International Journal of Production Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713696255

A new paradigm for rule-based scheduling in the wafer probe centre

T. C. Chiang [a]; Y. S. Shen [a]; L. C. Fu [a]
[a] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

First Published:August2008

To cite this Article Chiang, T. C., Shen, Y. S. and Fu, L. C.(2008)'A new paradigm for rule-based scheduling in the wafer probe centre',International Journal of Production Research,46:15,4111 — 4133

To link to this Article: DOI: 10.1080/00207540601137199

URL: http://dx.doi.org/10.1080/00207540601137199

International Journal of Production Research,
Vol. 46, No. 15, 1 August 2008, 4111–4133

Taylor & Francis
Taylor & Francis Group

A new paradigm for rule-based scheduling in the wafer probe centre

T. C. CHIANG, Y. S. SHEN and L. C. FU*

Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan, 106

(*Revision received August 2005*)

This paper addresses the scheduling problem in the wafer probe centre. The proposed approach is based on the dispatching rule, which is popularly used in the semiconductor manufacturing industry. Instead of designing new rules, this paper proposes a new paradigm to utilize these rules. The proposed paradigm formulates the dispatching process as a 2-D assignment problem with the consideration of information from multiple lots and multiple pieces of equipment in an integrated manner. Then, the dispatching decisions are made by maximizing the gains of multiple possible decisions simultaneously. Besides, we develop a genetic algorithm (GA) for generating good dispatching rules through combining multiple rules with linear weighted summation. The benefits of the proposed paradigm and GA are verified with a comprehensive simulation study on three due-date-based performance measures. The experimental results show that under the proposed paradigm, the dispatching rules and GA can perform much better than under the traditional paradigm.

Keywords: Paradigm; Scheduling; Dispatching rules; Wafer probe

1. Introduction

Semiconductor manufacturing is among the most complicated and capital-intensive manufacturing industries in the world. To survive in the highly competitive market, the wafer-manufacturing industry needs to utilize the resources very well not only to meet internal performance criteria such as cycle time and throughput but also to satisfy customers' requests including delivery date and quantity. Thus, scheduling becomes an inevitable task in this industry (Johri 1992). Wafer manufacturing consists of four stages, and wafer probing is the second stage. Wafers are fabricated in the first stage, and then tested for functionality by means of electrical probes in the probe centre before packaging and final testing. In the literature, much more research has been carried out on scheduling in wafer-fabrication facilities than in the probe centres. However, the performance of the probe centre is found to be critical to the entire wafer-manufacturing process, and scheduling in the probe centre has started to attract the attention of more and more researchers.

Scheduling in the wafer probe centre can be viewed as a job-shop scheduling problem with many complex problem characteristics including dynamic job arrivals,

*Corresponding author. Email: lichen@ntu.edu.tw

International Journal of Production Research
ISSN 0020–7543 print/ISSN 1366–588X online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207540601137199

machine unavailability, sequence-dependent setup times, alternative machines, and batch-type processing. Since the classic job-shop scheduling problem was already proven to be NP hard, scheduling in the wafer probe centre is certainly a difficult problem, thus motivating research on developing effective and efficient scheduling approaches. For a thorough review of system characteristics and scheduling issues in the semiconductor manufacturing industry, see Uzsoy *et al.* (1992, 1994). In the past, most literature about scheduling in the semiconductor manufacturing industry concentrated on cycle time-based objectives since production lots were usually not related to a particular order under the make-to-stock fashion. However, with Application Specific Integrated Circuit (ASIC) and speciality processors gaining more and more market share and production volume, the capability of meeting due dates is becoming a critical factor in the make-to-order environment.

In this work, we address the wafer-probing scheduling problem with three due-date-based objectives including the tardy rate, total tardiness, and maximum tardiness. A survey of previous research works is provided in section 2, in which we will see that the dispatching rule is the most popular approach for scheduling in the semiconductor manufacturing industry. In the literature, researchers devoted themselves to designing a good dispatching rule or generating a combined rule with some optimization algorithms. To our best knowledge, all of them applied the dispatching rules under a typical paradigm. In section 3, we will indicate the drawback of this traditional dispatching paradigm and will also propose a new paradigm to utilize the dispatching rules better. In addition to the dispatching paradigm, we develop a GA to generate appropriate dispatching rules according to the system status and performance criteria. Section 4 will detail the components of the proposed GA. Experiments and results are provided in section 5 to show the benefits of the proposed approach. Conclusions and future research directions are given in section 6.

2. Literature review

In the literature, several approaches have been proposed to solve the scheduling problem in the semiconductor testing facilities. Chen *et al.* (1995) presented a Lagrangian relaxation approach to the IC sort and test facility to minimize the total weighted tardiness. A class of preemptive scheduling problems was addressed in this work. Ovacik and Uzsoy (1996) proposed a decomposition method to minimize the maximum lateness. Their problem considered multiple stages, sequence-dependent setup times, and parallel machines. However, only two kinds of operations are modelled, and uncertainties such as machine breakdown were not taken into account. Yang and Chang (1998) also adopted the Lagrangian relaxation approach and tried to minimize weighted tardiness and flow time simultaneously. Their model contained only one stage and no setup times. Periodic rescheduling was used to cope with the uncertainties. Pearn *et al.* (2002) transformed the wafer-probing scheduling problem into a vehicle-routing problem and solved it with an existing approach. In their later work (Pearn *et al.* 2004), they extended the model from a single stage to multiple stages consisting of both serial and batch stages. Uncertainties, nevertheless, were still not taken into consideration. Ellis *et al.* (2004) provided a survey of more than 10 approaches and indicated that only their work studied the problem characteristics including multiple wafer types, multiple test stations, sequence-dependent setup times, and so on. Even though their problem incorporated many characteristics, they did not model uncertainties and batch stages.

As can be seen, scheduling in the semiconductor manufacturing system must take account of many problem characteristics such as sequence-dependent setup times, uncertainties, various types of operations, etc. Among the scheduling approaches to this kind of problem, dispatching rules could be the most popular in the industry. They are favoured by practitioners because of their ease of implementation, intuitive appeal, flexibility to incorporate domain knowledge, small computational requirements, and convenience of dealing with the dynamic environment. There are many reports from the industry to show successful applications of dispatching rules, such as Appleton-Day and Shao (1997) from AMD, Giegandt and Nicholson (1998) from Siemens, and Ham and Dillard (2005) from Samsung.

With the wide acceptance of dispatching rules in the industry, many researchers in academia have also devoted themselves to developing good dispatching rules. Lu *et al.* (1994) proposed a class of fluctuation smoothing policies to reduce the mean and variance of cycle time. The main idea is to reduce fluctuations such as the burstiness of arrivals to each buffer in the queueing network. Li *et al.* (1996) investigated the minimum inventory variability scheduling policy, also aiming to reduce mean and variance of cycle time. Their scheduling policy tried to reduce variability by introducing the maximum correlation between inter-arrivals and services. Kim *et al.* (1998b) proposed three rules for lot release control, mask scheduling, and batch scheduling to minimize mean flow time and work-in-process inventory and to maximize throughput rate. All these three rules were built based on the concept of load balancing. Hung and Chen (1998) explored a simulation-based dispatch rule and a queue-prediction dispatch rule for achieving the goal of reducing flow times while maintaining a high machine utilization. Their study emphasized that accurate prediction and exploitation of future flow times can improve the schedule performance. Sethi *et al.* (1999) improved the scheduling policy from Lu *et al.* by considering the issues of setup times and batch processing. Yoon and Lee (2000) proposed an operation-due-date rule to reduce the variance of the flow time. In their rule, the operation due date was set to be proportional to the utilization of the capable workstations. Most research works of applying dispatching rules for scheduling in the semiconductor manufacturing systems put the focus on the cycle time (flow time)-related performance measures, except the works by Kim *et al.* (1998a, 2001), whose objectives were to minimize the total tardiness. Due-date information of lots was put into their dispatching rules for lot release, lot scheduling, and batch scheduling. Other interesting research works include Lee *et al.* (2001), Lee *et al.* (2002), Rose (2002, 2003), Chern and Liu (2003), and Chen *et al.* (2005).

