The International Journal of Flexible Manufacturing Systems, 6 (1994): 235–253
© 1994 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

An Approach to Real-Time Flexible Scheduling

GEORGE CHRYSSOLOURIS, KRISTIAN DICKE AND MOSHIN LEE
Room 35-134, Massachusetts Institute of Technology, Laboratory for Manufacturing and Productivity, Cambridge, MA 02139

Abstract. Two types of flexibility are important in manufacturing scheduling in general and in real-time scheduling in particular. The first is flexibility with respect to the criteria that can be considered in the scheduling decisions. The second is flexibility with respect to the trade-off between decision quality and computational burden: that is, the ability to arrive at a solution that makes maximum use of the *available* computational capacity and computation time. This paper describes a procedure which meets the above requirements. The procedure is justified using a theoretical analysis based on probability. Experimental results of the procedure's performance are also presented. The results show that random selection (which is used in the procedure) can play a useful role in the real-time scheduling problem.

Key Words: decision, manufacturing, real-time, scheduling.

1. Introduction

In manufacturing, there are large numbers of decision-making problems in the areas of product design, control of manufacturing processes, and production control. Decision making in manufacturing is characterized by the time required for a decision to be made, the number of decisions that have to be made over a certain period of time, and the impact a decision will have on the manufacturing system. Based on the values of these attributes, manufacturing decisions can be classified as strategic, operational or detailed decisions, which are related in a hierarchical fashion (see figure 1).

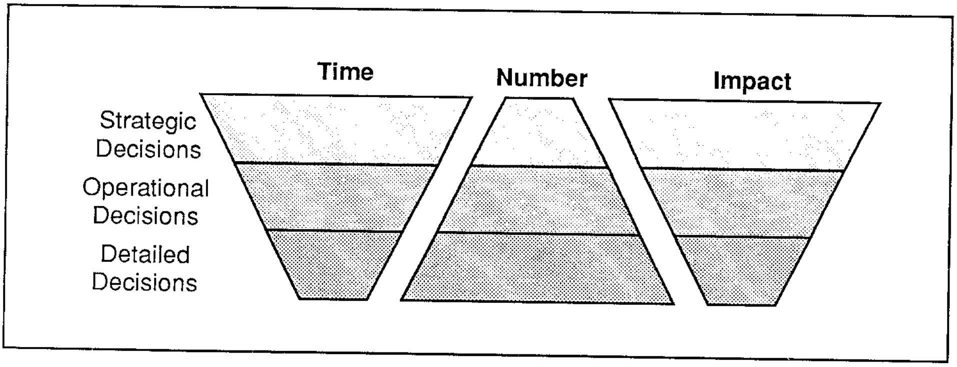


Figure 1. Characteristics of manufacturing decisions.

At the strategic level, the time which is available to make a decision is usually lengthy and the decisions have a large impact on the manufacturing organization. The frequency of decisions is low. An example of this class of decisions is the decision to construct a new manufacturing plant. Although this decision is made relatively rarely, the time required to make it is usually long, due to its complexity and possible impact.

At the operational level, the frequency of decisions tends to be greater, and each decision must be made in less time. Individual decisions have less impact on the organization. An example of this class of decisions is the master production scheduling decision that determines which portion of the planned workload in a manufacturing system should be allocated to each of several consecutive "time buckets."

Individual detailed decisions must be made quickly and have low impact on the manufacturing organization. However, since the frequency of these decisions is high, the aggregation of these decisions has a significant impact on the organization. The assignment of a production task to a particular machine is an example of this type of decision. When this assignment occurs in real time in reaction to events on the factory floor, it is called real-time scheduling, which is the subject of this paper.

In real-time scheduling, a *decision point* arises whenever one or more production *resources* become available after completing manufacturing tasks or after repair. A resource can be any individual production unit such as a single machine, an operator, or a manufacturing cell of machines grouped together with auxiliary devices (e.g., robots). Resources can be logically grouped into *work centers* according to common manufacturing function. The decision to be made in each work center at each decision point is which of the pending tasks (tasks that are ready to be performed) should be performed next on each of the available resources. The partitioning of this decision-making problem into work centers renders it more computationally manageable while still allowing the consideration of all feasible resource-task assignments. A feasible decision *alternative* is a list of resource-task assignments in which the resources are the ones that are available in a work center at the decision point and the tasks are selected from those pending at the work center at the decision point. For example, if two resources R_1 and R_2 are available and three tasks T_1, T_2, and T_3 are pending at a decision point, then the six possible alternatives are: ((R_1 T_1) (R_2 T_2)), ((R_1 T_2) (R_2 T_1)), ((R_1 T_1) (R_2 T_3)), ((R_1 T_3) (R_2 T_1)), ((R_1 T_2) (R_2 T_3)), and ((R_1 T_3) (R_2 T_2)). The schedule for a work center can be generated by deciding on alternatives at consecutive decision points over time; the schedule for an entire facility can be generated by combining the schedules of its constituent work centers.

Because of limits placed on computational burden in real-time scheduling, often a "good" decision must be made in lieu of the optimal decision. A "good" decision can be defined as one with a decision quality (DQ) which is within some interval Δ of the decision quality of the optimal decision. The efficiency of a manufacturing decision can be defined as the ratio of its quality to its computational burden.

Decision quality can be a function of multiple, possibly conflicting, criteria. These are generally different from organization to organization. An effective real-time scheduling procedure should be flexible in terms of the trade-off between decision quality and computational effort; that is, it should make full use of the available decision time in order to achieve the highest possible decision quality. The purpose of this paper is to motivate and describe such a procedure, and to present experimental results of its application.

A number of procedures have been developed to address the problem of real-time scheduling. One approach that has been extensively researched is the use of dispatch rules (Conway, Johnson, and Maxwell 1960; Panwalker and Iskander 1977; Stecke and Solberg 1981; Blackstone, Philips, and Hogg 1982; Malstrom 1983). A dispatch rule orders the tasks waiting to be assigned in a queue based on some task attribute. At each decision point, the dispatch rule is applied and the task at the head of the queue is selected for processing. If more than one resource is available at the decision point, the resource which will process the selected task must be chosen either randomly or heuristically. Thus dispatch rules partition the assignment problem into two parts, namely the selection of an available resource and the selection of a task to assign to the resource. Resources and tasks are not considered simultaneously, but rather *sequentially* at a decision point. Since dispatch rules consider only a limited amount of information (i.e., a single attribute of the pending tasks at a decision point), their computational burden is relatively slight. This computational burden cannot be adjusted to change decision quality because dispatch rules are "hard-wired" heuristics. The dispatch rule approach is limited in its ability to consider multiple criteria because most dispatch rules are designed to benefit a single criterion. In addition, although much research has been devoted to determining the conditions under which particular dispatch rules perform well (Elvers 1974; Rochette and Sadowski 1976), the results are difficult to generalize beyond the particular manufacturing systems and conditions studied, making the selection of a dispatch rule a difficult task.

