Selection of Dispatching Rules in Simulation-Based Scheduling of Flexible Manufacturing

Chuda Basnet
Department of Management Systems
The University of Waikato, Hamilton, New Zealand
chuda@waikato.ac.nz

Abstract— Discrete-event simulation has been proposed as a tool for the real-time selection of dispatching rules in the scheduling of flexible manufacturing systems (FMS). In this approach, a look-ahead simulation is used to predict the trajectory of the system under different dispatching rules. Then a rule is selected for use in the FMS, based on the performance criterion associated with each trajectory. However, this raises the issue of the criterion to be used for the selection of the rules. A simulation model of an FMS was used to test the effect of rule selection procedure on two performance measures. This simulator was run using four scheduling policies and the performance of the FMS was measured. The main finding of this research is that the use of global criteria gives better results than the use of local criteria.

Keywords- flexible manufacturing systems; simulation-based scheduling; dispatching rules; global control criterion

I. INTRODUCTION

An FMS consists of several computer-numerically-controlled (CNC) workstations that can process a variety of parts. These parts are mounted on pallets that are transported by automatic transportation devices. The CNC machines have tool magazines on which some tools may be stored for automatic changeover as needed. Processing can be done on alternate machines if the machines are suitably tooled. The whole system is directed by a supervisory computer. FMSs are particularly suitable when a variety of products is to be manufactured in small lots. However, operating an FMS is exceedingly complex due to the many factors that come into play: the availability of pallets, the alternate routing of parts, the availability of material handling devices, the capacity of the machines, the place available in tool magazines, and the dynamic arrival of jobs. Two modes of operation can be identified for the FMSs: dedicated and non-dedicated (random). In the dedicated mode of operation [8] the FMS processes a fixed set of part types usually in a fixed ratio. The FMS is tooled for these particular parts, and often even the routing is predetermined. In the non-dedicated mode of operation of FMS, the number of part types processed simultaneously by the FMS is much higher than that of the dedicated FMS. The set of parts is not fixed, nor is the production ratio of the parts. The problem of scheduling a random FMS is often decomposed into two subproblems. The first problem is of selecting (releasing) parts for manufacturing in a planning horizon (a shift). Parts are released in accordance with the production capacity and the tool-magazine capacity of the machines. The second problem is that of processing these parts within the bounds of the available pallets, material handling devices, and the numerically controlled machines. This later problem can be addressed by the use of dispatching rules, which are scheduling rules that are followed to allocate an available resource to a job from among the jobs waiting for this resource. For instance, a job with the least requirement of this resource may be selected.

Naturally, the question of the selection of these rules arises. Which of these rules should be used? Recourse may be made to the many studies of these rules. One approach under investigation is the use of a look-ahead discrete-event simulation to predict the trajectory of the system in hand under different dispatching rules. The rule may then be selected on the basis of the performance criteria obtained in each trajectory. This approach is gaining attention because of the increasing speed of the microcomputers and with the advent of distributed processing. This approach is especially attractive since a discrete-event simulation model can be built to any amount of detail to capture the aforementioned complexity of an FMS. In this approach, a simulation model of the FMS is built which closely mirrors the actual operating FMS. The simulation model is updated periodically with the latest real-time data captured from the FMS, and the simulator...
is operated under the candidate dispatching rules. The choice of the dispatching rule is then made following some performance criterion. The selected rule is used for the next period, and the cycle is repeated. Figure 1 illustrates this approach.

![Image: Simulation-Based Approach to Scheduling](image_url)

**Figure 1. Simulation-Based Approach to Scheduling**

There are many research questions engendered by this approach. How can statistically valid conclusions be derived from these look-ahead simulations? Which decision alternatives should be examined? What criteria should be used in the selection of the decision alternatives? How long should the trajectory be extended by way of simulation? Is this approach feasible for on-line decision making? This paper seeks to gain some insight into the question of the selection of the decision alternatives. A simulated environment of a hypothetical FMS was used to investigate the effect of the selection procedures on the steady state performance of an FMS. We present the results of this experiment.

The next section presents a review of the literature. Then the research problem is discussed. The experimental setup is described next. This is followed by the results of the experiment. Finally, concluding remarks are presented.