Besides devising good dispatching rules, there is another stream of research on automatic and intelligent selecting or combining dispatching rules. For example, Hsieh *et al.* (2001) proposed the ordinal optimization-based approach to select good rules under different initial states, performance indices, and time horizons. To generate a good combined dispatching rule by mixing multiple rules, a popular method is to associate each rule with a weight value and then to generate different combined rules by adjusting the weight values. The rule with a larger weight has a higher impact on the dispatching result obtained by the combined rule. Huang and Lin (1998) developed a human–computer interactive scheduler to adjust the weights of dispatching rules. In their scheduler, many rules were built to relate the criterion to be improved and the rule to be intensified. There are several other approaches to set the weights of rules to generate a combined rule. For instance, Chen *et al.* (2001) used the GA, Min and Yih (2003) took the neural network (NN), and Dabbas *et al.* (2001) and Lin *et al.* (2005) adopted the response surface method (RSM).

3. Proposed paradigm for rule-based scheduling

3.1 *Basic concept*

In the previous section, we have seen that using a dispatching rule is a common approach to solve the lot-scheduling problem in the semiconductor manufacturing industry. However, the typical paradigm for applying the dispatching rules in the existing works has several drawbacks. In this subsection, first we will describe the traditional dispatching paradigm and show its defects. Then, we will describe the basic concept of the dispatching paradigm proposed in this paper.

Under the traditional dispatching paradigm, the dispatching rule is invoked each time when a piece of equipment is released after finishing a testing operation or being repaired. We use figure 1(a) as an example. Eqp1 is released, and there are four waiting lots that it can process. By using the specified dispatching rule in the probe centre, Eqp1 will rank these four lots, and then the lot with the highest priority value will be the next one to be processed. In this way, the production managers and engineers can embed their expertise into the rule and use it to select the most suitable lot according to the status at the decision point. That is the reason why the dispatching rule is flexible and fit for the dynamic and complex environment, and therefore is attractive to the practitioners.

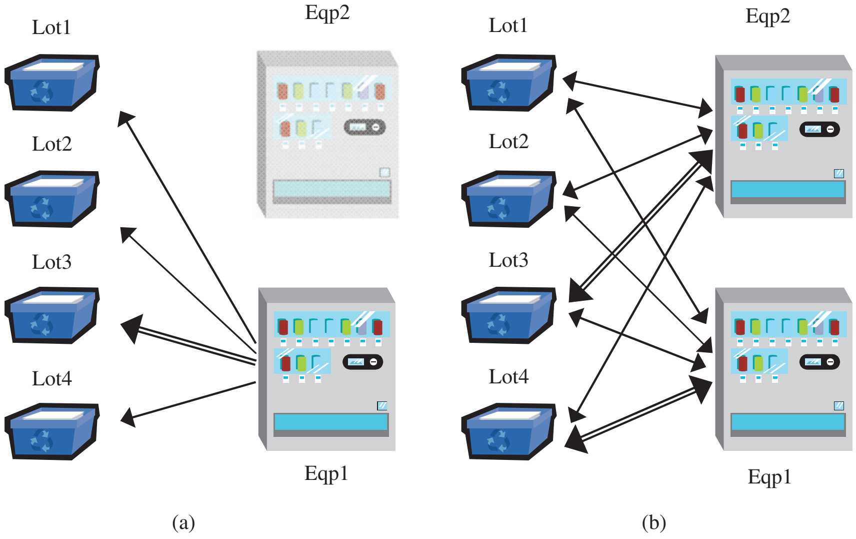


Figure 1. (a) Traditional dispatching paradigm. (b) Proposed dispatching paradigm.

However, there is still some space to make this paradigm perform better. First, the traditional paradigm uses information only from the viewpoint of the equipment initiating the invocation of dispatching rules. In the example in figure 1(a), Eqp1 selects its most preferred lot, say Lot3, regardless of the existence and status of other equipment such as Eqp2, which might be more suitable to process Lot3 than Eqp1. Second, the goal of the traditional paradigm is to find the best match between just one lot and one piece of equipment. But it should be a better way to search for the best set of matches among multiple lots and pieces of equipment simultaneously.

Based on our observations, we propose a dispatching paradigm in order to fix the above defects so that the dispatching rules can be utilized more effectively. The basic concept is illustrated in figure 1(b). Under the proposed paradigm, not only will the equipment rank the waiting lots by the dispatching rules, but the waiting lots will also rank the equipment by the machine selection rules at each decision point. In addition to the equipment that initiates the dispatching process, other pieces of equipment that have the same processing capability may also be involved. In figure 1(b), when there are two pieces of equipment and four lots involved in the dispatching process, a total of eight possible matches will be evaluated, and the matching preference of each possible match is calculated based on the ranking results from both sides of lot and equipment. Then, the best set of two matches is to be found so as to maximize the sum of matching preferences. Mathematically, we formulate the dispatching process as a 2-D assignment problem as follows:

$$\begin{aligned}
&\text{Maximize } \sum_{i\in S_1, j\in S_2} a_{ij}x_{ij}\\
&\text{subject to } \sum_{j\in S_2} x_{ij} = 1,\ \forall\, i = 1, \ldots |S_1|,\\
&\sum_{i\in S_1} x_{ij} \le 1, \quad \forall\, j = 1, \ldots |S_2|,\\
&0 \le x_{ij},\ \forall\, i \in S_1 \text{ and } j \in S_2.
\end{aligned} \tag{1}$$

Given two sets of lots and equipment involved in the dispatching process, S_1 denotes the set with smaller size, and S_2 denotes the other. The variable x_{ij} is 1 if the entities i and j are matched together; otherwise, x_{ij} is 0. The preference of a match of i and j is denoted by a_{ij}, and the goal is to find the set of matches that maximize the sum of matching preferences.

In the following subsections, section 3.2 describes how we choose the lots and equipment to be involved in the dispatching process, namely, how to form S_1 and S_2; section 3.3 details the calculation of matching preferences (a_{ij}); section 3.4 gives a brief description of an existing approach to solve the 2-D assignment problem, and finally section 3.5 presents the linkage of the results of the assignment problem and the dispatching decision.

3.2 *Collection of lot and equipment candidates*

Each time a piece of equipment is released, the first step is to collect the lot and equipment candidates that participate in the dispatching process. Intuitively, all waiting lots that can be processed by this equipment are collected. If there is any piece of idle equipment that has the same processing capability, it is also collected. The non-trivial part of collection is to collect the busy equipment that has the same capability, which we call the 'look-ahead' feature.

To process the waiting lots on the currently idle equipment may not always be the best choice. For example, suppose there are two pieces of equipment, say Eqp1 and Eqp2, and one waiting lot, say Lot1. Eqp1 is idle, and Eqp2 is busy and will finish its current operation after 10 min. Processing of Lot1 on Eqp1 requires 30 min for setup, since the previous operation on Eqp1 is different from the current operation of Lot1. However, processing of Lot1 on Eqp2 needs no setup because Eqp2 is testing the same operation as Lot1. In this situation, it seems more reasonable to let Lot1 wait until the release of Eqp2 than to process Lot1 immediately on Eqp1.