A second approach to real-time scheduling partitions the overall problem into independent decision-making problems at each resource. These problems are treated via a class of scheduling policies similar to dispatch rules which determine the processing sequence of part types in the input buffer of each resource (Perkins and Kumar 1989). The decisions are made solely on the basis of part type levels in the input buffer. These policies differ from standard dispatch rules in that parts of the selected *type* are processed until none remain in the input buffer. A different part type is then selected to be cleared from the input buffer. Such policies, implemented independently at each resource, enable a manufacturing system to achieve prespecified demand rates for each part type, and are stable in the sense that they can be implemented without exceeding preset, finite input buffer capacities. Since the decision making is very distributed and each decision process considers only inventory levels in one input buffer, the computational burden of the approach is very small, making it easily implementable in real time. However, the partitioning of the overall problem by resource means that this approach is not suitable for systems in which multiple resources may perform a single task. This approach cannot make use of extra computation time, if any, to improve decision quality. It is also incapable of addressing multiple criteria, since the decision making considers only inventory levels.

A third approach combines discrete simulation with dispatch rules (Wu and Wysk 1989). In this approach, discrete simulation of a manufacturing system model is used to evaluate the performance of a set of plausible dispatching rules over a short planning horizon, Δt. The rule with the best simulated performance in the planning horizon is then applied to the physical system. At the end of Δt, the state of the physical system is incorporated into the simulation model. The evaluation/application process is carried on repeatedly. The use of different dispatch rules at different times is designed to overcome the weaknesses of any single rule. The decision quality versus computational burden of this approach may be adjusted by varying Δt or by varying the number n of dispatch rules that are simulated.

Increasing Δt and n increases the decision quality but reduces the ability of the method to respond in real time. This approach is limited in its ability to consider multiple criteria for the same reasons as the dispatch rule approach: most dispatch rules are designed to benefit a single criterion; in addition, the success of a dispatch rule with respect to different criteria depends in an unknown way on the structure of the manufacturing system and on the particular conditions (e.g., workload) which hold there. This makes the selection of the set of dispatch rules to be simulated a difficult task.

2. A game theory approach

In this paper, game theory (Rajan and Nof 1990; Nof 1991) is proposed as a tool for determining a good alternative at each decision point. The decision-making problem at each decision point can be represented as a two-person non-zero sum game in which the decision maker is player 1 and "nature" is player 2. The decision alternatives $\{\text{Alt}_1, \text{Alt}_2, \ldots, \text{Alt}_m\}$ are associated with the rows of the game matrix, while the criteria $\{\text{Crit}_1, \text{Crit}_2, \ldots, \text{Crit}_n\}$ by which the alternatives to be evaluated are associated with the columns. The consequence c_{ij} of the i^{th} alternative with respect to the j^{th} criterion forms the element in the i^{th} row and j^{th} column of the game matrix (figure 2). The utility of the i^{th} alternative may be calculated as the weighted sum of its (normalized) consequence values:

$$U_i = w_1\hat{c}_{i1} + w_2\hat{c}_{i2} + \ldots + w_n\hat{c}_{in} \qquad (1)$$

Here each $\hat{c}_{ij}$ is a consequence value that has been normalized so that it is dimensionless, and the higher its value, the more favorable the consequence (Keeney and Raiffa 1976). The alternative with the highest utility is selected. The game theory approach proposed in this paper, allows multiple-criteria decision making and the trade-off between execution time and quality of solution. Furthermore, such an approach makes relatively few restrictive assumptions about the problem.

	CRITERIA					
ALTERNATIVES	Crit_1	Crit_2	…	Crit_j	…	Crit_n
Alt_1	c_{11}	c_{12}	…	c_{1j}	…	c_{1n}
Alt_2	c_{21}	c_{22}	…	c_{2j}	…	c_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Alt_i	c_{i1}	c_{i2}	…	c_{ij}	…	c_{in}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Alt_m	c_{m1}	c_{m2}	…	c_{mj}	…	c_{mn}

Figure 2. Game theory representation of the real-time decision-making problem.

Ideally, the evaluation of an alternative's utility should be based on all of the information available at a decision point. This means that it should be based not just on the quality of the assignments in the alternative, but also on the quality of the assignments that are possible once the assignments in the alternative are implemented. Consider a situation in which there are two available resources R_1 and R_2 and three pending tasks T_1, T_2, and T_3. Let the sole decision criterion be the average cost of performing a task. A feasible alternative is ((R_1 T_1) (R_2 T_2)), with T_3 left unassigned. If only the costs of the assigned tasks T_1 and T_2 are considered in the evaluation of this alternative, then information about the cost of T_3 will not have been utilized. In order to incorporate the information about the cost of T_3, one may evaluate the alternative by considering its *samples*—namely, the ways in which *all* of the pending tasks may be assigned, given that the assignments in the alternative are fixed. For the given alternative, there are two samples: ((R_1 T_1 T_3) (R_2 T_2)) and ((R_1 T_1) (R_2 T_2 T_3)). That is, T_3 may be performed on R_1 after the completion of T_1, or it may be performed on R_2 after the completion of T_2. We may then evaluate the alternative ((R_1 T_1) (R_2 T_2)) by computing the *average* cost per task of its two samples. In general, we may define the utility of an alternative to be the *average* utility of its samples.

Given this definition of an alternative's utility, we may make two observations for motivating a particular game theory approach to real-time decision making.

Observation 1

If some maximum number of alternatives (*MNA*) are formed randomly from a population of N possible alterntives, then the probability P_0 of forming the best alternative is given by *MNA*/*N*. For decisions of even moderate size, N may be prohibitively large. For example, if five resources are available and 20 tasks are pending, then N is

$$\frac{20!}{(20-5)!} = 1{,}860{,}480. \qquad (2)$$

Thus it is unlikely that P_0 can be made to approach the optimum value of 1 in a real-time scheduling application.

However, a more positive picture emerges if a "good" alternative will suffice. An alternative is "good" if its utility is within some Δ of the optimal utility $u_{\max}$. For a given decision, if we characterize the distribution of the alternatives' utility values by a continuous density function $f(x)$, then the probability P_Δ that the utility of one randomly formed alternative lies within Δ of the optimal utility $u_{\max}$ can be approximated as:

$$P_\Delta = \int_{u_{\max}-\Delta}^{\infty} f(x)\, dx \qquad (3)$$

Now, consider the fact that up to *MNA* alternatives can be formed at each decision point. If these are formed randomly, the probability $P(MNA, \Delta)$ of forming at least one good alternative is:

$$P(MNA,\ \Delta) = 1 - (1 - P_\Delta)^{MNA} \qquad (4)$$

This probability measures the quality of the alternatives formation process. Figure 3 shows an example of $P(MNA, \Delta)$ versus MNA assuming that $f(x)$ is a normal probability density function with a mean μ of 70 and a standard deviation σ of 10. The best utility $u_{\max}$ is defined to be $\mu + 3\sigma$, which is 100 in this case. Figure 4 shows $P(MNA, \Delta)$ versus Δ for three values of σ, the standard deviation of $f(x)$. Again this distribution is assumed to be normal with a mean μ of 70. Δ is assumed to be 5. Using the $\mu + 3\sigma$ definition, the three values of $u_{\max}$ in this case can be computed to be 76, 100, and 130. In these examples, P_Δ can be approximated as:

$$P_\Delta = \int_{u_{\max}-\Delta}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{5}$$

From equation (4) and figure 4, it can be seen that, regardless of the distribution which determines P_Δ, $P(MNA, \Delta)$ increases sharply at first as MNA is increased and then levels off. Thus, by forming a (relatively small) subset MNA out of N possible alternatives at random, just enough to reach the point at which $P(MNA, \Delta)$ levels off, the probability of forming a "good" alternative can be made to approach the probability that would result if a much greater number of alternatives are formed. This behavior is more pronounced for larger values of Δ (figure 3), meaning that as the definition of a "good" alternative becomes less stringent, the measure $P(MNA, \Delta)$ of the quality of the alternatives formation process increases. This behavior is also more pronounced for smaller values of σ (figure 4). As σ decreases, P_Δ increases, leading to an increase in the value of $P(MNA, \Delta)$. This means that as the distribution of the alternatives' utilities becomes narrower, the likelihood of forming a good alternative increases.