II. LITERATURE REVIEW

The mathematics of scheduling problems can be prohibitively complex [5]. However, dispatching rules are heuristics that permit practical resolution of the complexity. Many such rules exist [6, 10, 12, 14]. It is also widely accepted that no single dispatching rule is the “best” [1], that it should be selected on the basis of the operating conditions, and the rule should be even changed dynamically as the operating conditions change [2, 11, 17].


The use of on-line simulation for operating FMSs represents an innovative use of simulation in the area of scheduling. Here, simulation is used primarily to select dispatching rules to be used in the FMS. Use of simulation as a tool to evaluate dispatching rules is not entirely new. Many studies of dispatching rules are based on simulation [12]. The difference is in the suggestion of using data from the actual FMS, on-line, for selection of dispatching rules switching the rules as the operating conditions warrant. Obviously, a large amount of data gathering is required by this approach to initialize the model. Davis and Jones [4] propose concurrent simulation to carry out production scheduling. In their scheme, multiple simulators of a production facility are initialized to the latest state of an FMS. These simulators are stopped after some time. The simulations are then analyzed as terminating simulations to choose the best rule to use.

Synergism between expert systems and simulation is used in an on-line scheduling system called ESS (Expert System Scheduler). Jain et al. [9] describe the development of a scheduling system that communicates on-line with the factory control system, generating schedules in real-time. The scheduling decisions are based on the expertise of an experienced scheduler. The system is based on LISP, and uses object-oriented ideas for both the expert systems and simulation. It is possible to run the simulation backward in time to obtain starting time-windows for jobs.

Wu and Wysk [16, 17] report on a multi-pass expert control system (MPECS) which uses discrete-event
C. BASNET: SELECTION OF DISPATCHING RULES

simulation for on-line control and scheduling in flexible manufacturing systems. In their system, simulation is used to evaluate dispatching rules. An expert system is employed to compile the set of candidate dispatching rules [16]. This expert system has a learning module to learn from past decisions. The expert system generates the candidate set based on current system objectives, system status, and the characteristics of ongoing operations. A 'Flexible Simulation Mechanism' (FSM) collects all the data on the current system status. A simulation model is then generated based on this data. A series of simulation runs is carried out starting from the current state using each of the candidate dispatching rules for the next short time period (dt), selected by the user. FSM provides the performance criteria for each run. The rule that results in the best performance is used to generate a series of commands to the real-time control system of the FMS. The FMS is then run for time dt under the 'best' dispatching rule. Compared to single-pass heuristic scheduling, Wu and Wysk report an improvement of 2.3%-29.3% under different simulation windows (= dt) and measures of performance.

Ishii and Talavage [7] propose a transient based algorithm for determining the length of the simulation window. This is done on the basis of the system state, which is evaluated by a measure similar to the load on the FMS. Strategies are proposed to select the dispatching rule avoiding the problem of censored data with arbitrary simulation windows. Improvements in performance measures of up to 16.5% are reported. Ishii and Talavage also grapple with the issue of determination of length of the simulation runs. Their approach is to use a prior simulation to determine the point in future when there is least load in the system. The look-ahead simulations are then carried out to this point in time using each candidate dispatching rule. While this approach counters the problem of censored data, it adds significant overhead of simulation. Furthermore, the determination of time to least load by one rule does not guarantee that the same time is required for the other rules.

Ramaswamy and Joshi [13] suggest the use of response surface methodology in developing dispatching rules for FMS. In their methodology, offline simulation is used to develop regression weights relating performance measures to significant dispatching factors. These weights are then used with the dispatching factors to sequence jobs in machines.

Very few articles have been published in the area of scheduling based on online simulation [15]. These papers have created interest in this field of investigation, but many questions remain to be answered. The criteria to be used for the selection of dispatching rule, the associated problems of suboptimization, and the issue of censored data in the simulation runs merit further study.

III. RESEARCH PROBLEM

The research described in this paper seeks to contribute to the use of online look-ahead simulation for scheduling. The rationale of a simulation look-ahead lies in the attempt to find a suitable dispatching rule as the situation in an FMS changes dynamically.