The problem is how to justify if it is worth holding the lots to wait for the busy equipment. In this work, we propose a heuristic to do the judgement. Denote the set of idle equipment by E_I, the set of waiting lots by L, and the processing time and sequence-dependent setup time of lot i on equipment j by p_{ij} and s_{ij}. For each busy equipment k, denote the time to its finishing time of current operation by w_k. Then, a piece of busy equipment k is collected if and only if there exists at least one $i \in L$ such that $p_{ik} + s_{ik} + w_k < \max_{j \in E_I}\{p_{ij} + s_{ij}\}$.

The rationale behind the heuristic is that a piece of busy equipment is worthy of being collected if it is possible to finish the processing of at least one waiting lot earlier than at least one piece of idle equipment.

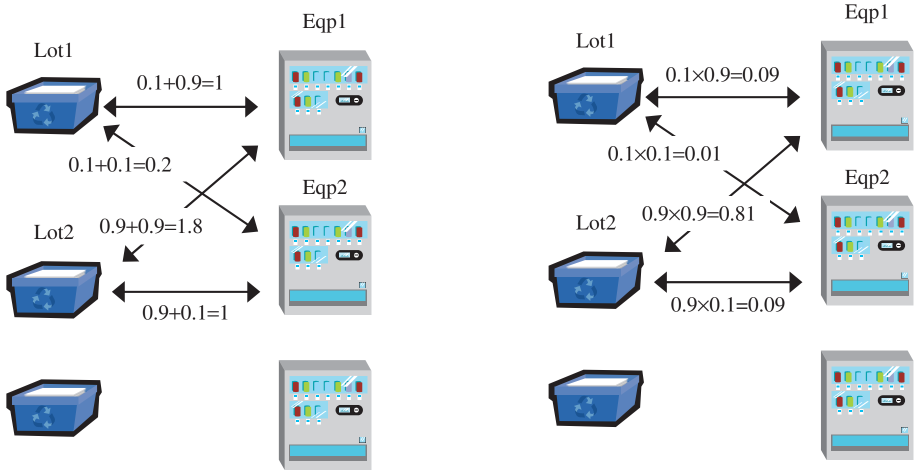
3.3 *Calculation of matching preferences*

After collecting the lot and equipment candidates, the next step is to evaluate all possible matches. The matching preference of the match of lot i and equipment j is composed of two parts, the priority of choosing i from the perspective of j and the priority of choosing j from the perspective of i. These two priority values are calculated by the dispatching rules and machine selection rules. Denote E as the set of equipment candidates, L as the set of lot candidates, $R_j(i)$ as the index value of choosing lot i from the perspective of equipment j by the dispatching rule R, and $R_i^{'}(j)$ as the index value of choosing equipment j from the perspective of lot i by the machine selection rule R'. Then, the matching preference (a_{ij}) of the match of lot i and equipment j is defined as

$$a_{ij} = \frac{R_j(i) - \min_{x \in L}\{R_j(x)\}}{\max_{x \in L}\{R_j(x)\} - \min_{x \in L}\{R_j(x)\}} \cdot \frac{R_i^{\prime}(j) - \min_{y \in E}\{R_i^{\prime}(y)\}}{\max_{y \in E}\{R_i^{'}(y)\} - \min_{y \in E}\{R_i^{'}(y)\}}. \tag{2}$$

Here, we assume that a larger index value means a higher priority. In brief, the matching preference is the product of the normalized index values obtained by the dispatching rule and machine selection rule. The index value is normalized into the same interval [0, 1], since different rules could have different ranges of index values.

Note that the matching preference is the product of two normalized index values, not the sum of them. This is to realize an 'urgent-to-efficient' strategy. Take figure 2 as an example. Suppose we use the sum of the normalized index values as the matching preference, as illustrated in figure 2(a); we can not distinguish the matching results {(Lot1, Eqp1), (Lot2, Eqp2)}, and {(Lot1, Eqp2), (Lot2, Eqp1)}, since the sums of matching preferences associated with these two results are both equal to two.



(a) (b)

Figure 2. (a) Condition using summation to calculate the matching preferences. (b) Condition using product to calculate the matching preferences.

In practice, the lot/equipment with greater normalized index value usually means that it is more urgent/efficient. Thus, we often favour the later matching result, since the urgent lot (Lot2) is processed by the efficient equipment (Eqp1). By using the product of the normalized index values as the matching preference, as illustrated in figure 2(b), we can clearly distinguish the above two matching results because the sum of matching preferences of the former one is 0.18, while that of the later one is 0.82.

There is no restriction on the dispatching rules and machine selection rules to be used. Common examples of dispatching rules include earliest due date (EDD), critical ratio (CR), and least slack (SLACK). Machine selection rules were rarely discussed in the literature. In this paper, we propose to use the SSPT rule, whose definition is given in table 1. For practitioners who want to adopt the proposed dispatching paradigm, it is flexible for them to use the rules that are already implemented in their plant.

3.4 *Solving the 2-D assignment problem*

The next step, after preparing an instance of the 2-D assignment problem, is to find the best set of matches. Here, we adopt a well-known algorithm, the auction algorithm, developed by Bertsekas (1991). The auction algorithm solves the assignment problem by the idea from the auction process in the real world. There are two parties—persons and objects. A benefit is defined for matching a certain person to a certain object, and the goal is to maximize the total benefits under the constraints that one person should be matched to exactly one object, and one object should be matched to at most one person. The auction algorithm is an iterative process, and the persons (bidders) will raise the price of a preferred object by a

Table 1. Descriptions of dispatching rules and machine selection rule.

	Dispatching rules		
Rule	Index value $Z=$	Rule	Priority value $Z=$
FIFO	a	CR	$(d-t)/\mathrm{r}$
EDD	d	CR + SPT	$\max\{(d-t)/r\cdot p,\ p\}$
MDD	$\max\{d,\ t+r\}$	S/RPT + SPT	$\max\{(s/r)\cdot p,\ p\}$
ODD	$s'+p$	COVERT	$(1/p)\cdot(1-s/(k_b\cdot r))^+$
MOD	$\max\{s'+p,\ t+p\}$	ATC	$(1/p)\cdot\exp(-(s'/(k_a\cdot l))^+)$
SLACK	s		
Notations			
a	Lot arrival time	p	Processing time of current stage
d	Lot due date	s'	$d-t-c\cdot(r-p)-p$
t	System time	c	Parameter for calculating s'
r	Remaining processing time	k_b	Parameter of COVERT
s	$d-t-r$	k_a	Parameter of ATC
	Machine selection rules		
Rule	Index value $Z=$		
SSPT	$w+p+s$		
Notations			
p	Processing time of the lot on the equipment	s	Sequence-dependent setup time on the equipment
w	Waiting time (non-zero only if the equipment is busy at the decision point)		

Note: For all rules except COVERT and ATC, the smaller the index value is, the higher the priority is.

bidding increment. The bidding increments and price increases spur competition by making the bidder's own preferred object less attractive to other potential bidders. Bidding from the viewpoint of persons is called the forward auction. By a similar idea, objects can also bid for persons, which is the main idea of the reverse auction. In Bertsekas (1991), a combined forward/reverse auction algorithm was shown to be much more efficient than the forward version. Therefore, we use the combined auction algorithm in our approach. For more details about the auction algorithms, readers are invited to Bertsekas (1991).