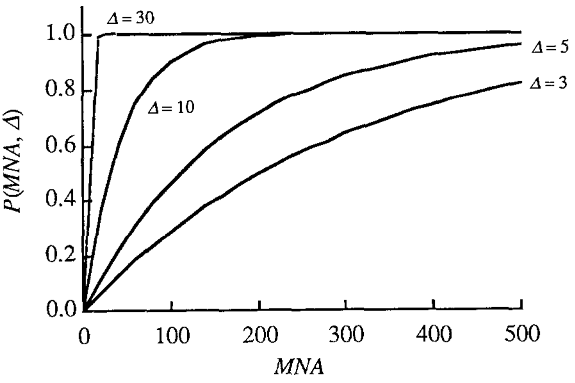


Figure 3. $P(MNA, \Delta)$, the probability of forming at least one good alternative, versus MNA for several values of Δ.

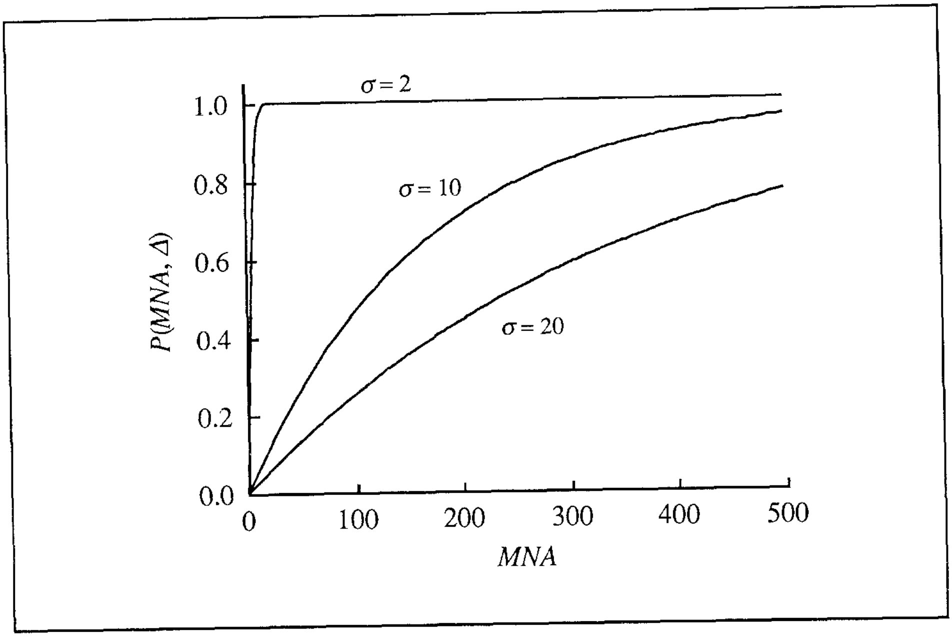


Figure 4. P(*MNA*, Δ) versus *MNA* for several values of σ.

Observation 2

Although the mean utility of *all* samples of an alternative defines the alternative's true utility, the mean utility of a fewer number of samples *SR* can be used as an *estimate* of the alternative's utility. The goodness of this estimate is given by the probability that the estimated utility lies within some δ of the true utility. If we characterize the distribution of the utility values of a particular alternative's samples by a continuous probability density function $g(x)$, then the probability P_δ that the estimated utility lies within δ of the true utility $\bar{u}_0$, if only *one* sample is used to estimate $\bar{u}_0$, can be written as:

$$P_\delta = \int_{\bar{u}_0-\delta}^{\bar{u}_0+\delta} g(x)\, dx \tag{6}$$

Note that $\bar{u}_0$ is by definition the mean of the distribution defined by $g(x)$.

Now consider the case where the mean utility of *SR* samples is used as the estimate of $\bar{u}_0$. Evaluation of this probability requires a knowledge of the distribution of estimated utilities $\bar{u}_{SR}$ with a density function $g'(x)$. This distribution can be approximated by a normal distribution for two reasons. First, the utilities of an alternative's samples tend to be normally distributed because of the additive procedure by which they are calculated. Consider an example. Let the available resource be R_1 and the pending tasks be T_1, T_2, T_3, T_4, T_5, T_6, T_7, and T_8. A feasible alternative is (R_1 T_1), and one sample (sample 1, say) of this alternative is (R_1 T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8). The utility of such a sample is usually defined as the mean value of some combination of criteria for the tasks in the sample. For example, if the criterion is tardiness, then the utility of this sample would be calculated as

$$u_1 = \frac{\mathrm{Tard}_{1,1} + \mathrm{Tard}_{1,2} + \mathrm{Tard}_{1,3} + \mathrm{Tard}_{1,4} + \mathrm{Tard}_{1,5} + \mathrm{Tard}_{1,6} + \mathrm{Tard}_{1,7} + \mathrm{Tard}_{1,8}}{8} \qquad (7)$$

where $\mathrm{Tard}_{1,j}$ represents the tardiness of task j if the resource i, processes the tasks in the sequences specified in the sample. The utilities of the alternative's other samples are similarly calculated. Irrespective of the distribution of the $\mathrm{Tard}_{i,j}$'s, utilities of the form in equation (7) tend to be normally distributed by the central limit theorem. The second justification for this assumption comes from a separate application of the central limit theorem. Let σ_0 be the standard deviation of the distribution $g(x)$ of the utilities of all of the alternative's samples; for sufficiently large SR, $\bar{u}_{SR}$ is approximately a normal random variable with mean $\bar{u}_0$ and standard deviation $\sigma_0/\sqrt{SR}$ (central limit theorem). Therefore, the probability $P(SR, \delta)$ that the estimated utility $\bar{u}_{SR}$ is within δ of the true utility $\bar{u}_0$ is approximately

$$P(SR, \delta) = \int_{\bar{u}_0-\delta}^{\bar{u}_0+\delta} \frac{\sqrt{SR}}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{SR\,(x-\bar{u}_0)^2}{2\sigma_0^2}\right) dx \qquad (8)$$

This is a measure of the quality of the evaluation of an alternative. Figure 5 shows $P(SR, \delta)$ versus SR assuming that the distribution of the alternative's sample utilities has a mean ($\bar{u}_0$) of 70 and a standard deviation (σ_0) of 10. Figure 6 shows $P(SR, \delta)$ versus SR for several values of σ_0, assuming that δ is 5.

