\left\{\frac{-2(1-\rho)}{3\rho} \cdot \frac{(N-1)^2}{N}\right\}. \tag{6}$$

From (4), (5), and (6) we have

$$\text{MFT} = \frac{(N-1)\pi}{2\rho} + \frac{\pi\rho}{2(1-\rho)} \exp\left\{\frac{-2(1-\rho)}{3\rho} \cdot \frac{(N-1)^2}{N}\right\} + \pi N$$

or

$$\frac{\text{MFT}}{\pi} = \frac{(N-1)}{2\rho} + \frac{\rho}{2(1-\rho)} \exp\left\{\frac{-2(1-\rho)}{3\rho} \cdot \frac{(N-1)^2}{N}\right\} + N. \tag{7}$$

The plot of the dimensionless variable MFT/π as a function of N is depicted in Figure 3 for the uitlization levels of 50 percent, 70 percent and 90 percent. This graph shows MFT/π to be monotonically increasing in N. In addition, there are crossovers among the utilization levels with an increase in batch size N. This crossover is attributable to the fact that, as N increases, the mean batching time MBT (or equivalently, MBT/π) increases while the mean queueing time MBWT (or MBWT/π) decreases. Also, MFT/π is attributable primarily to MBT/π at low utilization levels, and to MBWT/π at high utilization

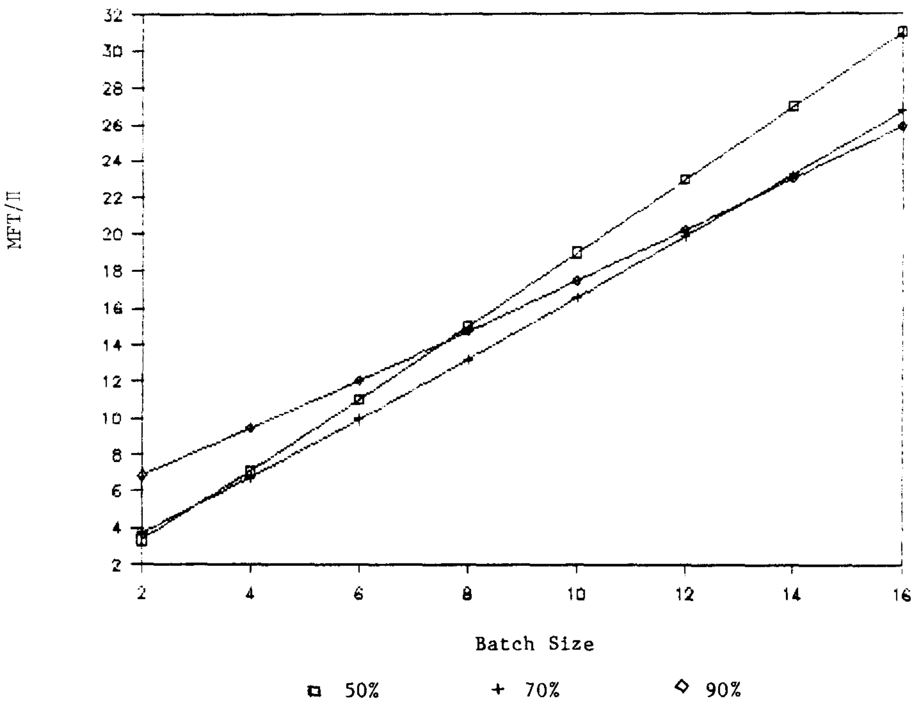

Figure 3. Graph of mean flow time/processing time vs. batch size for the single part type case

levels. Consequently, the lowest values of MFT/π are obtained at 50 percent utilization in the graph for small N, and for large N, 90 percent utilization yields the best results.

We will now extend the investigation to the case of two part types.

3.2. Two part types

Because the batching processes of individual part types are independent of one another, the time between two successive batches of (each part type) continues to follow an Erlang distribution with parameter N and $\lambda/2$. Hence, the expression for mean batching time given by Equation (5) is valid in this case as well, if we substitute λ by $\lambda/2$ and π by $\bar{\pi}$, where $\bar{\pi}$ is the average processing time (per job) of the two part types.

However, the introduction of the second part requires that we consider the setup changes as well. Specifically, the mean batch processing time in this case is given by

$$\text{MBPT} = N\bar{\pi} + \theta S = N\bar{\pi}\left(1 + \theta \frac{S}{N\bar{\pi}}\right),$$

where θ is the steady state probability that two successive batches are of different part types. However, if the setup time S is small compared to the batch machining times $N\pi_1$ and $N\pi_2$, as we would expect in most automated machining centers, we can approximate the mean batch processing time by

$$\text{MBPT} \simeq N\bar{\pi},$$

which is similar to the expression derived for a single part type.

The second part type introduces two other major considerations. First, the input to the machining process is not the output of an Erlang process; instead, it is the superposition of two independent and identical Erlang processes. Second, the batch processing time is not a constant anymore. It is a random variable with mean $N\bar{\pi}$ and a coefficient of variation, say C_{BPT} (≥ 0).

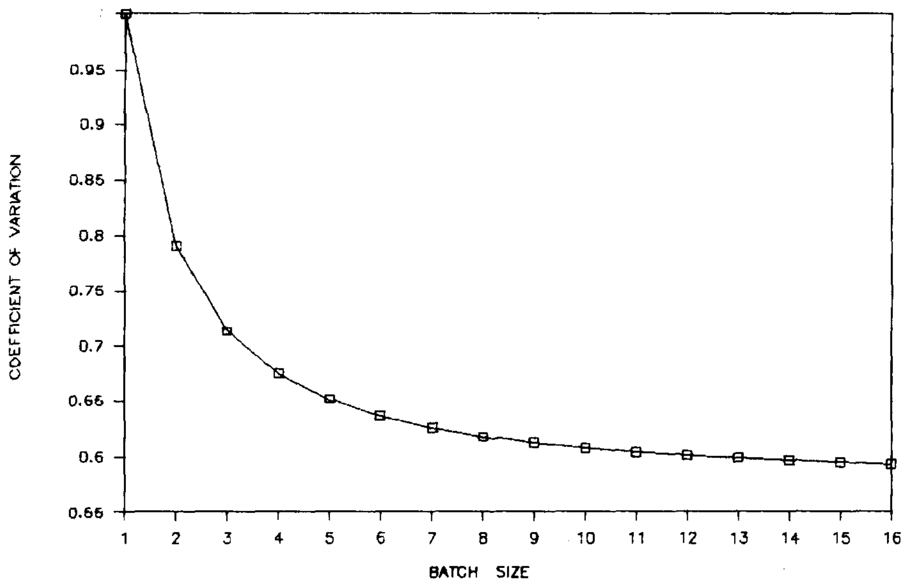

Figure 4. Graph of C_Z vs. batch size for two part types

Raman (1988) shows that the first two moments of the random variable Z, the time between the formation of two successive batches resulting from the superposition of two independent and identical Erlang processes, each with parameters N and $\lambda/2$, are given by

$$E(Z) = \frac{N}{\lambda}, \tag{8}$$

and

$$E(Z^2) = \frac{4}{\lambda^2} \sum_{r=0}^{N-1} \sum_{s=0}^{r} \binom{N+s}{s} (N+s+1) 2^{-(N+s+2)} + \frac{4}{N} \sum_{r=0}^{N-1} \sum_{s=0}^{N-1} \binom{r+s}{p} \frac{(r+s+1)(r+s+2)}{\lambda^2} 2^{-(r+s+3)}. \tag{9}$$