3.5 *Linking of assignment result and dispatching decision*

Finally, the assignment problem is solved, and the best set of matches is found. We examine all the matches and pick up the lots that are matched with idle equipment. These lots will be dispatched to their matched equipment and start processing or setup right away. As for the other lots, which are matched with busy equipment, they will be held until they are matched to idle equipment in later invocations of the dispatching process. Note that we do not reserve the busy equipment according to the matching result. For example, suppose the lot Lot1 is matched to the busy equipment Eqp1 after solving the assignment problem; it does not mean that Eqp1 will definitely process Lot1 right after it finishes the current operation. Instead, it will initiate the dispatching process after its current processing, and the most suitable lot to be processed next is determined based on the system status at that time. This strategy can ensure that the decision is made with respect to the latest information in the dynamic environment.

4. Optimization of rules by GA

In the previous section, we propose a paradigm for utilizing the dispatching rules more effectively. However, there is another issue to be solved—which rule is to be used under the paradigm. In the literature, it was often reported that no single rule is dominant in all conditions. For example, the cost over time (COVERT) rule usually performs well on minimizing the total tardiness, while the EDD rule usually provides a good performance on minimizing the maximum tardiness. To combine the advantages of multiple rules, a popular methodology is to mix them by linear weighted summation. For more details of this methodology, see Chen *et al.* (2001), Dabbas *et al.* (2001), and Min and Yih (2003). The weight of each rule refers to its importance in the mixed rule. With different combinations of weights, different mixed rules can be generated. In this work, we develop a GA in order to determine the appropriate weights of rules automatically and intelligently.

GA (Goldberg 1989) is a population-based search algorithm mimicking the natural evolutionary process by artificial genetic operators. It starts with a set of individuals forming the initial population. Each individual in the population is often termed a genome. The genome encodes the solution to the target problem in a form that is convenient for the genetic operations. During the execution of GA, individuals in the population evolve generation by generation toward optimal or near-optimal ones. In each generation, the individuals are evaluated by decoding back to the corresponding solutions and using the fitness function to calculate their fitness values. Based on the fitness values, individuals are processed by mating selection, crossover, mutation, and environmental selection to produce offspring and the new population. The above process repeats until the stopping criterion is reached, and the best individual in the final population is taken as the best solution to the target problem. The following subsections will detail each component in our GA.

4.1 *Genome encoding and decoding*

As mentioned, the goal of our GA is to determine the weights of rules to generate the mixed rule. Hence, the genome is encoded as a string of weight values, as illustrated in figure 3.

There are two kinds of genes in the genome. One kind of gene, g, represents the weight vector of dispatching rules, and the other kind of gene, g', represents the weight vector of machine selection rules. If X dispatching rules and Y machine

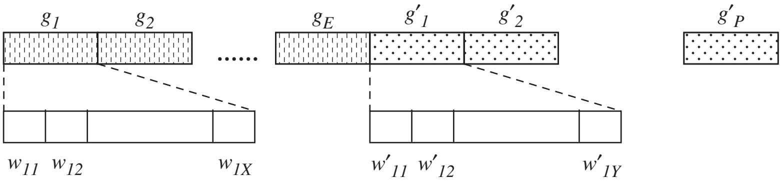


Figure 3. Genome encoding scheme.

selection rules are adopted to generate the mixed rules, genes g and g' are actually a vector of X and Y real numbers, respectively. Given E equipment groups and P product types of lots, the length of genome is $E+P$. In other words, we set for each equipment group and product type a distinct weight vector for mixing rules.

To decode a genome to a solution, a simulator is used. During simulation, each time the dispatching decision is to be made, the mixed dispatching and machine selection rules should be used to calculate the matching preferences (section 3.3) under the proposed paradigm. For example, if we want to calculate the matching preference for a lot of product type 2 and a piece of equipment in equipment group 3, the weight vectors recorded in g_2 and g'_3 will be taken to generate the mixed dispatching and machine selection rules, respectively, and then to calculate the priority values and finally the preference. After simulation, a complete schedule is derived, and then the concerned performance measures can be obtained. These measures will be used in the fitness function to calculate the fitness of a genome.

4.2 *Fitness function*

In our GA, two types of fitness functions are used. The first type is to calculate the sum of the scores of multiple performance measures according to the pre-specified weights, while the second type is to calculate the product of these scores according to the weights. Given K performance measures, these two fitness functions are defined as follows:

$$F_1(G) = \sum_{i=1}^{K} w_i \cdot h_i(f_i(G)).$$

$$F_2(G) = \prod_{i=1}^{K} (h_i(f_i(G)))^{w_i}$$

In the above equations, G denotes the genome to be evaluated, $f_i()$ denotes the ith performance measure, and $h_i()$ is the function to calculate the score with respect to the ith performance measure. Here the function $h_i()$ is a normalization function to convert the performance measure into the range of [0, 1]. Its purpose is to make the selection of the value of w_i more intuitively and reasonably. After normalization, if we assign w_1 with a value greater than w_2, it reflects that the first performance measure is more important than the second measure in the fitness function. In the current implementation, the value of weight w_i associated with each performance measure is open to users so that they can decide the importance of different measures based on their opinions.

4.3 *Mating selection, crossover, and mutation*

After evaluating all individuals with the fitness function, mating selection is responsible for selecting individuals as parents for doing crossover and mutation to produce the offspring. In our GA, the two-tournament selection is used as the mating selection mechanism. By the two-tournament selection, two individuals are picked randomly, and the one with the higher fitness value is selected as the parent. For the crossover operator, we adopt the arithmetic crossover. In brief, each value in the gene of the offspring is the linear weighted summation of the corresponding values of the parents, and the weight is based on the fitness of the parents. An example is given in figure 4. We choose this crossover operator since its performance is better than the two-point crossover in our preliminary experiments. The single-gene substitution mutation serves as our mutation operator. By this operator, a gene is randomly picked and then is replaced by a vector of random real values whose sum is equal to one.

4.4 *Environmental selection*

As we know, ‘survival of the fittest’ is an important principle to be carried out in the GA. The environmental selection mechanism is the component to deal with this issue. In the literature, two kinds of environmental selection mechanisms were usually used. The first the direct replacement mechanism, put the offspring produced after mating selection, crossover, and mutation directly into the next population. The second, called the $n/2n$ mechanism, puts the best n individuals among n individuals in the current population and n produced offspring into the next population. The symbol n denotes the population size. In our experience, the direct replacement mechanism has weak converging power, while the $n/2n$ mechanism loses population diversity quickly. Thus, we propose using another mechanism.

In our mechanism, the best two of two parents and two offspring will replace the parents. Keeping the best two from four individuals, this mechanism has a much greater converging power than the direct replacement mechanism. Besides, replacing

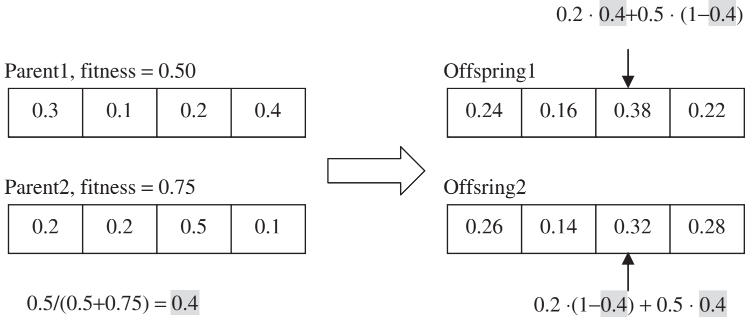


Figure 4. Example of the arithmetic crossover.