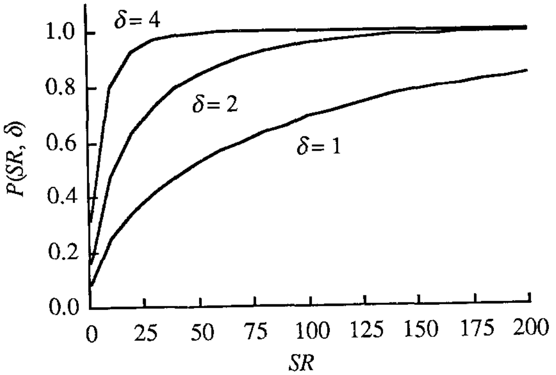


Figure 5. $P(SR, \delta)$ versus sampling rate SR for various values of δ.

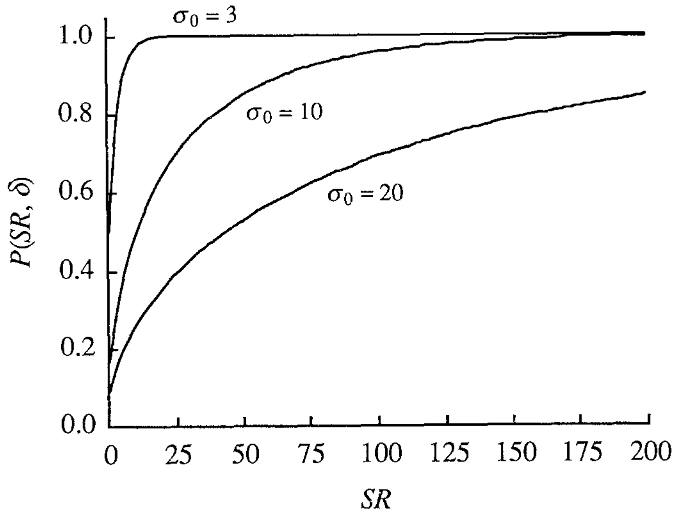


Figure 6. $P(SR, \delta)$ versus sampling rate SR for various values of σ_0.

As the sampling rate (SR) increases, $P(SR, \delta)$ increases rapidly at first and then levels off. Thus although the quality of the evaluation of alternatives improves as SR increases, the amount of improvement levels off after a certain level of SR is reached. This behavior is more pronounced for larger values of δ (figure 5). It is also more pronounced for smaller values of σ_0 (figure 6), meaning that as the distribution of the utilities of an alternative's samples becomes more concentrated about their mean (the true utility), the likelihood that the estimated utility will be within δ of the true utility $P(SR, \delta)$ increases. The quality of the evaluation of alternatives, as measured by $P(SR, \delta)$, has the same type of behavior as the quality of the formation of alternatives, as measured by $P(MNA, \Delta)$. That is, the quality of evaluation at lower values of SR can rival that at much higher values of SR.

Based on the above observations, the following procedure is proposed for making the assignment decisions required in real-time scheduling (Chryssolouris, Wright, Pierce, and Cobb 1988; Chryssolouris, Pierce and Dicke 1991, 1992; Chryssolouris, Lee, and Dicke 1991; Chryssolouris, Dicke and Lee 1992).

1. Form alternatives
   - Establish the maximum number of alternatives (MNA) that can be considered out of the total set of N alternatives $\{\mathrm{Alt}_i\}$.
   - If $MNA \leq N$, then randomly form (without replacement) MNA alternatives out of $\{\mathrm{Alt}_i\}$ and proceed with step 2; skip the step directly below.
   - If $MNA > N$, then form all N alternatives and proceed with step 2.

2. Establish criteria
   - Establish the decision-making criteria by which the formed alternatives will be evaluated.
3. Evaluate alternatives
   - For each alternative formed construct *SR* samples, and calculate the utility of each sample.
   - Calculate the mean utility of the *SR* samples and use this as an estimate of the utility of the alternative.
4. Select the best alternative
   - Implement the alternative with the highest estimated utility.

3. Statistical aspects

Since $P(MNA, \Delta)$ (figures 3 and 4) represents the quality of the formation of alternatives, and $P(SR, \delta)$ (figures 5 and 6) represents the quality of the evaluation of alternatives, we can define the probability

$$\begin{aligned} P_x &= P(MNA, \Delta)\, P(SR, \delta) \\ &= [1 - (1 - P_\Delta)^{MNA}] \int_{\bar{u}_0-\delta}^{\bar{u}_0+\delta} \frac{\sqrt{SR}}{\sqrt{2\pi}\,\sigma_0} \exp\left(- \frac{SR\,(x - \bar{u}_0)^2}{2\sigma_0^2}\right) dx \qquad (9) \end{aligned}$$

to be a measure of the quality of the decision-making process as a whole. P_x is the probability that among the maximum of *MNA* alternatives that are formed and evaluated at a decision point, at least one alternative has a utility that is within Δ of the utility ($u_{\max}$) of the best alternative, and, in addition, for each alternative that is evaluated, the value of the utility is estimated to within δ of the true value. We assume, for the purposes of calculating P_x, that a single representative standard deviation σ_0 characterizes the distribution of sample utilities for each and every alternative. In other words, $P(SR, \delta)$ is the same for all alternatives. Although this is a rough approximation, it will suffice for the drawing of qualitative predictions from the resulting P_x.

In figure 7, P_x is plotted versus both the maximum number of alternatives *MNA* and the sampling rate *SR*. The two plots can be viewed as theoretically derived decision quality (DQ) versus computational burden (CB) curves for the proposed decision-making procedure. As CB is increased (by increasing *MNA* or *SR*), the increase in DQ (P_x) is rapid at first but then levels off.

An observation can be made with regard to the relative effectiveness of increasing DQ via increasing *MNA* or increasing DQ via increasing *SR*. Qualitatively, figure 7 shows that increasing *MNA* is more effective in increasing DQ than increasing *SR*. This can be quantitatively expressed by observing the *saturation point:* the CB at which the DQ (as measured by P_x) attains 95% of its highest value. Below this point, increasing CB appreciably increases DQ; above this point, little increase in DQ can be achieved, no matter how much CB is increased. The saturation points for the P_x versus *MNA* plots occur at 460 units or greater (depending on *SR*), while they occur at less than 20 units for the P_x versus *SR*