The research problem addressed in this paper is: how to select the dispatching rule in simulation-based scheduling. Common measures for performance of scheduling algorithms are sojourn time, lateness, or tardiness. (To avoid confusion in what follows, the criterion used for evaluation of dispatching rules during look-ahead simulations is called a control criterion. The measure used to evaluate the performance of the real FMS is called a performance measure. The reason why they could be different is discussed presently). In the extant literature [7, 17] the criterion used in the selection of dispatching rule from look-ahead simulations is the same as the desired performance measure for the FMS (control criterion is the same as the performance measure). That is, if it is sought to minimize weighted tardiness for an FMS, the particular rule is selected that minimizes weighted tardiness in each selection of dispatching rules via simulation runs. It should be noted that this by no means guarantees minimum weighted tardiness in the long run since each particular look-ahead simulation run will suboptimize. (The fact that a measure is minimized in each of many small time intervals does not guarantee that the measure will be minimized in the total time frame). This is especially true for the very short time windows used to achieve online selection of dispatching rules (this is the censored data problem mentioned earlier). Thus, there is not only the problem of censored data but also of suboptimization.

We propose that global control criteria such as machine utilizations may be used to alleviate the problem of suboptimization. Machine utilization is a global measure in the sense that it is additive over the entire run of the FMS. This cannot be said of the other criteria mentioned above. Hence it can be conjectured that global control criteria may be able to avoid the penalty of suboptimization more than local control criteria. Thus we propose the use of machine utilization as a criterion for the selection of dispatching rule in online look-ahead simulations. Thus the hypothesis to be tested in this research is that in the selection of dispatching rules, the use of a global control criterion, such as utilization, will give better results than local criteria such as flow time or tardiness, in terms of the usual performance measures of sojourn time or tardiness.
IV. SIMULATION EXPERIMENT

A simulation experiment was carried out to test the hypothesis mentioned above. A highly detailed simulator was developed which permits the modeling of the machines, transport devices, fixtures, and tools. These physical components can have different types and exist in limited quantities. It is possible to represent the limited queue capacity of the machines. The machines can have individual distributions of time between failure and time to repair. The events of arrival of work part, mounting on the needed pallet, its movement within the FMS, retrieval of pallet, and eventual exit from the FMS are modeled in complete detail. The machines, in turn, are allowed to fail, to get repaired, to get blocked by work pieces that cannot move, and to get the tools in their tool magazines changed.

A hypothetical FMS cell was modeled, which consisted of 8 numerically controlled machines, a load/unload station, a tool crib, pallets and 4 transport devices (see Figure 2). The input and output queue capacities at each machine were fixed at 2. The processing time for each operation was distributed uniformly from 5 to 25 minutes. The processing time sampling was done at the time of creation of the order by the order generator. This time was assigned to the order and was used throughout the life of the order. The number of machines visited by a part is distributed uniformly from 1 to 4. The number of alternate machines for each operation is distributed uniformly from 1 to 2. Transportation time from a station to another was uniformly distributed from 1.5 to 4 minutes. Each machine was equipped with a tool magazine of capacity 10 slots. The tool needed for a process was randomly assigned from a total of 100 tool types in the tool crib. It was assumed that two copies of each tool are available. A setup time of 30 minutes was incurred in mounting/dismounting of tools for the 8 machines and in transportation from/to the tool crib. One tool occupies one slot, and one operation requires only one tool. Orders for a random (3 to 8) number of parts of a specific part type were generated with the inter-arrival time of 37.2 minutes, distributed exponentially. (This gives a traffic density of 70.34%). Each part type requires one pallet type. There were 50 pallet types, with 5 copies of each type being available. Due dates were assigned to each order on the basis of the total work content. A factor of 5 is used, which gives a tight due date. That is, if an order is generated at current time, due date = current time + total processing time * order size * 5

The machines failed with an exponential distribution of time between failures (500 hours) and uniform distribution of time between repairs (5 hours to 10 hours). When the failure of a machine occurs, the work part under process is not scrapped. Instead, it is assumed to need only the remaining part of its processing time.

Data on two performance measures were gathered: mean sojourn time and mean weighted tardiness. Each order could have two priorities: high and low, high with the probability of 0.1. The tardiness of high priority orders was weighed by a factor of two, while the tardiness of low priority orders was given the weight of one. The length of simulation window (dt) was 120 minutes of simulated time.

A. Dispatching rules

The dispatching rules considered for selection in a time window were the following. These were chosen because of their wide popularity in the dispatching literature.