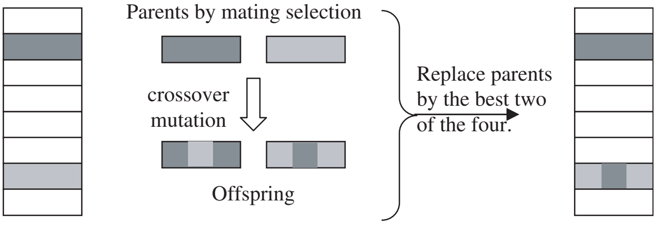


Figure 5. Environmental selection mechanism.

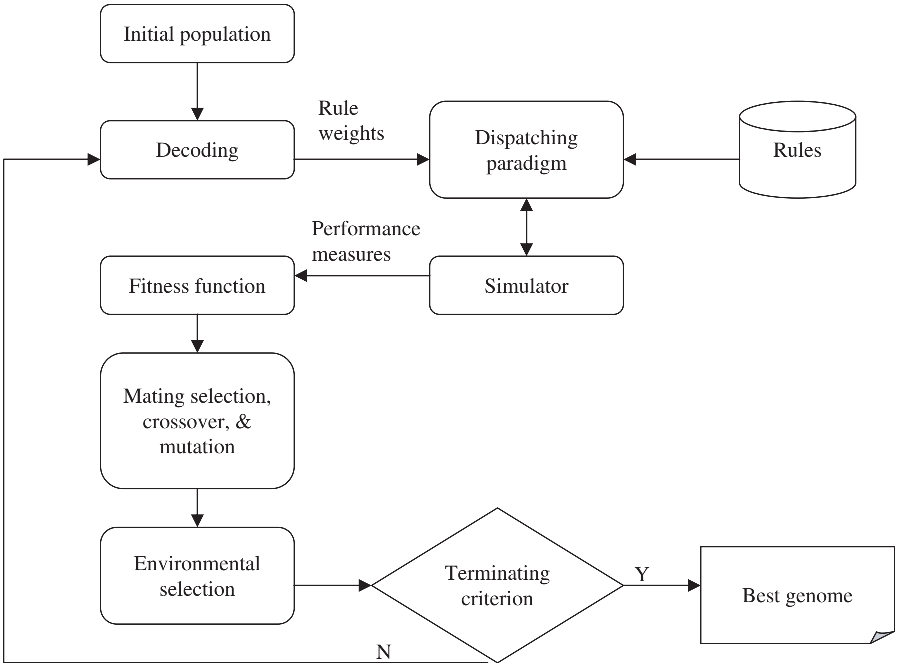


Figure 6. Architecture of the proposed approach.

the parents but not the worst two individuals in the population, we can avoid the rapid loss of population diversity. Figure 5 is a graphical explanation.

4.5 *Initialization and terminating criterion*

Individuals in the initial population are generated randomly. Each gene of an individual is a vector of random real values whose sum is equal to one. In current implementation, the precision of real values is 0.1. The GA terminates when the number of generations reaches a predefined generation number.

In this section, we introduced the proposed GA for generating appropriate mixed rules to be used under the proposed dispatching paradigm. The architecture of the entire proposed approach including the dispatching paradigm and the GA is illustrated in figure 6.

5. Experiments and results

5.1 *System descriptions and experimental settings*

In the tested wafer probe centre, ten types of products are considered. There are seven kinds of operations, and the route of each product consists of two to seven stages. Each operation corresponds to an equipment group, which contains one to three pieces of equipment. There are five operations of serial processing, and the processing times range from 20 to 150 min. The processing times of the other two batch-type operations range from 900 to 1800 min. The sequence dependent setup times range from 10 to 60 min. The processing time depends on the product type and the operation, while the setup time depends on not only the previous two factors but also the product type of the previous processed lot. The mean time between failures (MTBF) and mean time to repair (MTTR) of equipment depend on the equipment types.

The lots are released into the probe centre after they arrive, and the inter-arrival times follow the exponential distribution. The arrival rate is adjusted to keep the equipment utilization to at least 80%. The due date of a lot is set as $t_a + r \cdot P$, where t_a is the arrival time, P is the raw total processing time, and r is a random real value uniformly distributed in the interval [1, 4].

We used the batch means method to collect the simulation output data. The warm-up period was set as one month based on the observations of the curves of average cycle time and WIP level. Eleven batches were collected, and each batch contained data in one month. Hence, each run of experiment simulated for one year. The probe centre was assumed to be empty at the beginning of simulation. In this work, our focus is on the dispatching of stages of serial processing, and we use the first-in-first-out (FIFO) policy for the batch-type stages.

Three due date-based performance measures are considered, including the tardy rate, total tardiness, and maximum tardiness. Denote C_i and D_i as the completion time and due date of a lot i, respectively. For each lot i, U_i is 1 if C_i is greater than D_i; otherwise, U_i is 0. Let L denote the set of lots completed during a batch. The three measures are defined as follows:

$$\text{tardy rate } (T\%) = \frac{1}{|L|}\sum_{i\in L} U_i,$$

$$\text{total tardiness } (\Sigma T) = \sum_{i\in L} \max\{0, C_i - D_i\},$$

$$\text{maximum tardiness } (T_{\max}) = \max_{i\in L}\{\max\{0, C_i - D_i\}\}.$$

The entire system was implemented using C++ language, and the experiments were conducted on a personal computer with a 1-GHz CPU and 256 MB of RAM. Simulation over a year took about 4 s.

5.2 *Experiment 1: effects of dispatching paradigms with different dispatching rules*

In the first experiment, we want to examine the benefit of the proposed dispatching paradigm with 11 popular dispatching rules. The definitions of these rules are summarized in table 1. The machine selection rule is fixed as the SSPT rule.

Table 2. Average tardy rate under different paradigms and dispatching rules in scenario 1.

	P1	P2		P3		P4	
	T%	*T*%	±%	*T*%	±%	*T*%	±%
FIFO	21.87	21.42	−2.1	20.50*	*−6.3*	20.74*	*−5.2*
EDD	14.37	14.12	−1.8	13.87	−3.5	13.97	−2.8
MDD	14.03	14.07	0.3	13.80	−1.7	14.01	−0.1
ODD	14.53	14.14	−2.7	13.81*	*−5.0*	14.32	−1.4
MOD	14.45	13.97	−3.3	13.64*	*−5.6*	13.68*	*−5.3*
SLACK	14.24	14.27	0.2	13.84*	−2.8	14.15	−0.6
CR	14.42	14.15	−1.8	13.93	−3.4	14.19	−1.6
CR + SPT	14.29	14.21	−0.5	13.48*	*−5.6*	14.10	−1.3
SLK/RPT + SPT	13.99	13.71	−2.0	13.45	−3.8	**13.43**	−4.0
COVERT	14.19	13.85	−2.4	13.66	−3.7	13.62*	−4.0
ATC	14.03	13.58	−3.2	13.63	−2.8	13.45*	−4.1

*Performance under P2/P3/P4 is significantly different from that under P1 at significant level 5%.
Bold value signifies the best performance measure in the table.

For the COVERT and ATC rules, which have parameters, we tested ten values from 1 to 10 in increment of 1 for k_a and k_b and picked the best performance measures. Two scenarios were tested. In scenario 1, each equipment group consists of identical equipment, namely, the processing times on all equipment in a group are the same. The assumption of parallel identical equipment is common in the literature. In scenario 2, all pieces of equipment in a group are identical except one piece of faster equipment. The processing time on the faster equipment is 10% shorter than that on the others. In the real-world cases, an equipment group usually contains identical equipment initially. After running the manufacturing system for a while, new (and faster) equipment is often bought to increase manufacturing capacity or to replace the old equipment. Therefore, we used scenario 2 to reflect the practical situation in the industry.