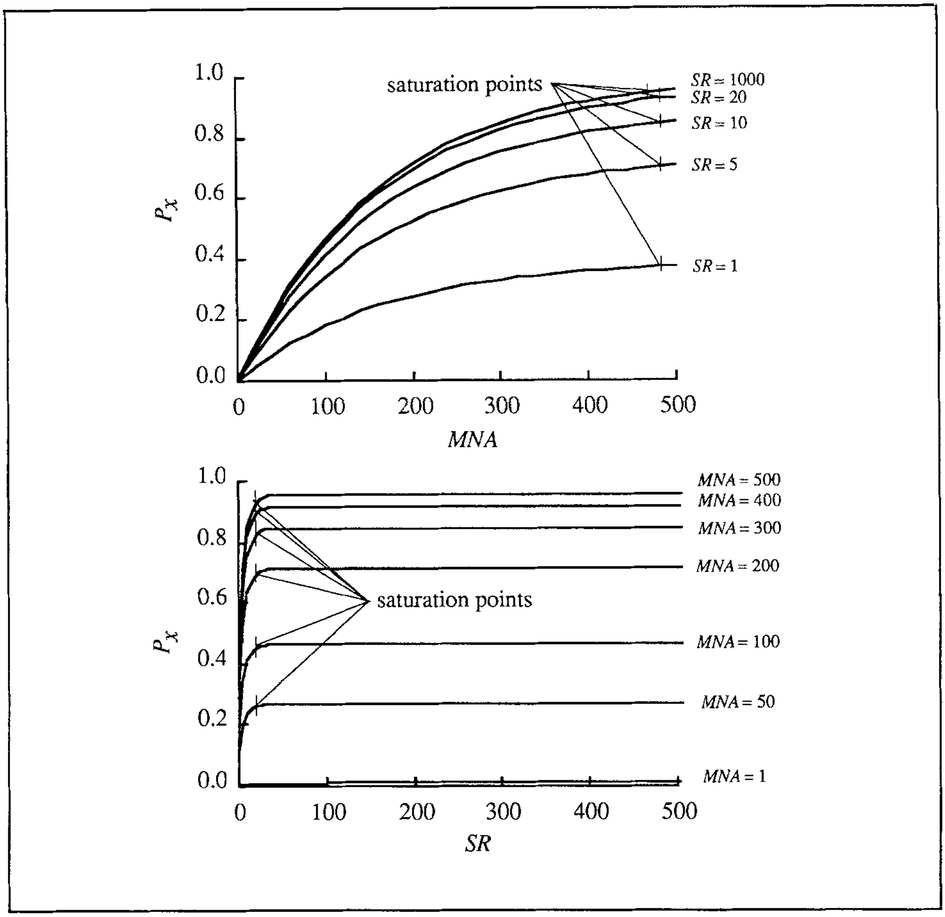
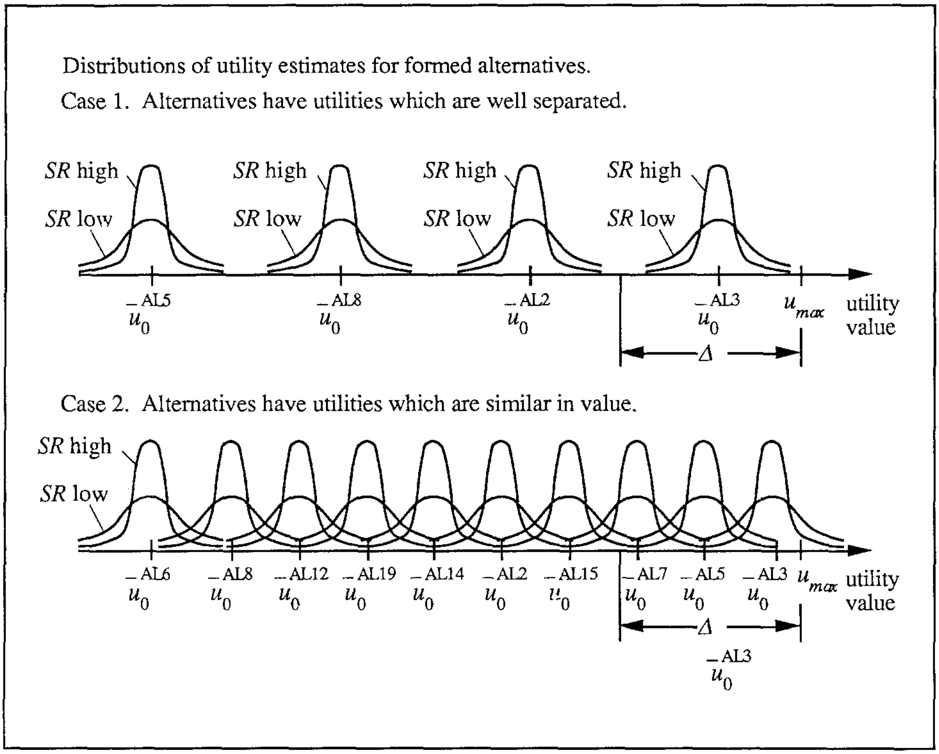


Figure 7. P_x versus *MNA* and P_x versus *SR*. The distribution of alternatives' utilities are normal with mean 70 and standard deviation 10. Δ is 5. The distribution of sample utilities for each alternative is normal with standard deviation 10. δ is 5.

plots (depending on *MNA*). Thus *a given computational capacity is best invested by devoting most of it to forming alternatives (increasing MNA) and relatively little of it to evaluating alternatives (increasing SR)*. In actuality, formation of alternatives (increasing *MNA*) takes much less time than evaluation of alternatives (increasing *SR*). This only strengthens the above conclusion.

Another way to examine the effect of the sampling rate *SR* on DQ is the following. A measure of DQ is the probability P_y that the alternative that is implemented will have a utility $\bar{u}_0$ that is within a desired range Δ of the maximum utility $u_{\max}$. Figure 8 shows the distributions of the estimated utilities in two representative cases. In the first case, the formed alternatives have utilities which are far apart. Since the distributions have little overlap for a wide range of values of *SR*, the selected alternative will very likely be the alternative with the highest $\bar{u}_0$. As this utility lies within the desired range, P_y will be very


Distributions of utility estimates for formed alternatives.

Case 1. Alternatives have utilities which are well separated.

Case 2. Alternatives have utilities which are similar in value.

Figure 8. The effect on decision quality of the decision parameter *SR* (sampling rate).

close to 1. In this case the influence of the sampling rate *SR* is relatively low. In the second case, the formed alternatives have utilities which are close together. Here, the overlap of the distributions of estimated utilities, when *SR* is small, makes it likely that any of the few top alternatives will be selected. Increasing *SR* makes the distributions narrower, reducing the amount of overlap and making the selection of the top alternative more likely. However, since the few top alternatives are all within the desired range, P_y is again not greatly affected by *SR*.

4. Experimental procedure

Simulation tests were conducted to verify the predicted characteristics of the proposed game theory-based approach to real-time scheduling.

Test 1

The first characteristic of the approach that was verified was its flexibility with respect to the trade-off between decision quality and computational burden. Specifically, the influence of the decision parameters maximum number of alternatives (*MNA*) and sampling rate (*SR*) on decision quality and on computational burden were investigated.

Simulation runs were performed on a system consisting of one work center with three resources. The workload of the system consisted of 20 independent tasks, all of which arrived at time 0. Each time one or more resources became available during the course of the simulation, the game theory procedure was used to assign a task to each available resource. The processing times of the tasks were integers generated uniformly at random from the range [1, 59]. All resources were capable of processing all tasks. Each simulation was run until all 20 tasks were completed.

The values of the decision parameters *MNA* and *SR* were varied across the simulation runs. Ten values of *MNA* were used: 3, 5, 10, 20, 30, 50, 100, 200, 500, and 1000. Ten values of *SR* were used: 1, 2, 4, 8, 12, 18, 25, 30, 40, and 50. There were thus 100 combinations of decision parameters. Each combination was simulated five times, and the results of these simulations were averaged, yielding a total of 500 simulation runs for test 1.