Slack till due date (SLACK): When using this rule, every time there is a choice of operations to be carried out on a machine, the job with the least slack is given preference. Slack is the time remaining till due date after allowing for processing time, i.e.:

\[ \text{Slack} = \text{Due date} - (\text{Current time} + \text{Remaining process time}) \]

Shortest processing time (SPT): In using this rule, the job with the shortest current operation is picked. However, the version of this rule as implemented here gives preference to tardy jobs. If a job is late for more than 100 minutes, it has a higher priority over jobs that are not as late. Among these late jobs, the priority is by the lowest slack, as defined above.

Earliest due date (DUEDATE): In this rule, a machine always picks the job that is due first.

Maximum operations remaining (MOPR): Here the number of remaining operations on the candidate jobs is compared to choose the job with the maximum remaining operations. Among the jobs with the same number of remaining operations, the priority goes to the first job to arrive at the machine.
B. Experimental factors

The object of the experiments with the simulator was to compare the following procedures for the selection of the dispatching rules to be used in scheduling the FMS:

1. Single-pass scheduling. In these experiments, only one of the above four dispatching rules were used throughout each experiment. That is, there is no simulation based choice of the dispatching rule. A given dispatching rule is used all the time. See Figure 3. These policies could be viewed as the controls in the sense that there is no look-ahead simulation for these.

2. Multi-pass scheduling [17]. In these experiments, simulation was used to select the dispatching rules. Two control criteria were used: mean sojourn time, and mean weighted tardiness (of course, for one replication only one criterion was used). Every time window (dt = 120 minutes), the 'real' FMS was stopped, and the trajectory of the 'real' FMS was extended, by simulation, for another time window (120 minutes). This simulation was done for each of the above dispatching rule. The rule which performed best in terms of the selected control criterion within the simulations was then used for the next time window for the 'real' FMS. It is important to note that the simulation look-ahead uses only a copy of the 'real' FMS to project it ahead using the dispatching rules (deterministically, without the occurrence of stochastic events). Once the selection of dispatching rule is made, the trajectory of the 'real' FMS is continued from the stopped state allowing the occurrence of stochastic events, such as job arrivals, and machine failures, as mentioned previously. This process continued to the end of the replication. Thus the 'real' FMS would use different dispatching rules through the evolution of its operation, the selection of these rules being achieved by simulation performed every time-window (120 minutes). See Figure 3.

3. Transient-based scheduling. In these experiments, the third strategy of Ishii and Talavage [7] was used. This strategy, which showed the best overall performance in their experiments, consists of selecting a single best rule for the entire manufacturing planning horizon (a shift, for example, 480 minutes, in our case) first (see Figure 4). Still using the entire planning horizon for the simulation period, candidate rules are used for the next time window (dt = 120 minutes), and the best rule from the first simulation for the rest of the planning horizon. The candidate rules are selected on the basis of the control criterion achieved at the end of the planning horizon, and is used for the next time window (dt). Again, the two control criteria of mean sojourn time, and mean weighted tardiness were used separately.
C. BASNET: SELECTION OF DISPATCHING RULES

Figure 3 Single-Pass and Multi-Pass Scheduling

Figure 4. Transient-Based Scheduling
4. Utilization-based scheduling. These experiments follow the same process as multipass scheduling, but the control criterion for selection of the dispatching rule is different. In these experiments, a fixed time window of 120 minutes was used, and the dispatching rule that gave the highest total utilization over this period was selected for next 120 minutes.

Twelve replications of each experiment were carried out, each using a set of unique seeds for random numbers. Each simulation run terminated after the attainment of steady state, as evidenced by the number of parts in the system.

C. Results of the experiment

The response variables measured in the experiments were the performance measures of mean sojourn time and mean weighted tardiness. The tardiness of the jobs was weighted, as explained earlier. The independent variables in the experiments were the scheduling policies. Four dispatching rules were used within the single pass policy (SPT, SLACK, MOPR, and DUEDATE). Within the multipass scheduling policy, two control criteria were experimented with: mean sojourn time and mean weighted tardiness. These will be denoted hereafter as MULTI-SOJOURN and MULTI-TARDINESS. The same control criteria of mean sojourn time and mean weighted tardiness were used within the transient based scheduling. These are called TRANSIENT-SOJOURN and TRANSIENT-TARDINESS. The utilization based scheduling is called UTILISATION. Thus a total of 9 scheduling policies were evaluated. The twelve replications of the experiments were analyzed using the analysis of variance (ANOVA) procedure, using the following model.

\[ X_{ij} = \mu + C_i + R_j + e_{ij} \]

where,
- \( X_{ij} \) = \{mean sojourn time, mean weighted tardiness\}. This is the response variable - the performance measure achieved by the FMS.
- \( \mu \) = Average response over all the populations
- \( C_i \) = Effect of the scheduling policy (i = 1, 2, ...9)
- \( R_j \) = Effect of the replication block (j = 1,2, ..12)
- \( e_{ij} \) = Random error for replication j and scheduling policy i.