The coefficient of variation of Z, C_Z, is depicted in Figure 4 as a function of N. It can be seen that, similar to the case of a single part type, C_Z is decreasing in N.

On the other hand, it can easily be shown that C_{BPT} is independent of N as long as both part types have the same batch size. Also, $0 \leqslant C_{\text{BPT}} \leqslant 1$. Using Kraemer and Langenbach-Belz approximation formula for this GI/G/1 model, we obtain

$$\text{MBWT} = \frac{N\bar{\pi}\rho}{2(1-\rho)} \left(C_Z^2 + C_{\text{BPT}}^2 \right) \exp\left\{ \frac{-2(1-\rho)}{3\rho} \cdot \frac{\left(1 - C_Z^2\right)^2}{\left(C_Z^2 + C_{\text{BPT}}^2\right)} \right\}.$$

Substituting the values of MBT, MBWT and MBPT in Equation (4), we have

$$\text{MFT} = \frac{(N-1)\bar{\pi}}{\rho} + \frac{N\bar{\pi}\rho}{2(1-\rho)} \left(C_Z^2 + C_{\text{BPT}}^2 \right) \exp\left\{ \frac{-2(1-\rho)}{3\rho} \cdot \frac{\left(1 - C_Z^2\right)^2}{\left(C_Z^2 + C_{\text{BPT}}^2\right)} \right\} + N\bar{\pi}$$

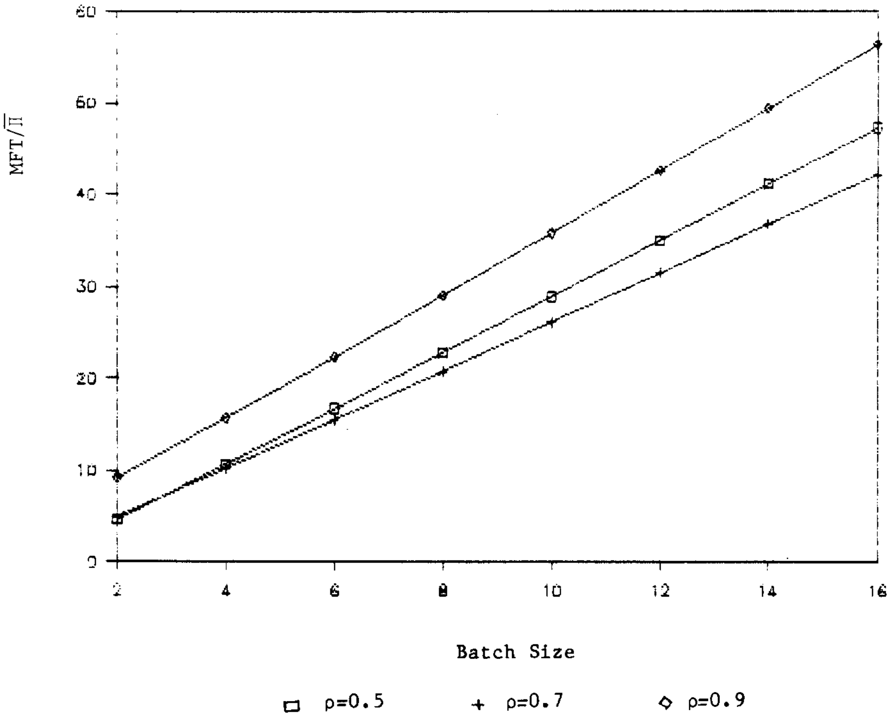

Figure 5. Graph of MFT/$\bar{\pi}$ vs. batch size for two part types, $C_{BPT} = 0$

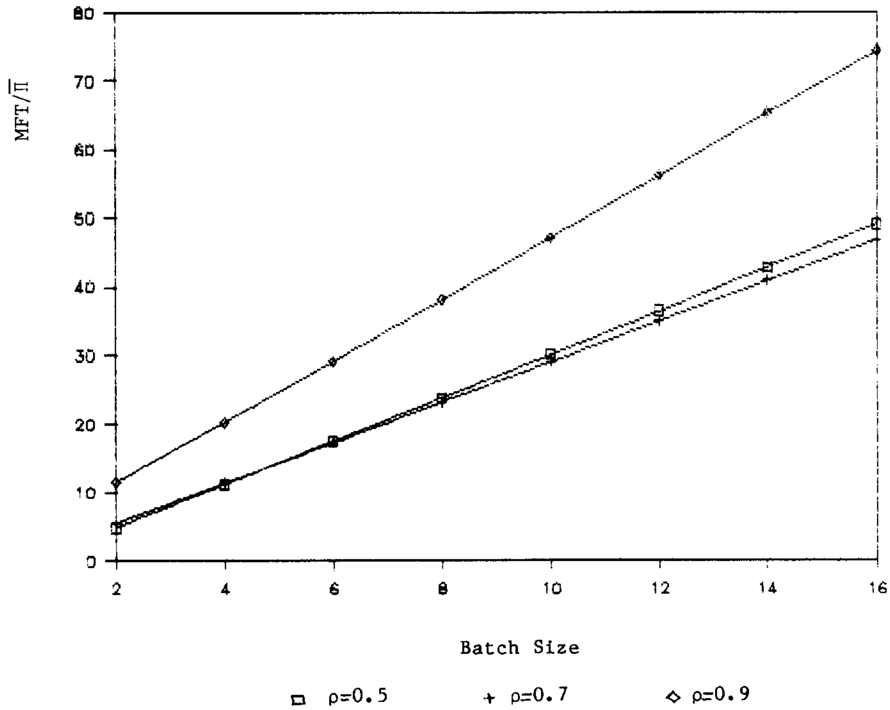

Figure 6. Graph of MFT/$\bar{\pi}$ vs. batch size for two part types, $C_{BPT} = 0.5$