Four dispatching paradigms were tested in the experiments:

- P1: the traditional dispatching paradigm (section 3.1);
- P2: the proposed dispatching paradigm but without the look-ahead feature (section 3.2);
- P3: the proposed dispatching paradigm;
- P4: the proposed dispatching paradigm but using 'summation' when calculating the matching preferences (section 3.3).

Given two scenarios, four dispatching paradigms, and 11 dispatching rules, a total of 88 runs of simulation were conducted. For each simulation run, the average tardy rate was calculated over 11 batches, and so were the average total tardiness and average maximum tardiness. The results are then summarized in tables 2–7. Taking table 2 as an example, the average tardy rate by the FIFO rule under paradigm P1 in scenario 1 is 21.87%. By changing the paradigm from P1 to P2, the average tardy rate is reduced to 21.42%, and the improvement percentage over P1 is 2.1%. In all these tables, values in italics indicate an improvement percentage over 5%. Each value in bold is the best performance measure in the table. In addition, we also conducted the paired *t*-test to check if there is any statistically significant difference

Table 3. Average total tardiness under different paradigms and dispatching rules in scenario 1.

	P1	P2		P3		P4	
	ΣT	ΣT	±%	ΣT	±%	ΣT	±%
FIFO	1508.5	1548.8	2.7	1461.7	−3.1	1434.0	−4.9
EDD	575.1	590.0	2.6	569.6	−0.9	596.1	3.7
MDD	570.9	615.6	7.8	578.5	1.3	587.0	2.8
ODD	618.8	601.5	−2.8	**546.8***	*−11.6*	584.0	*−5.6*
MOD	633.7	595.3	*−6.1*	561.0	*−11.5*	569.8	*−10.1*
SLACK	608.9	613.7	0.8	558.2	*−8.3*	595.2	−2.2
CR	614.2	590.0	−3.9	584.0	−4.9	616.5	0.4
CR + SPT	601.9	645.1	7.2	569.9	*−5.3*	607.3	0.9
SLK/RPT + SPT	621.5	626.9	0.9	584.4	*−6.0*	585.8	*−5.8*
COVERT	595.8	564.6	*−5.2*	552.2	*−7.3*	575.9	−3.3
ATC	582.7	550.6	*−5.5*	554.2	−4.9	568.9	−2.4

*Performance under P2/P3/P4 is significantly different from that under P1 at significant level 5%.
Bold value signifies the best performance measure in the table.

Table 4. Average maximum tardiness under different paradigms and dispatching rules in scenario 1.

	P1	P2		P3		P4	
	$T_{\max}$	$T_{\max}$	±%	$T_{\max}$	±%	$T_{\max}$	±%
FIFO	37.2	36.0	−3.2	33.8	*−9.0*	33.1	*−11.1*
EDD	29.1	28.2	−3.0	31.7	8.7	30.5	4.9
MDD	28.2	37.7	33.8	37.3	32.1	30.9	9.4
ODD	38.0	34.3	*−9.8*	28.4	*−25.3*	33.8	*−11.2*
MOD	37.1	28.0*	*−24.4*	31.8	*−14.2*	28.4*	*−23.5*
SLACK	35.6	30.8	*−13.4*	32.0	*−9.9*	34.6	−2.6
CR	30.8	31.0	0.7	33.2	7.8	30.3	−1.4
CR + SPT	34.6	31.5	*−8.8*	31.8	*−8.1*	37.5	8.6
SLK/RPT + SPT	30.6	36.3	18.6	30.8	0.6	31.3	2.4
COVERT	27.6	27.4	−0.6	27.1	−1.6	31.6	14.4
ATC	**25.3**	27.4	8.2	28.6	12.9	28.3	11.6

*Performance under P2/P3/P4 is significantly different from that under P1 at significant level 5%.
Bold value signifies the best performance measure in the table.

between the performance under the new paradigms (P2, P3, and P4) and under the traditional paradigm (P1). For example, the average tardy rate obtained by using the ODD rule under the paradigm P3 is significantly different from that under the paradigm P1 at a significance level of 5%.

Based on these experimental results, we have the following observations. First, in scenario 1, where each equipment group consists of identical equipment, the proposed paradigm P3 can provide the highest improvement percentage over the traditional paradigm P1. By using P3, the average tardy rates of all 11 dispatching rules become improved, and four of them become improved by more than 5%. The result from a paired *t*-test shows that five rules have a significantly different

Table 5. Average tardy rate under different paradigms and dispatching rules in scenario 2.

	P1	P2		P3		P4	
	T%	T%	±%	T%	±%	T%	±%
FIFO	19.80	19.70	−0.5	19.19	−3.1	18.74*	*−5.4*
EDD	13.55	13.29	−2.0	12.48*	*−7.9*	12.92	−4.7
MDD	13.35	13.23	−0.9	12.57*	*−5.9*	12.65*	*−5.3*
ODD	13.51	13.55	0.3	12.94	−4.2	12.73*	*−5.8*
MOD	13.29	12.73*	−4.2	12.65*	−4.8	12.95	−2.6
SLACK	13.53	13.12	−3.1	12.46*	*−7.9*	12.85*	*−5.0*
CR	13.35	13.29	−0.4	12.80*	−4.2	13.05	−2.3
CR + SPT	13.31	13.28	−0.2	12.67*	−4.8	12.96	−2.7
SLK/RPT + SPT	13.30	13.05	−1.9	12.65*	−4.8	12.53*	*−5.8*
COVERT	13.20	13.01	−1.4	12.40*	*−6.0*	12.53*	*−5.0*
ATC	13.17	12.88	−2.2	**12.38***	*−6.0*	12.46*	*−5.4*

*Performance under P2/P3/P4 is significantly different from that under P1 at significant level 5%. Bold value signifies the best performance measure in the table.

Table 6. Average total tardiness under different paradigms and dispatching rules in scenario 2.

	P1	P2		P3		P4	
	ΣT	ΣT	±%	ΣT	±%	ΣT	±%
FIFO	1313.8	1324.0	0.8	1315.8	0.2	1268.4	−3.5
EDD	569.1	571.9	0.5	493.8	*−13.2*	543.0	−4.6
MDD	543.8	541.8	−0.4	**466.0***	*−14.3*	494.0	*−9.2*
ODD	550.2	532.8	−3.2	542.5	−1.4	560.3	1.8
MOD	534.8	530.2	−0.9	498.3	*−6.8*	567.0	6.0
SLACK	558.5	542.9	−2.8	506.2	*−9.4*	538.9	−3.5
CR	550.4	573.1	4.1	503.2*	*−8.6*	552.4	0.3
CR + SPT	563.5	549.3	−2.5	519.0	*−7.9*	558.7	−0.9
SLK/RPT + SPT	573.0	559.4	−2.4	551.3	−3.8	514.4	*−10.2*
COVERT	520.8	511.2	−1.8	485.1*	*−6.8*	486.3	*−6.6*
ATC	539.4	521.4	−3.3	493.5	*−8.5*	485.7	*−10.0*

*Performance under P2/P3/P4 is significantly different from that under P1 at significant level 5%. Bold value signifies the best performance measure in the table.

performance from their performance under P1. As for the average total tardiness, the performance of ten rules is improved, and six of them have more than 5% improvement by changing from the paradigm P1 to P3. However, only one rule is identified to have a significantly different performance. By observing the total tardiness values of 11 batches obtained by applying the SLACK rule, we found that in some batches its performance is worse under P3 than under P1, although the average performance over all batches is 8.3% better under P3 than under P1. The occasional performance degradation causes a higher variance of the performance difference and therefore makes the performance difference under two paradigms harder to distinguish statistically. Similar observations were made from the data of other rules including CR, CR + SPT, and so on. By using the paradigm P3,

Table 7. Average maximum tardiness under different paradigms and dispatching rules in scenario 2.