The results of the ANOVA procedure indicated that there were significant differences between the scheduling policies (both models were found significant with the level of significance of 0.0001). To discover the differences, the Duncan Multiple Range Test for means analysis was carried out at a significance level of 0.05. Figure 5 shows the means analysis for the response variable of mean weighted tardiness and Figure 6 shows the analysis for the response variable of mean sojourn time.

### Duncan Grouping Scheduling Policy

<table>
<thead>
<tr>
<th>Duncan Grouping</th>
<th>Scheduling Policy</th>
<th>Mean Sojourn Time</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUEDATE</td>
<td>1047.13</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>SLACK</td>
<td>1035.01</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>TRANSIENT-TARDINESS</td>
<td>959.95</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>MOPR</td>
<td>954.67</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>TRANSIENT-SOJURN</td>
<td>949.77</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>MULTI-SOJURN</td>
<td>946.01</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>MULTI-TARDINESS</td>
<td>933.40</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>SPT</td>
<td>900.06</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>UTILIZATION</td>
<td>853.42</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 5 Mean Weighted Tardiness Comparisons**

Among the single-pass scheduling rules, the SPT rule performs the best, for both the performance measures of flow time and tardiness. Note that the usual SPT rule was modified in these experiments to give preference to tardy jobs. This scheduling policy compares well against any other scheduling algorithm.
Multipass scheduling performed about the same or better than single-pass scheduling rules. An interesting observation is that the control criterion of sojourn time gives a better performance than the control criterion of tardiness when the performance measure is actually tardiness, and vice versa. That is, when the dispatching rule giving the least tardiness is selected in individual simulation runs, the tardiness in the actual FMS is more than when the dispatching rule giving the least sojourn time is selected. This emphasizes the concern about suboptimization raised earlier.

Transient-based scheduling seems to perform better than the multipass scheduling algorithm for the mean weighted tardiness measure, but not for the mean sojourn time criterion. However, the behavior of the rules is consistent.

Utilization-based scheduling performs best for both the performance measures of sojourn time and tardiness. It appears that both multipass and transient scheduling are not able to overcome the problem of local optima: the selection of one rule in the next period may be the best for that period, but it may worsen the overall situation in the following periods. This is seen dramatically for the multipass scheduling policy. The control criterion of total machine utilization also selects the dispatching rule period by period, but it appears to suffer much less in terms of the performance measures of mean sojourn time and mean weighted tardiness. This lends support to the hypothesis that the use of a global criterion, such as total machine utilization, gives better results than the use of local criteria, such as sojourn time or lateness.

However, although UTILIZATION gives the best performance for both the measures of performance of tardiness and sojourn time, SPT and UTILIZATION do not appear to be different in terms of statistical significance, as seen in the Duncan groupings. This reinforces the superiority of SPT as a dispatching rule that has been seen repeatedly in the scheduling literature [3].

V. Conclusion

Use of simulation for scheduling purposes is gaining increasing acceptance, especially in view of the computational complexity of scheduling problems. Simulation is primarily used in this case to evaluate scheduling alternatives. We experimented with a number of procedures for the selection of dispatching rules in simulation-based scheduling of flexible manufacturing systems. These included single-pass and variations of multi-pass scheduling. Using the performance measures of steady state average sojourn time and average weighted tardiness, it is seen that the criterion of machine utilization performs best among
the evaluated procedures for selection of dispatching rules. Although statistical difference from the popular dispatching rule of shortest processing time was not found, utilization-based scheduling was statistically better in this experimental environment than the other scheduling policies examined. This lends support to the hypothesis that global criteria may be superior to local criteria in the look-ahead simulations used in simulation-based scheduling of FMSs. We worked within a narrow set of dispatching rules, performance measures, and simulation windows. More extensive work is needed in the identification of suitable control criteria, set of dispatching rules to examine, and the duration of the simulation runs. Another avenue of research is to attempt to combine the use of global and local criteria.

REFERENCES