A single decision criterion, mean flow time, was used for all simulations. The flow time of a task is the difference between its time of completion and its time of arrival into the work center (time 0, for these simulations). At each decision point, the utility of each alternative was estimated to be the mean flow time of the tasks in *SR* randomly formed samples. For example, if *SR* was 2, and one resource R_1 was available with four pending tasks T_5, T_9, T_{11}, and T_{20} remain to be assigned, then the utility of the alternative (R_1 T_5) was found by first forming two samples at random, say (R_1 T_5 T_{20} T_9 T_{11}) and (R_1 T_{20} T_5 T_{11} T_9), and then calculating the average flow time of the tasks in these samples:

$$u(\mathrm{R}_1\ \mathrm{T}_5) = \frac{\mathrm{Flow}_{1,5} + \mathrm{Flow}_{1,9} + \mathrm{Flow}_{1,11} + \mathrm{Flow}_{1,20} + \mathrm{Flow}_{2,5} + \mathrm{Flow}_{2,9} + \mathrm{Flow}_{2,11} + \mathrm{Flow}_{2,20}}{8} \tag{10}$$

where: $\mathrm{Flow}_{i,j}$ ≡ flow time (difference between completion time and arrival time) of task j in sample i.

The decision quality of a final schedule was defined to be the mean flow time (MFT) of all 20 tasks in the schedule:

$$\text{decision quality (DQ)} = \text{MFT} = \frac{\sum_{j=1}^{20} (t_j^{\text{comp}} - t_j^{\text{arr}})}{20} \tag{11}$$

where: t_j^{comp} ≡ completion time of task j;
t_j^{arr} ≡ arrival time of task j (= 0).

Test 2

The second aspect of the approach that was investigated was the dependence of its behavior on processing time variance. Intuitively, as processing time variance becomes smaller, the differences in the mean flow times (equation (11) of different schedules should become less significant. (In the limit of equal processing times for all tasks, mean flow time is the same for all schedules.) Thus the decision quality should be fairly constant across a wide range of computational effort (across a wide range of values of *MNA* and *SR*).

In the theoretical analysis, this type of behavior is predicted for the case in which the variance σ^2 of the distribution of the alternatives' utilities at a decision point (defined by $f(x)$ in equation (3) is small relative to the distance Δ from the optimal utility which defines a "good" alternative. Therefore simulation experiments were conducted to verify the following statements:

1. A narrow (wide) processing time distribution results in a narrow (wide) distribution of the alternatives' utilities at a decision point.
2. If statement 1 is true, then, in accordance with the theoretical analysis, a narrow processing time distribution creates a situation in which the decision quality remains fairly constant across a wide range of computational effort (across a wide range of values of *MNA* and *SR*).

In order to verify statement 1, two simulations were performed using the same facility as test 1. Again the single decision criterion mean flow time (equation (10)) was used. For the first simulation, the workload was identical to that of test 1. The processing times of the 20 tasks were integers uniformly distributed in the range [1, 59]. For the second simulation, the processing times of the 20 tasks were more narrowly distributed, being integers uniformly distributed in the range [25, 35]. For both simulations, 1000 (*MNA*) alternatives—out of a possible 20!/(20-3)! or 6840—were formed at the first decision point, and the utility of each of these alternatives was estimated as the mean flow time of 50 (*SR*) randomly formed samples. These utilities were accumulated in a histogram in order to show the distribution of their values. One histogram was construced for each simulation.

In order to verify statement 2, the experiments of test 1 were repeated, with the difference that the processing time distribution of the 20 tasks was changed from uniform in the range [1, 59] to uniform in the range [25, 35].

5. Results and discussion

Test 1: Influence of the decision parameters on the decision quality and the computational burden

Figure 9 shows the computational burden, as measured by the CPU time required to execute a simulation in test 1, versus *MNA* and versus *SR*. Each data point of the CPU time versus *MNA* graph (figure 9(a)) is the average CPU time across all 10 values of *SR* for the given value of *MNA*. Similarly, each data point of the CPU time versus *SR* graph (figure 9(b)) is the average CPU time across all 10 values of *MNA* for the given value of *SR*. The graphs show linear relationships between CPU time and the values of *MNA* and *SR*. This justifies the use of *MNA* and *SR* as surrogate measures of computational burden in the theoretical analysis. It also shows that the computational burden per unit increase in *MNA* is an order of magnitude less than the computational burden per unit increase in *SR*.

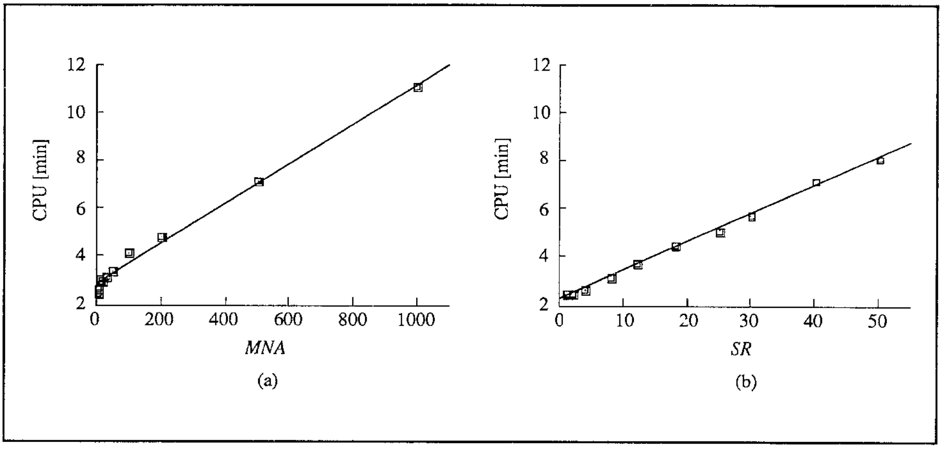


Figure 9. Computational burden, as measured by CPU time, versus *MNA* and *SR*.

Figure 10 shows decision quality, as measured by mean flow time MFT (equation (11)), versus *MNA* and versus *SR*. The lower the MFT, the higher the decision quality. Each point in figure 10(a) is the mean of the MFTs across all 10 values of *SR* for the given value of *MNA*. Similarly, each point in figure 10(b) is the mean of the MFTs across all 10 values of *MNA* for the given value of *SR*. Both graphs show that as computational burden (*MNA* or *SR*) is increased, the decision quality increases sharply at first and then levels off. This is consistent with the behavior predicted in the theoretical analysis (figure 7). The "leveling off" of decision quality occurs at a lower value of *SR* than one of *MNA*. This indicates that decision quality is more effectively increased by increasing *MNA* than by increasing *SR*, which confirms the theoretical analysis.