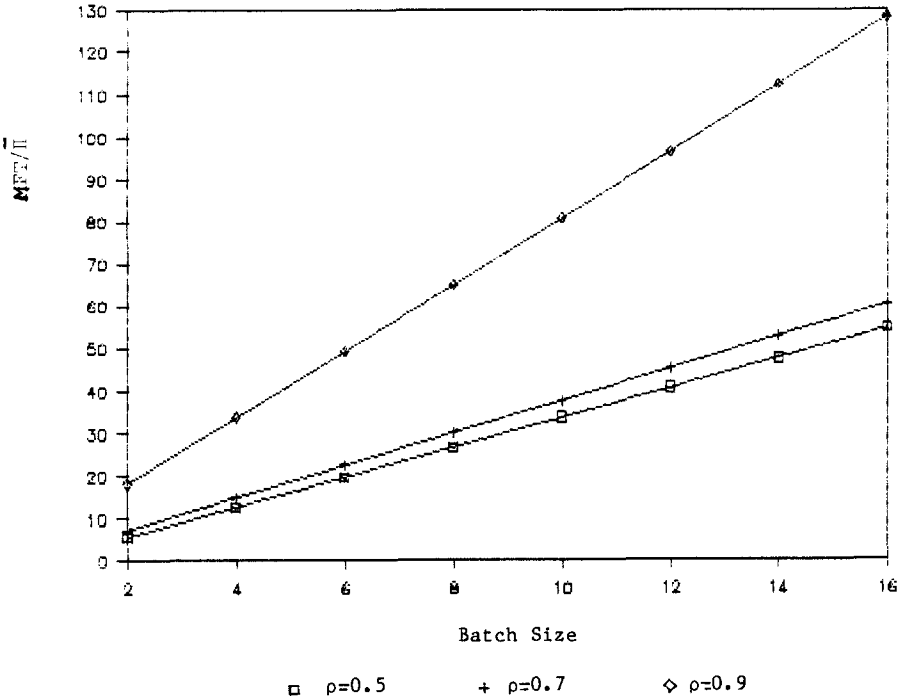

Figure 7. Graph of MFT/$\bar{\pi}$ vs. batch size for two part types, $C_{\mathrm{BPT}} = 1.0$

or

$$\frac{\mathrm{MFT}}{\bar{\pi}} = \frac{(N-1)}{\rho} + \frac{N\rho}{2(1-\rho)}\left(C_Z^2 + C_{\mathrm{BPT}}^2\right) \exp\left\{\frac{-2(1-\rho)}{3\rho} \cdot \frac{\left(1-C_Z^2\right)^2}{\left(C_Z^2 + C_{\mathrm{BPT}}^2\right)}\right\} + N. \tag{10}$$

In this case, the dimensionless variable MFT/$\bar{\pi}$ is a function of N, ρ and C_{BPT}^2. [Note that C_Z^2 is a function of N.]

Figures 5 through 7 depict the behavior of MFT/$\bar{\pi}$ with respect to N and ρ at three values of C_{BPT} — 0, 0.5 and 1. All three figures show MFT/$\bar{\pi}$ to be increasing in N. As depicted in Figure 5, when C_{BPT} equals zero, we obtain crossovers among the three utilization levels similar to the case of one part type. However, as shown in Figures 6 and 7, with an increase in C_{BPT}, the relative performance at 70 percent and 90 percent utilization levels deteriorates. In fact, when C_{BPT} equals 1, there is no crossover. This phenomenon is attributable to the fact that the mean queueing time MBWT/$\bar{\pi}$ is an increasing function in C_{BPT}. Therefore, when C_{BPT} is sufficiently high, MBWT/$\bar{\pi}$ does not decrease rapidly with an increase in N. Consequently, at high utilization levels, MFT/$\bar{\pi}$ remains high. Also note that the absolute value of MFT/$\bar{\pi}$ increases with an increase in C_{BPT} for all three utilization levels.

This analysis cannot be easily extended to three or more part types because it is difficult to derive closed form expressions for $E(Z)$ and $E(Z^2)$. However, in Section 4, we present the results of a simulation experiment in which we address the dependence of MFT on utilization levels and batch sizes.

4. Dynamic scheduling

The dynamic nature of the real-time scheduling problem requires implementation of scheduling procedures for selecting the batch to be processed next whenever the ATS becomes available. We describe the scheduling rules investigated in Section 4.1. Section 4.2 describes the simulation experiment conducted to evaluate the effectiveness of these rules. The experimental results are given in Section 4.3.

4.1. Description of the dispatching procedures

This paper investigates nine dispatching procedures. The first procedure is based on the first-come–first-serve discipline. The next four procedures—the revised modified due date rule, the modified myopic rule, the modified Montagne's rule and the revised earliest due date rule are modifications of heuristic methods found to be effective for the regular single-machine and/or job-shop problems by various researchers. The modifications have been necessitated by the need to incorporate the batching of jobs and changeover time. The sixth procedure is based on the Weighted Batch Processing condition discussed in Section 2.1. The seventh procedure is derived from the necessary condition for local optimality stated in Theorem 3, while the eighth procedure is a combination of the necessary condition and the Modified Myopic rule. The final procedure tested is based on the implicit enumeration approach described in Section 2.2. These procedures are now discussed.

(i) First-Come–First-Serve (FCFS) method

The First-Come–First-Serve Rule sequences the batches in the order in which they are formed. This rule essentially provides a benchmark for evaluating the relative effectiveness of other procedures investigated.

(ii) Revised Modified Due Date (RMDD) method

For the single-machine and job-shop tardiness problems, Baker and Bertrand (1982), Baker and Kanet (1983), and Baker (1984) found the modified due date (MDD) to be quite effective relative to other heuristics. The robustness of MDD lies in effectively combining the shortest processing time (SPT) rule, which has been shown to be quite effective when the due dates are set very tightly, and the EDD rule which performs well when the due dates are lax. Rachamadugu (1987) showed that, for the regular single-machine tardiness problem, there exists at least one optimal sequence which satisfies the MDD rule.

For the present study, the MDD rule has been modified as follows. The revised modified due date RMDD_{bp} of a batch b of part type p is defined as,

$$\text{RMDD}_{bp} = \sum_{j=1}^{N_p} \max\left(t + \delta_p S + \pi_p N_p,\ d_{pbj}\right),$$

where t is the time at which the scheduling decision has to be made.