	P1	P2		P3		P4	
	$T_{\max}$	$T_{\max}$	±%	$T_{\max}$	±%	$T_{\max}$	±%
FIFO	38.2	35.7	*−6.6*	40.7	6.4	39.2	2.7
EDD	36.1	33.0	*−8.5*	32.1	*−11.0*	33.5	*−7.0*
MDD	32.4	33.1	2.1	**25.4**	*−21.7*	32.0	−1.4
ODD	32.1	27.2	*−15.3*	29.8	*−7.3*	38.8	20.7
MOD	27.3	30.4	11.5	30.3	11.2	41.2	51.1
SLACK	35.2	29.1	*−17.4*	32.2	*−8.6*	33.6	−4.4
CR	32.1	34.7	8.2	30.4	−5.2	38.4	19.5
CR + SPT	37.6	35.9	−4.3	31.7	*−15.6*	30.8	*−18.0*
SLK/RPT + SPT	37.0	32.1	*−13.3*	28.2	*−23.7*	33.1	*−10.4*
COVERT	28.2	27.9	−1.0	27.4	−2.9	29.0	2.6
ATC	27.7	28.1	1.1	27.2	−1.8	28.7	3.5

Bold value signifies the best performance measure in the table.

the maximum tardiness of six rules is reduced, and five of them have more than 5% improvement. However, none of them can be shown to be significantly different under P3 from under P1. The reason is similar to what we have found for the total tardiness.

Regarding the effect of the proposed paradigm on the three concerned performance measures, the proposed paradigm is the most beneficial for the tardy rate, followed by the total tardiness, and finally the maximum tardiness. The relatively smaller benefit for the later two performance measures could be due to the idle time of equipment. Let us use figure 7 as an example. In figure 7(a), the equipment Eqp1 is released and there is a waiting lot Lot1. The equipment Eqp2 is busy now and will be released after 20 time units. The processing times of Lot1 on Eqp1 and Eqp2 are both 60 time units, and the setup times on Eqp1 and Eqp2 are 30 and 0, respectively. With the proposed paradigm, Lot1 will be assigned to Eqp2 since it is found to be finished earlier on Eqp2 than on Eqp1. However, suppose that a lot Lot2, which is at the same step as Lot1, comes after 20 time units. Since it takes too long to wait for Eqp2 (60 time units), Lot2 will be processed on Eqp1. With the due dates of Lot1 and Lo2 as 85 and 100, respectively, the schedule built under the proposed paradigm has a total tardiness of 10. Figure 7(b) shows the schedule built under the traditional paradigm. Lot1 is assigned to be processed on Eqp1, since there is only one piece of idle equipment at the time to dispatch Lot1. This decision might not look so good for Lot1, but the entire schedule considering both Lot1 and Lot2 is better than the schedule in figure 7(a). The total tardiness in the schedule in figure 7(b) is only 5.

The above example shows that under our proposed paradigm, the saving of setup times sometimes accompanies the equipment idle time and consequently the capacity loss. This kind of capacity loss will reflect the occasional performance degradation on the total tardiness and the maximum tardiness. For the total tardiness, the performance degradation is infrequent, since the increment of tardiness of some lots caused by capacity loss is usually compensated by the decrement of tardiness of other lots by the saving of setup times. Therefore, the overall performance is still

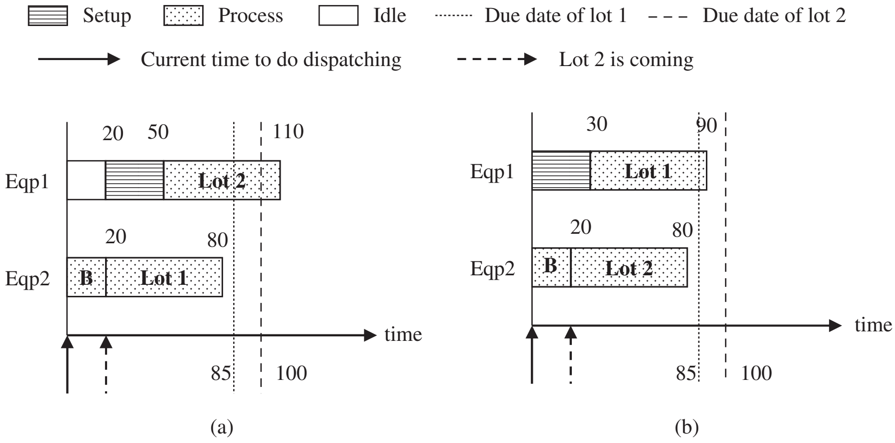


Figure 7. Example of the potential limitation of the proposed paradigm.

improved for most rules under the proposed paradigm. However, the occasional performance degradation will make it difficult to statistically distinguish the performance under the proposed and the traditional paradigm, as we have mentioned earlier. The capacity loss from equipment idle time is more detrimental on the maximum tardiness because the increment of the maximum tardiness cannot be compensated like the total tardiness. Hence, in table 4 we can find that only six rules are improved on the maximum tardiness under the proposed paradigm, and none of them shows a statistically significant performance difference between the traditional and the proposed paradigms. To enhance the proposed paradigm, we can consider the incoming lots and/or use a machine selection rule that involves the equipment idle time. This will be left to future works.

Second, in scenario 2, where there is a piece of faster equipment in each equipment group, the proposed paradigm P3 is still the best paradigm with regard to all concerned performance measures. Under the traditional paradigm P1, the best performance values among 11 rules on the tardy rate, total tardiness, and maximum tardiness are 13.17 (ATC), 520.8 (COVERT), and 27.3 (MOD), respectively. By using the paradigm P3, the best performance values are all improved, with the values as 12.38 (ATC), 466.0 (MDD), and 25.4 (MDD), respectively. The number of rules whose performance is getting improved is 11, 10, and 9 on the average tardy rate, total tardiness, and the maximum tardiness, respectively. In other words, almost all rules are improved. There are 5, 8, and 6 rules with more than 5% improvement on the three performance measures, respectively. Besides, 9 and 3 rules are shown to have significantly different performance under paradigms P1 and P3 on the average tardy rate and the average total tardiness. Similar to what we have observed in the experimental results of scenario 1, the effect of the proposed paradigm on reducing the tardy rate is the most significant, and the effect on reducing the maximum tardiness is the least significant. Comparing the results of scenario 1 and 2, the proposed paradigm is more effective in scenario 2. This larger performance gain is understandable, since in scenario 2, the proposed paradigm can improve the performance by not only saving the setup times but also exploiting the flexibility of a piece of faster equipment in each equipment group.

Third, the paradigm P3 is better than the other two paradigms P2 and P4. This result can verify the benefits of the 'look-ahead' feature and the 'urgent-to-efficient' strategy in the design of our paradigm. One noteworthy observation is that the paradigm P3 is much better than P2 on the tardy rate and total tardiness, but the advantage is less obvious on the maximum tardiness. This observation indicates again that the capacity loss from equipment idling caused by the 'look-ahead' feature should be dealt with properly, especially when the maximum tardiness is concerned.