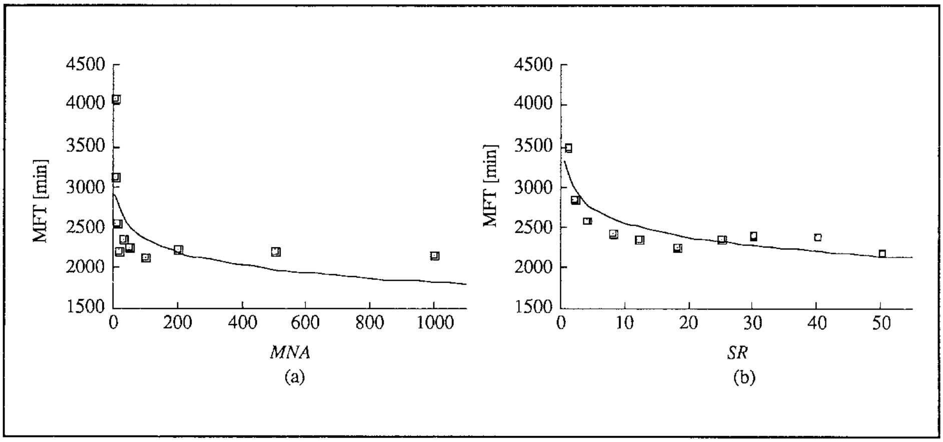


Figure 10. Decision quality, as measured by mean flow time, versus *MNA* and *SR* for the [1, 59] uniform processing time distribution.

This point can be quantitatively made by graphing the efficiency of the decision-making process versus *MNA* and versus *SR* (figure 11). Efficiency is defined as the ratio of decision quality (the reciprocal of MFT) to computational burden (CPU time):

$$\text{Efficiency} = \frac{1}{\text{MFT} \cdot \text{CPU Time}} \tag{12}$$

Each data point in figure 11(a) is the average efficiency across all 10 values of *SR*, for the given value of *MNA*. Similarly, each data point in figure 11(b) is the average efficiency across all 10 values of *MNA*, for the given value of *SR*. Peak efficiency is attained at about *MNA* = 20 and at about *SR* = 3, confirming the higher *MNA*, lower *SR* prescription.

Test 2: Influence of the processing time distribution on the relationship of decision quality to computational burden

In order to establish the influence of the processing time distribution on the relationship of decision quality to computational burden, we first establish its relationship to the distribution of the utilities of the alternatives (defined at $f(x)$ in equation (3)) at a decision point. Figure 12 shows the task processing time distributions used in two simulations (uniform with minimum = 1, maximum = 59 and uniform with minimum = 25, maximum = 35) along with the resulting distributions $f(x)$ at the first decision point in each simulation. The graphs confirm that narrow (wide) processing distributions result in narrow (wide) utility distributions at a decision point.

According to the theoretical analysis, when distributions of the alternatives' utilities at decision points are narrow, the decision quality remains fairly constant across a wide range of computational effort (across a wide range of values of *MNA* and *SR*). Therefore we would

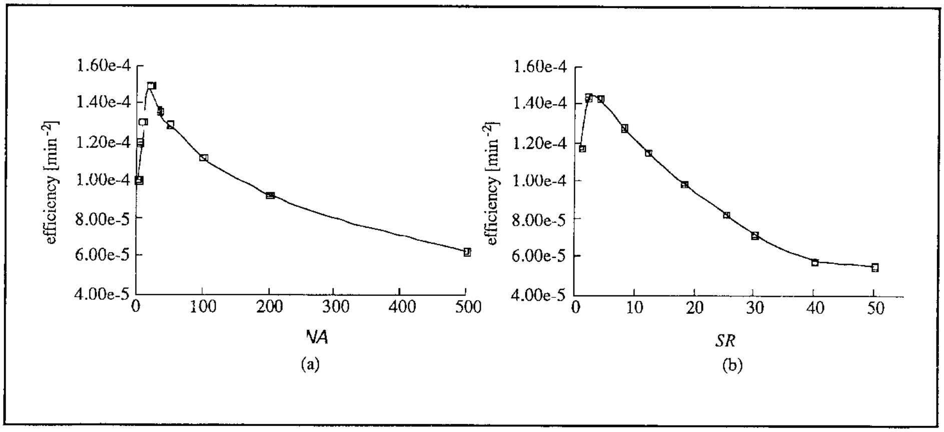


Figure 11. Efficiency versus *MNA* and *SR* for the [1, 59] uniform processing time distribution.

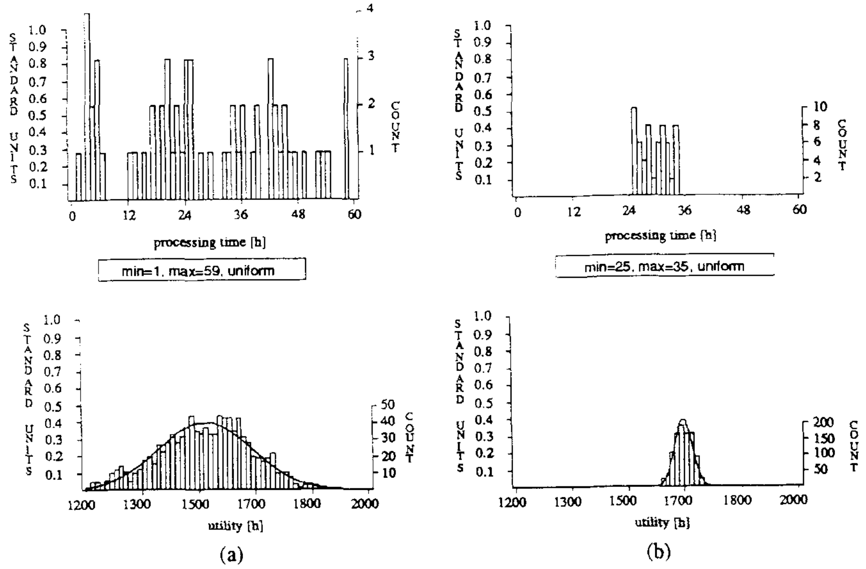


Figure 12. Influence of processing time distribution on the distribution of alternatives' utilities at a decision point.

expect that the effect of the invested computational effort should be minimal when the processing time distribution is narrow. This is shown in figure 13. By contrast, the behavior of decision quality versus *MNA* and *SR* in the case of a wide processing time distribution is shown in figure 10.

When the processing time distribution is wide, there is a decision quality benefit to be obtained by increasing the computational burden up to a certain point (figure 10). However, if the processing time distribution is narrow, then decision quality is fairly uniform for across the entire range of computational burdens and it makes sense only to invest the minimum computational burden possible (figure 13). This is reaffirmed by the efficiency versus computational burden (*MNA* and *SR*) graphs for the narrow processing time distribution case (figure 14). Peak efficiency is achieved at the lowest possible values of *MNA* and *SR* as opposed to *MNA* = 20 and *SR* = 3 for the wide processing time distribution case (figure 11).