According to the RMDD rule, when the ATS becomes available at time t, the batch with the earliest modified due date RMDD_{pb} is selected for processing. While it can be shown that the RMDD rule is not necessary for local optimality, it is nevertheless worth investigating, given the strength of the MDD rule for regular single-machine problems.

(iii) Modified Myopic (MYOP) method

Morton and Rachamadugu's (1982) study of the single-machine weighted tardiness problem found the Myopic rule to be quite effective compared to the other rules tested. Like MDD, the myopic rule reduces to SPT when all jobs have nonpositive slack, and to EDD when they have substantial slack. In the intermediate range, however, the priority assigned to a job increases exponentially with decreases in its slack. The modified myopic rule assigns priority MYOP_{bp} to a batch b of part type p, where

$$\text{MYOP}_{bp} = \sum_{j=1}^{N_p} \text{Pr}_{bpj}$$

and

$$\text{Pr}_{bpj} = \frac{1}{\left(\delta_p S + \pi_p N_p\right)} \exp\left[-\left(d_{pbj} - \left(t + \delta_p S + \pi_p N_p\right)\right)^{+} / \pi_{\text{ave}}\right],$$

where π_{ave} is the mean processing time of a batch of jobs, averaged over all part types, and $(x)^{+}$ denotes $\max(x, 0)$.

(iv) Modified Montagne's (MONT) method

In their experimental investigation of the regular single-machine tardiness problem, Baker and Martin (1974) found the dispatching procedure suggested by Montagne (1969) to perform well. In our study, this procedure has been revised to yield the Modified Montagne's rule which sequences the available batches in the non-decreasing order of MONT_{bp}, where for batch b of part type p,

$$\text{MONT}_{bp} = \sum_{j=1}^{N_p} \left[\pi_p N_p \Big/ \left(\sum_{p=1}^{B_p} \left(t + \pi_p N_p + S \right) - d_{pbj} \right) \right].$$

The implementation of RMDD, MYOP and MONT rules, automatically incorporates the precedence relationship shown in Figure 1. Computational requirements are thereby reduced by restricting the evaluation of priorities, for each part type, to only the batch with the lowest sum of job due dates.

(v) Revised Earliest Due Date (REDD) method

The REDD method sequences the batches in the non-decreasing order of the sum of the due dates of jobs in the batch. Operationally, this rule is equivalent to the earliest due date rule for the regular single-machine problem.

(vi) Weighted Batch Processing Time (WBPT) method

The WBPT method sequences the batches in the non-decreasing order of their weighted batch processing times $S/N_pB_p + \Pi_p$. Ties between batches are broken by sequencing the batch with the lower sum of job due dates first.

(vii) Necessary Condition-based (NC) method

This method implements the condition stated in Theorem 3 as a heuristic procedure. Though this theorem is restrictive in its scope, it is obvious from the discussions presented in Section 2.2 that the application of this theorem to the more general case, in which there is more than one batch of each part type, should lead to minor degradation of the solution quality, as long as the changeover time S is small relative to the batch processing times $\pi_p N_p$.

The NC method considers the schedulable batches two at a time and establishes the precedence relationship between them. As discussed in Section 2.2, Theorem 3 may not always be able to establish the precedence relationship between two batches. In such cases, the NC method determines the precedence in the favor of the batch with the lower sum of job due dates. In general, since this precedence relationship need not be transitive, cycles may be formed while applying this procedure. In this study, however, the formation of such cycles is avoided by enforcing transitivity in the order in which the batches are considered.

(viii) Revised Necessary Condition-based (RNC) method

This procedure is similar to NC, the underlying difference being the use of the Modified Myopic rule, instead of the sum of the job due dates, to break ties between two batches when Theorem 4 fails to establish precedence.

The precedence relationships imply that, while implementing NC and RNC, the evaluation of the condition stated in Theorem 3 can be restricted to the batches due the earliest for each part type.

(ix) Implicit enumeration (BB) method

This procedure implements the branch and bound method discussed in Section 2.2 to solve the static tardiness problem optimally for selecting the batch to be processed next. Unlike the procedures mentioned earlier, this method provides an ordered sequence of *all* batches. Clearly, if no batch has been formed since the previous implementation of this procedure (in the previous cycle), the same sequence can be retained for this cycle as well. Operationally, this reduces the number of times the static optimal solution must be found.

4.2. The simulation model

A brief description of the simulation model used in this study is given below. For further details, the reader is referred to Rachamadugu et al. (1986).

Orders arrive at the ATS following a Poisson distribution. The number of jobs in an order has a uniform distribution. Each job is equally likely to belong to any one of the 10 part types that the ATS currently manufactures. Upon its arrival, each job is assigned a due date based on a modification of the Total Work Content rule. The due date d_{pj} of job j of part type p is given by

$$d_{pj} = a_j + F\pi_p N_p + \beta_p, \tag{11}$$

where a_j is the arrival time of job j, F is the flow allowance factor and β_p is the average batching time for a job of part type p. For the simulation sums, a range of due date tightness was achieved by using flow allowance factors of 2, 4, 8, 12, 16 and 20. β_p was determined by the following relationship,

$$\beta_p = \frac{\Lambda(N_p - 1)}{2\theta_p \theta_{\text{ave}}}, \tag{12}$$

where Λ is the mean interarrival time of orders, θ_p is the probability that an arriving job is of part type p, and θ_{ave} is the average order size. In the simulation study, the same batch sizes were used for all part types. Four values of batch sizes—2, 4, 8 and 16, and three levels of ATS utilization—50 percent, 70 percent, and 90 percent were considered. Hence, a total of 72 simulation runs were conducted. The processing times of the part types yielded a C_{BPT} of 0.5.

4.3. Experimental results

The results of the simulation experiment are given below for each performance criterion. [1]

Mean Flow Time (MFT)

Figure 8 depicts the plot of MFT against batch size for FCFS procedure at the three levels of utilization investigated. Note that at 90 percent utilization, there is a reduction in MFT as batch size is increased from 2 to 4. This reduction does not, however, occur for any procedure other than FCFS and REDD. The plot of MFT against batch size for other scheduling rules typically follows the pattern shown in Figure 9 for the BB method.

Among the scheduling procedures, WBPT is dominant at all utilization levels and batch sizes. MONT, BB and MYOP yield comparable results; BB and MYOP are more effective when due dates are set tightly.

Mean Tardiness (MT)

Figure 10 presents a typical graph of MT against batch size for FCFS. It can be seen that while MT increases monotonically with increases in batch size for 50 percent and 70 percent utilizations, at 90 percent utilization, we obtain a decrease in MT with increase in batch size initially.

Figure 11 shows the plot of MT against batch size for various scheduling rules at 90 percent utilization and flow allowance of 4. BB yields the best overall results though RNC, NC and MYOP are comparable. The relative performance of these rules remains unchanged at other values of flow allowances.

Proportion of jobs Tardy (PT)

The graph of PT against batch size is shown in Figure 12. Among the scheduling procedures, WBPT dominates all others across all scenarios. MONT yields comparable results while BB and MYOP perform

[1] For greater clarity in presentation, the results of those dispatching procedures which have been consistently dominated by others are omitted.

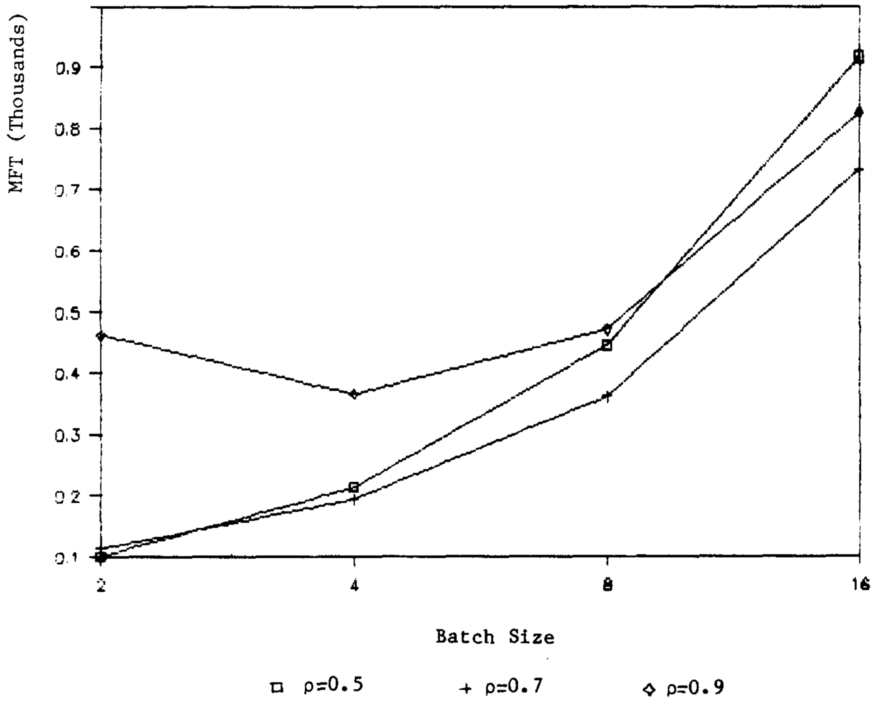

Figure 8. Graph of MFT vs. batch size, FCFS rule

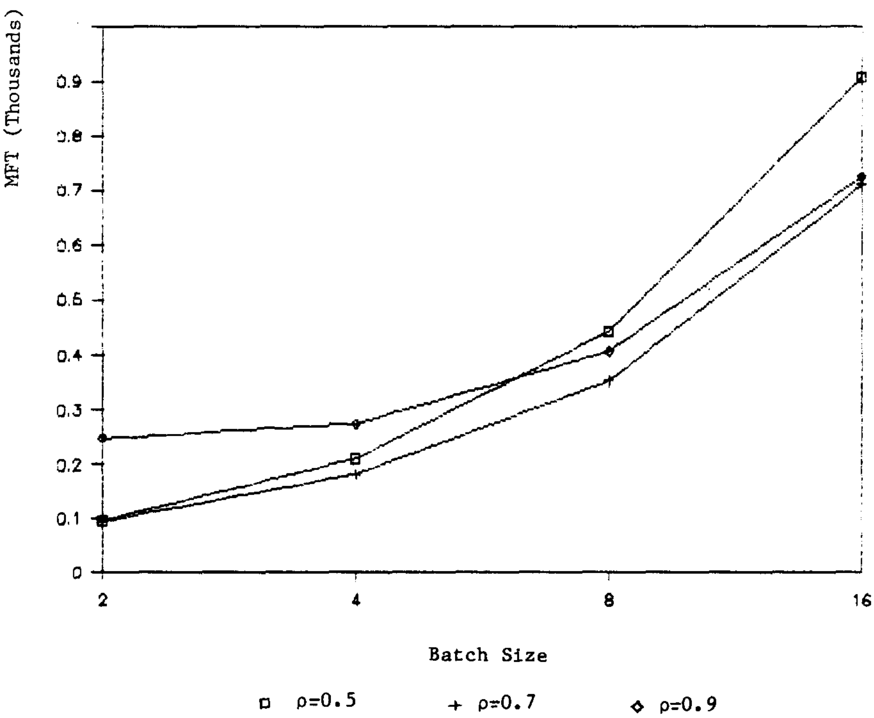

Figure 9. Graph of MFT vs. batch size, BB rule

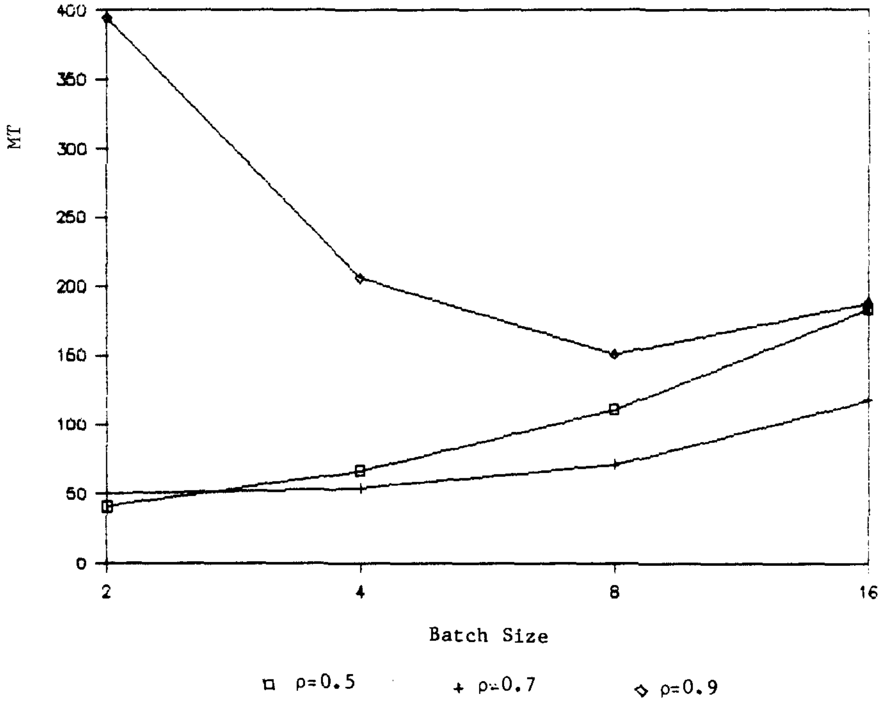

Figure 10. Graph of MT vs. batch size, FCFS rule

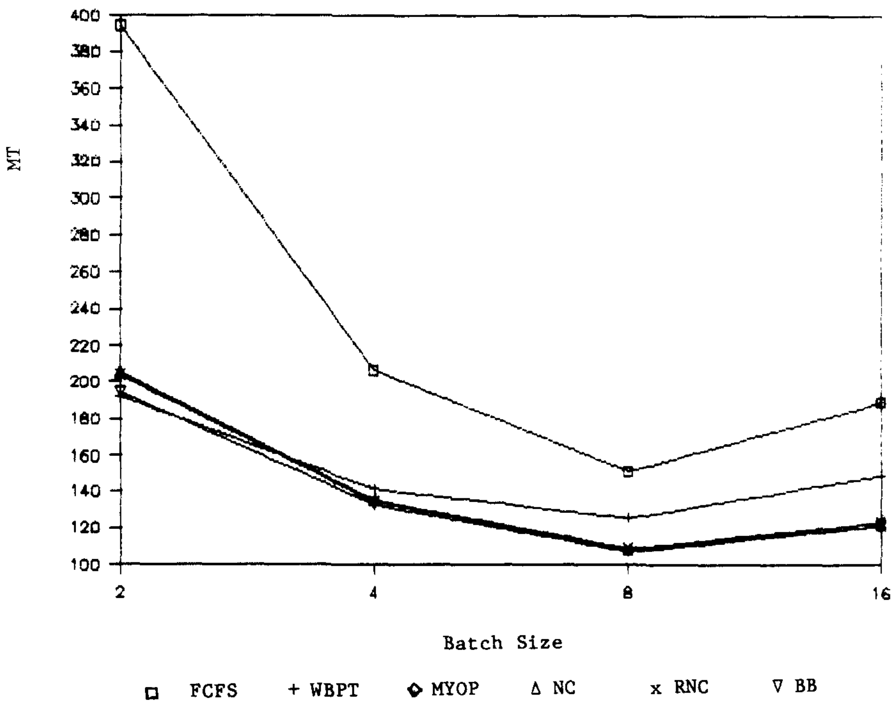

Figure 11. Graph of MT vs. batch size, $F = 4$, 90% utilization

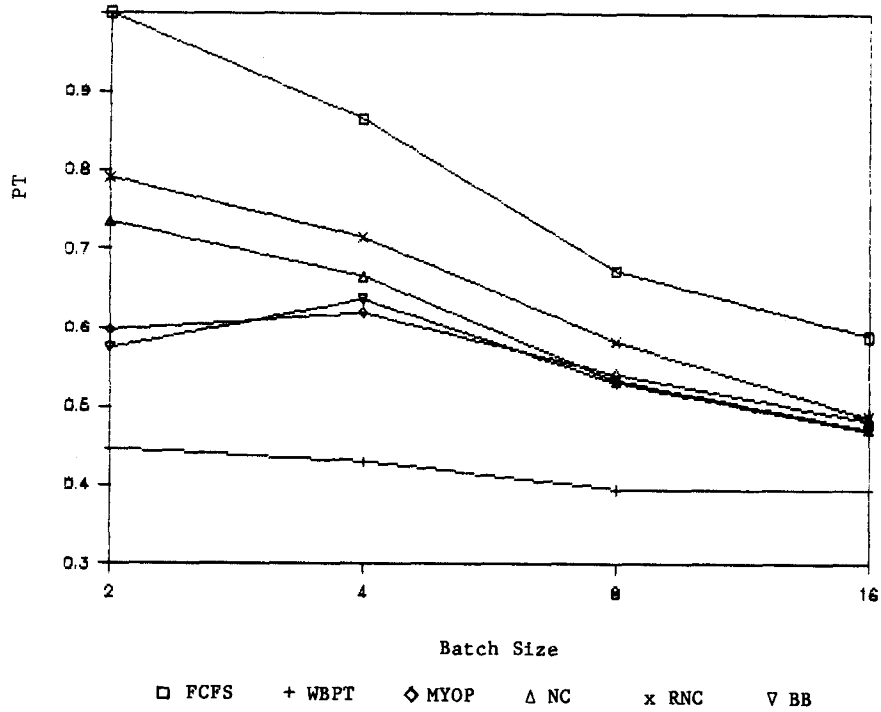

Figure 12. Graph of PT vs. batch size, $F = 4$, 90% utilization

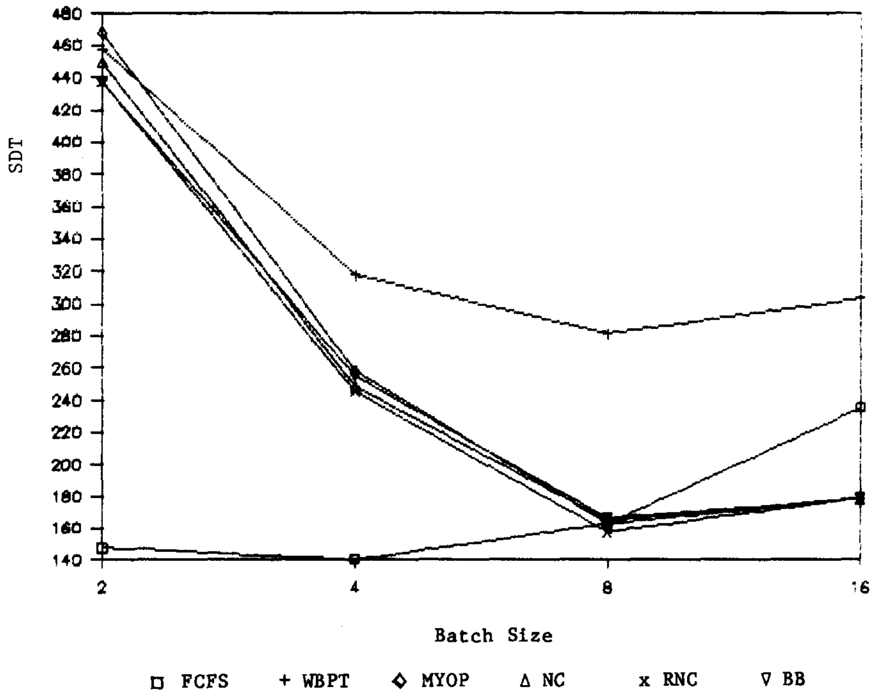

Figure 13. Graph of SDT vs. batch size, $F = 4$, 90% utilization

moderately well. In general, PT increases with an increase in batch size while it decreases at high utilization.