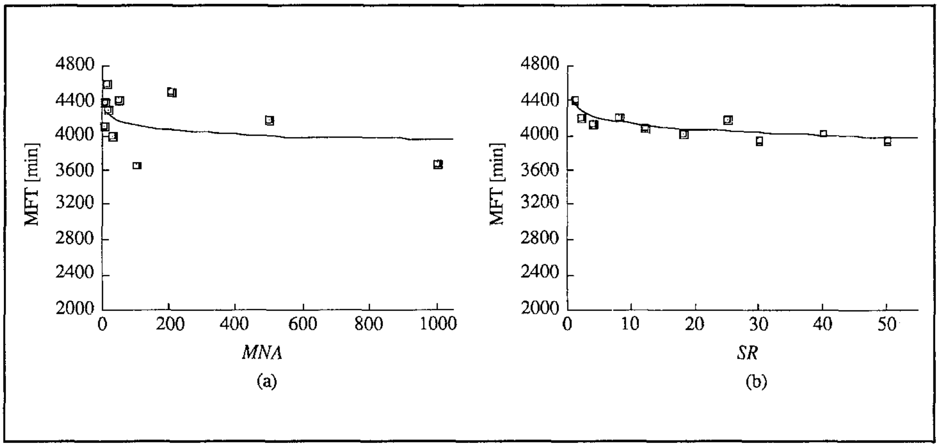


Figure 13. Decision quality, as measured by mean flow time, versus *MNA* and *SR* for the [25, 35] uniform processing time distribution.

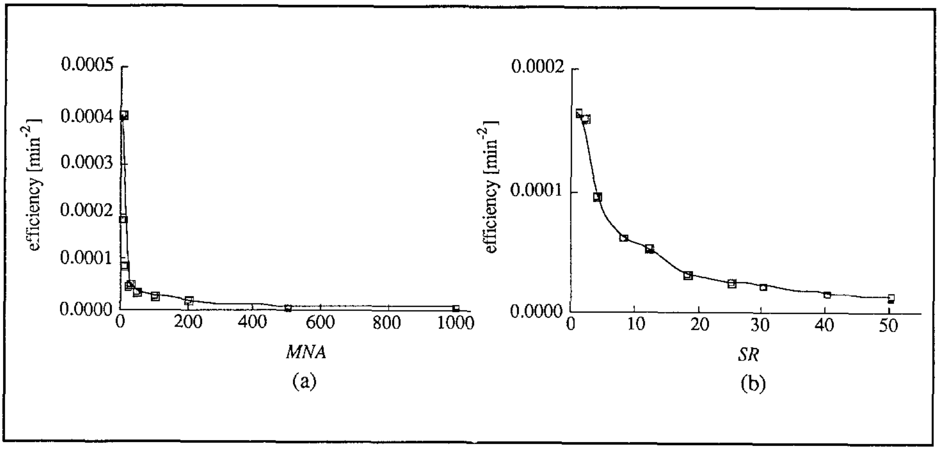


Figure 14. Efficiency versus computational burden for the narrow processing time distribution [25, 35].

6. Conclusions

The experimental results confirm the main points of the theoretical analysis regarding the proposed real-time scheduling procedure. Although the analysis is based on the consideration of probabilities at a single decision point, it successfully describes the behavior of the procedure over a sequence of decision points. The game theory procedure is flexible both with respect to the number and types of scheduling criteria and with respect to the decision quality versus computational burden trade-off. It can be tailored, via the decision parameters *MNA* and *SR*, to attain the best decision quality possible for a given computational capacity.

The real-time scheduling procedure is largely based on the power of random selection. As demonstrated in the theoretical analysis, formation and evaluation of a small fraction of the total number of alternatives at a decision point can yield a decision quality that is comparable to that which is obtained from an exhaustive and (likely) computationally infeasible enumeration of all possible alternatives. This is an intrinsic property of the real-time decision-making problem which is fully exploited by the proposed procedure.

References

Blackstone, J.H., Philips, D.T., and Hogg, G.L., "A State of the Art Survey of Dispatching Rules for Manufacturing Job Shop Operations," *International Journal of Production Research*, Vol. 20, No. 1, pp. 27–45 (1982).

Chryssolouris, G., Dicke, K., and Lee, M., "On the Resources Allocation Problem," *International Journal of Production Research*, Vol. 30, No. 12, pp. 2773–2795 (1992).

Chryssolouris, G., Lee, M., and Dicke, K., "An Approach to Short Interval Scheduling for Discrete Parts Manufacturing," *International Journal of Computer-Integrated Manufacturing*, Vol. 4, No. 3, pp. 157–168 (1991).

Chryssolouris, G., Pierce J., and Dicke, K., "An Approach for Allocating Manufacturing Resources to Production Tasks," *Journal of Manufacturing Systems*, Vol. 10, No. 5, pp. 368–382 (1991).

Chryssolouris, G., Pierce, J., and Dicke, K., "A Decision-Making Approach to the Operation of Flexible Manufacturing Systems," *International Journal of Flexible Manufacturing Systems*, Vol. 4, Nos. 3/4, pp. 309–330 (June 1992).

Chryssolouris, G. Wright, K., Pierce, J., and Cobb, W., "Manufacturing Systems Operation: Dispatch Rules Versus Intelligent Control," *Robotics and Computer-Integrated Manufacturing*, Vol. 4, Nos. 3/4, pp. 531–544 (Spring 1988).

Conway, R.W., Johnson, B.M., and Maxwell, W.L., "An Experimental Investigation of Priority Dispatching," *Journal of Industrial Engineering*, Vol. 11, No. 3, pp. 221 (1960).

Elvers, D.E., "The Sensitivity of the Relative Effectiveness of Job Shop Dispatching Rules with Various Arrival Distributions," *Transactions of the American Institute of Industrial Engineers*, Vol. 6, pp. 41 (1974).

Keeney, R. and Raiffa, H., *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, John Wiley and Sons, New York, NY (1976).

Malstrom, E.M., "A Literature Review and Analysis Methodology for Traditional Scheduling Rules in a Flexible Manufacturing System," *Final Technical Report performed under CAM-I Contract LA-83-FM-01* (1983).

Nof, S.Y., "Game Theoretic Models for Planning Cooperative Robotic Work," in *Proceedings of the 17th NSF Design and Manufacturing Systems Conference*, Austin, TX, pp. 553–556 (January 1991).

Panwalker, S.S. and Iskander, W., "A Survey of Scheduling Rules," *Operations Research*, Vol. 25, No. 1, pp. 45–61 (1977).

Perkins, J.R. and Kumar, P.R., "Stable, Distributed, Real-Time Scheduling of Flexible Manufacturing/Assembly/Disassembly Systems," *IEEE Transactions on Automatic Control*, Vol. 34, No. 2, pp. 139–148 (February 1989).

Rajan, V.N. and Nof, S.Y., "A Game-Theoretic Approach for Co-operation Control in Multimachine Workstations," *International Journal of Computer Integrated Manufacturing*, Vol. 3, No. 1, pp. 47–59 (1990).

Rochette, R. and Sadowski, R.P., "A Statistical Comparison of the Performance of Simple Dispatching Rules for a Particular Set of Job Shops," *International Journal of Production Research*, Vol. 14, p. 63 (1976).

Stecke, K.E. and Solberg, J.J., "Loading and Control Policies for a Flexible Manufacturing System," *International Journal of Production Research*, Vol. 19, No. 5, pp. 481–490 (1981).

Wu, S.Y. and Wysk, R., "An Application of Discrete-event Simulation to On-line Control and Scheduling in Flexible Manufacturing," *International Journal of Production Research*, Vol. 27, No. 9, pp. 1603–1623 (1989).