Standard Deviation of Tardiness (SDT)

The plot of SDT against batch size is shown in Figure 13. For small batches, FCFS and REDD yield the best results. For larger batches, RNC, RMDD, NC and MYOP perform reasonably well. In general, SDT increases with an increase in batch size at 50% and 70% utilizations, while at 90% utilization, there is an initial reduction followed by subsequent increase.

5. Discussion and summary

5.1. *Impact of utilization level and batch size*

As noted in Section 3, the flow time of any job comprises its batching time, and the queueing and processing time of the batch to which it belongs. At low utilization levels and for large batch sizes, the job flow time is largely attributable to its batching time. On the other hand, at high utilization levels and for small batch sizes, the queueing time is the major component of job flow time. Among the three utilization levels studied, the best compromise between batching and queueing times was obtained at 70 percent utilization level as shown analytically for the case of one and two part types as well as the results of the simulation study for 10 part types. In general, however, mean job flow time is also dependent upon the coefficient of variation of the processing times C_{BPT}, the scheduling rule used, and the number of part types produced P.

Since a job is schedulable only after it has been batched, scheduling decisions impact the outcome of only the machining process. Consequently, the difference between the various scheduling rules is especially significant when the queueing times are relatively high because of small batch sizes and/or high utilization levels.

A comparison of Figure 10, which depicts the variation of mean flow time with respect to batch size for 10 part types based on the simulation experiment, with Figure 6 which shows the relationship between these two variables for two part types at the same value of C_{BPT} indicates that the behavior of mean flow time also depends upon P, the number of different part types produced.

Mean tardiness values are likely to follow mean flow time values closely at high utilization levels, and for small batch sizes and tight due dates, because in these cases most jobs are likely to be tardy. In general, however, MT also depends upon the difference between the realized and expected batching times. This difference is significant at low utilization levels. As utilization levels decrease, the mean and, therefore, the variance of job interarrival times (because they follow an exponential distribution) increase which leads to greater variability in the due dates of jobs within a batch. In turn, the difference between the actual and expected batching times is high. Consequently, some jobs (which would be the initial jobs in their respective batches) tend to become tardy. This partially explains why, especially for large batch sizes, MT values at 50 percent utilization are comparable to those at 90 percent utilization even though the PT values are significantly smaller and SDT values are much higher.

5.2. *Evaluation of the dispatching procedures*

Under the TWK due date setting rule, BB and MYOP perform uniformly well for all four performance measures across all utilization levels and flow allowances. Unlike some of the procedures which are optimum for the static problem but which are not effective in a dynamic environment [for example, see Chand (1982) for a discussion on the degradation in the performance of the Wagner–Whitin algorithm when it is applied in a rolling-horizon environment], the effectiveness of BB for the static problem is extended to the dynamic case as well. The effectiveness of MYOP, which is essentially a single-pass heuristic, is however, surprising. Recall that the BB procedure implements MYOP for determining the initial solution. An analysis of the solutions generated by BB for the static problem has indicated that the initial solution provided by MYOP is quite often optimal; whenever it has been improved upon, the difference in the solution values have been marginal, generally in the order of 1 percent or less. NC and RNC appear to be quite effective for due date based performance measures. However, they yield high values of mean flow time at low utilization levels and/or high flow allowances; their performance for this measure is also very sensitive to the actual flow allowance factor used for setting job due dates. As expected, WBPT exhibits several 'SPT-like' properties such as low mean flow time, low proportion of tardy jobs and high tardiness variance.

Note that the TWK rule presupposes that the due dates are set deterministically when the jobs arrive and the same flow allowance is used for *all* jobs. To ascertain the robustness of these dispatching procedures with respect to variability in the values of F, another set of simulation runs was conducted. In this set, F was assumed to be uniformly distributed between 2 and 20; the due date was determined from Equation (10) using the sampled value of F. Under this due-date setting procedure, the correlation between the arrival time of a job and its due date is greatly reduced. The results of these runs indicate that the BB rule remains superior, while MYOP, NC and RNC continue to yield comparable results.

Finally, a comment will be made about the relative computational attractiveness of the various rules. All the rules except BB are essentially list processing algorithms. BB is exponential and hence is used operationally with a computational or iteration time trap. If the optimal solution is not found within this trap, the best solution found is used. Given the large size of the experiment used in this research, the simulation was carried out on an AMDAHL V8 computer using SIMAN and FORTRAN programs. The time trap for BB was 5 seconds. The simulation programs incorporating BB, MYOP and RNC were subsequently downloaded on a standard IBM-XT microcomputer with 640 kilobytes of core memory, and the execution times of these three procedures for solving static problems (which were generated during the simulation run) were monitored. The number of jobs in these static problems varied between 3 and 75. For batch size of 8, the average time taken to solve a static problem was found to be less than 3 seconds for all three procedures. On the IBM-XT, while MYOP and RNC were able to schedule consistently within 4 seconds, the maximum execution time for BB (with the built-in iteration trap for curtailing enumeration) was 10 seconds.

5.3. Summary

This paper extends the previous research on single-machine scheduling to the case where the jobs need to be processed in batches with a constant sequence-dependent changeover time. For the static mean tardiness problem, a solution procedure based on implicit enumeration is proposed. For the dynamic scheduling problem involving implementation of scheduling decisions in real-time, nine dispatching procedures are presented and evaluated.

This research highlights two issues of critical importance for an automated manufacturing center or a traditional job shop where job batching and sequence dependent setup times occur. First is the existence of an optimum utilization level for a given batch size. The results of this study indicate that for a given set of job and batching characteristics and for a given flow allowance factor, the mean flow time of all jobs is minimized at a utilization level which affects the best compromise between the batching and the queueing times of individual jobs. Operating at this utilization level highlights the need for appropriate material handling devices. Such a utilization level is also optimal for due date based scheduling measures when the flow allowances are small. In this study, for the given data set, the batch sizes considered and the utilization levels investigated, operating at 70 percent utilization optimized mean flow time as well as the three due date based scheduling measures for low to medium flow allowances across all dispatching rules. The results of Section 3 also indicate that the mean job flow time is increasing in the coefficient of variation of batch processing times C_{BPT}. Selecting appropriate batch sizes of individual part types is of critical importance because, for given part processing times, C_{BPT} depends only on the relative values of these batch sizes.

Secondly, this study suggests that, for such a manufacturing environment, serious consideration should be given to the dispatching rules BB, MYOP, NC and RNC, for their high expected system performance, and the ability to use them on today's microcomputers for real-time scheduling. [2]

References

Baker, K.R. (1974), *Introduction to Sequencing and Scheduling*, Wiley, New York.

Baker, K.R., and Martin, J.B. (1974), "An experimental investigation of solution algorithms for the single-machine tardiness problem", *Naval Research Logistics Quarterly* 21, 187–199.

Baker, K.R., and Schrage, L.E. (1978), "Finding an optimal sequence by dynamic programming: An extension to precedence-related tasks", *Operations Research* 26, 111–120.

Baker, K.R., and Bertrand, J.M.W. (1982), "A dynamic priority rule for sequencing against due-dates", *Journal of Operations Management* 3 (1), 37–42.

Baker, K.R. and Kanet, J.J. (1983), "Job shop scheduling with modified due-dates", *Journal of Operations Management* 4 (1), 11–22.

Baker, K.R. (1984), "Sequencing rules and due-date assignments in a job shop", *Management Science* 30, 1093–1104.

Blazewicz, J., Cellary, W., Slowinski, R., and Weglarz, J. (1986), *Scheduling under Resource Constraints: Deterministic Models*, J.C. Baltzer, Basel.

Chand, S. (1982), "A note on dynamic lot sizing in a rolling-horizon environment", *Decision Sciences* 13 (1), 113–119.

Coffman, E.G. (ed.) (1975), *Computer and Job-Shop Scheduling Theory*, Wiley, New York.

Conway, R.W., Maxwell, W.L., and Miller, L.W. (1967), *Theory of Scheduling*, Addison-Wesley, Reading, MA.

Dobson, G., Karmarkar, U.S., and Rummel, J.L. (1987), "Batching to minimize flow times on one machine", *Management Science* 33, 784–799.

French, S. (1982), *Sequencing and Scheduling — An Introduction to the Mathematics of the Job-Shop*, Ellis Harwood, Chichester, UK.

Graves, S. (1981), "A review of production scheduling", *Operations Research* 29, 646–675.

Groover, M.P. (1980), *Automation, Production Systems and Computeraided Manufacturing*, Prentice-Hall, Englewood Cliffs, NJ.

Jones, A. and McLean, C.J. (1986), "A Proposed hierarchical control model for automated manufacturing systems", *Journal of Manufacturing Systems* 5 (1), 15–25.

Karmarkar, U.S. (1987a), "Lot-sizes, lead times and in-process inventories", *Management Science* 33, 409–418.

Karmarker, U.S. (1987b), "Lot-sizing and sequencing delays", *Management Science* 33, 419–423.

Karmarkar, U.S., Kekre, S., and Kekre, S. (1984), "Lot-sizing in multi-item, multi-machine job shops", Working Paper no. QM8402, University of Rochester, Rochester, NY.

Kraemer, W., and Langenbalch-Belz, M. (1976), "Approximate formulae for the delay in the queueing system GI/G/I", Congressbook, 8th International Teletraffic Congress, Melbourne, Australia.

Lawler, E.L., Lenstra, J.K., and Rinnooy Kan, A.H.G. (1982), "Recent developments in deterministic sequencing and scheduling: A Survey", in: *Deterministic and Stochastic Scheduling*, edited by M.A.H. Dempster, J.K. Lenstra and A.H.G. Rinnooy Kan, Reidel, Dordrecht, 35–73.

Lenstra, J.K., and Rinnooy Kan, A.H.G. (1985), "Scheduling theory since 1981: An annotated bibliography", in: *Combinatorial Optimization: Annotated bibliographies*, edited by M. O'Eigeartaigh, J.K. Lenstra and A.H.G. Rinnooy Kan, Wiley, Chichester.

Montagne, E.R. (1969), "Sequencing with time delay costs", *Industrial Engineering Research Bulletin*, Arizona State University.

Morton, T.E., and Rachamadugu, R.V. (1982), "Myopic heuristics for the single machine weighted tardiness problem", Technical Report no. CMU-RJ-TR-83-9, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Naddef, D., and Santos, C. (1984), "Batching algorithms for single operation problems", Working Paper, Department of Combinatorics and Optimization, University of Waterloo, Ontario, Canada.

Picard, J.C., and Queyranne, M. (1978), "The time-dependent traveling salesman problem and its application to the tardiness problem in one-machine Scheduling", *Operations Research* 26(1), 86–110.

Potts, C.N., and Van Wassenhove, L.N. (1985), "A branch and bound algorithm for the total weighted tardiness problem", *Operations Research* 33(2), 363–377.

Rachamadugu, R.V. (1987), "A note on weighted tardiness problem", *Operations Research* 35(3), 450–453.

Rachamadugu, R.V., Raman, N., and Talbot, F.B. (1986), "Due-date based scheduling in a flexible manufacturing system (The ATS)", Working Paper, National Bureau of Standards, Gaithersburg, MD.

Raman, N. (1988), "Real-time scheduling problems in a general flexible manufacturing system", PhD dissertation, University of Michigan, Ann Arbor, MI.

Rinnooy Kan, A.H.G. (1976), *Machine Scheduling Problems: Classification, Complexity and Computations*, Martinus Nijhoff, The Hague.

[2] The authors thank Drs. Richard Jackson and Albert Jones of the National Bureau of Standards, Gaithersburg, Maryland, for their valuable assistance in this research effort. Comments and suggestions of two anonymous referees have contributed significantly to improving the quality of this paper.

Rinnooy Kan, A.H.G., Lageweg, B.J., and Lenstra, J.K. (1975), "Minimizing total costs in one-machine scheduling", *Operations Research* 23, 908–927.

Santos, C., and Magazine, M.J. (1985), "Batching in single operation manufacturing systems", Working Paper, University of Waterloo, Waterloo, Canada.

Shwimer, J. (1972), "On the n-job, one-machine, sequence independent scheduling problem with tardiness penalties: A branch-bound solution", *Management Science* 18, B301–B313.

Simpson, J.A., Hocken, R.J., and Albus, J.S. (1982), "The automated manufacturing research facility of the National Bureau of Standards", *Journal of Manufacturing Systems* 1(1), 17–32.

Smith, W.E. (1956), "Various optimizers for single-stage production", *Naval Research Logistics Quarterly* 3, 59–66.

Srinivasan, V. (1971), "A hybrid algorithm for the one machine sequencing problem to minimize total tardiness", *Naval Research Logistics Quarterly* 18, 317–327.

Zipkin, P.H. (1986), "Models for design and control of stochastic multi-item batch production systems", *Operations Research* 34, 91–104.