5.3 *Experiment 2: effects of dispatching paradigms with GA optimization*

After verifying the benefits of the proposed paradigm by using only dispatching rules, the second experiment was conducted to see its benefit with GA optimization. In the experiment, the population size, generation number, and mutation rate are 30, 50, and 0.05, respectively. Based on the result of experiment 1, four dispatching rules that perform the best on the three performance measures are encoded, including MDD, ODD, SLK/RPT + SPT, and ATC. The machine selection rule is fixed as the SSPT rule. Two versions of GA were tested:

- GA-Sum: the fitness function is F_1 in section 4.2. The weights are all equal, with a value of 0.33.
- GA-Product: the fitness function is F_2 in section 4.2. The weights are all equal, with a value of 1.

We ran each version of GA four times. The average performance measures of the best genomes over four runs are summarized in tables 8 and 9. In each table, values in bold are the best performance values for the three concerned measures. In order to see if the performance of GA under the traditional and proposed paradigms is significantly different, the *t*-test was performed. In these tables, we also provide the best performance values obtained by using only the dispatching rules.

Based on the results, we have the following two observations. First, although the GA can improve the performance under either dispatching paradigm P1 or P3, the benefits under P3 are much more significant than under P1. Take scenario 1 for

Table 8. Performance measures under different paradigms and genetic algorithms in scenario 1.

	P1			P3		
	$T\%$	ΣT	$T_{\max}$	$T\%$	ΣT	$T_{\max}$
Rule with the best $T\%$	13.99	621.5	30.6	13.45	584.4	30.8
Rule with the best ΣT	14.03	570.9	28.2	13.81	546.8	28.4
Rule with the best $T_{\max}$	14.03	582.7	25.3	13.45	568.9	28.3
GA-Sum	14.05	555.6	25.9	11.41*	396.3*	**23.5***
GA-Product	14.05	563.2	25.7	**11.40***	**395.7***	24.3*

*Performance under P3 is significantly different from that under P1 at significant level 5%.
Bold value signifies the best performance measure in the table.

Table 9. Performance measures under different paradigms and genetic algorithms in scenario 2.

	P1			P3		
	$T\%$	ΣT	$T_{\max}$	$T\%$	ΣT	$T_{\max}$
Rule with the best $T\%$	13.17	539.4	27.7	12.38	493.5	27.2
Rule with the best ΣT	13.20	520.8	28.2	12.57	466.0	25.4
Rule with the best $T_{\max}$	13.29	534.8	27.3	12.57	466.0	25.4
GA-Sum	13.11	500.6	24.8	10.72*	369.7*	**23.1***
GA-Product	13.09	504.7	26.3	**10.64***	**357.1***	23.2*

*Performance under P3 is significantly different from that under P1 at significant level 5%.
Bold value signifies the best performance measure in the table.

example, the GA can improve the performance by 16% on the tardy rate, 28% on the total tardiness, and 17% on the maximum tardiness under paradigm P3, but it can generate no more than 6% improvement on any performance measure under paradigm P1. Based on the result of the *t*-test, the performance of GA under the proposed paradigm is significantly different from its performance under the traditional paradigm at the significance level 5%. This result is interesting and encouraging. As mentioned in section 2, many research works were done on optimizing the dispatching rules, for example, by combining multiple rules with weights tuned by GA, NN, or RSM. On the one hand, the results show that the paradigm under which the dispatching rules are used also has a significant impact on the performance and deserves more research on this topic. On the other hand, the results also demonstrate the potential of improving the performance of aforementioned approaches by replacing the tradition dispatching paradigm inside them by our proposed paradigm. Second, the performance of two versions of GA is similar. For the practitioners, this result can serve as a reference that either fitness function is good when the importance of multiple performance measures is equal. More experiments are required to explore the difference between these two fitness functions when the importance of performance measures is different.

6. Conclusions and future research

Scheduling in the semiconductor manufacturing industry is often achieved by using the dispatching rules. Previous works in the literature focused on the design of dispatching rules, but none of them noticed the paradigm of using the rules. This paper proposes a new paradigm to utilize the dispatching rules with two features: first, multiple lots and equipment are considered simultaneously when evaluating the possible matches; second, the dispatching decision is made by seeking for the best set of matches, rather than just a single best match. These two features extend the scope of not only the information used in the dispatching process but also the dispatching process itself. In addition to the dispatching paradigm, we also develop a GA to generate the appropriate dispatching rules automatically and intelligently. Benefits of both the proposed paradigm and GA are verified with a model of wafer probe centre on three due-date-based performance measures.

The experimental results show that the proposed dispatching paradigm can improve the performance of the traditional one by $5 \sim 25\%$ when using dispatching rules only. When the proposed GA is applied to enhance the rules, the performance can be improved further by $16 \sim 28\%$, which is higher than under the traditional paradigm by at least 10%.

In the future, there are several directions in which to continue this research:

(1) Besides busy equipment, in-process lots could also be considered in the dispatching process. Just like the proposed heuristic of collecting busy equipment in this paper, we need to figure out a good way to collect the in-process lots.

(2) Under the proposed dispatching paradigm, lots may be held to wait for the busy equipment. Sometimes holding the lots could cause the waste of manufacturing capacity. In future studies, we will need a sophisticated algorithm to justify this kind of waiting.

(3) In the proposed GA, users are given the flexibility to set the weights of performance measures in the fitness function. On the other hand, setting the weights could be a burden for them. Recently, a research stream of multi-objective evolutionary algorithm has been growing rapidly. We will try to apply this kind of technique to deal with the issue of multiple objectives in the next generation of our approach.

Acknowledgement

This research is sponsored by National Science Council under Research Grant with No. NSC 94-2752-E-002-007-PAE.

References

Appleton-Day, K. and Shao, L., Real-time dispatch gets real-time results in AMD's Fab 25, in *Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 1997, pp. 444–447.

Bertsekas, D.P., *Linear Network Optimization: Algorithms and Codes*, 1991 (MIT Press: Cambridge, MA).

Chen, J.-H., Fu, L.-C., Lin, M.-H. and Huang, A.-C., Petri-net and GA-based approach to modeling, scheduling, and performance evaluation for wafer fabrication. *IEEE Trans. Robot. Autom.*, 2001, **17**, 619–636.

Chen, Q., Xi, L. and Wang, Y., The impact of release times, lot size, and scheduling policy in an AT&T facility. *Int. J. Adv. Manuf. Technol.*, 2005, **29**, 577–583.

Chen, T.-R., Chang, T.-S., Chen, C.-W. and Kao, J., Scheduling for IC sort and test with preemptiveness via lagrangian relexation. *IEEE Trans. Syst. Man Cybern.*, 1995, **25**, 1249–1256.

Chern, C.-C. and Liu, Y.-L., Family-based scheduling rules of a sequence-dependent wafer fabrication system. *IEEE Trans. Semicond. Manuf.*, 2003, **16**, 15–25.

Dabbas, R.M., Chen, H.-N., Fowler, J.W. and Shunk, D., A combined dispatching criteria approach to scheduling semiconductor manufacturing systems. *Comput. Ind. Eng.*, 2001, **39**, 307–324.

Ellis, K.P., Lu, Y. and Bish, E.K., Scheduling of wafer test processes in semiconductor manufacturing. *Int. J. Prod. Res.*, 2004, **42**, 215–242.

Giegandt, A. and Nicholson, G., Better dispatch application—a success story, in *Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1998, pp. 396–399.

Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, 1989 (Addison-Wesley: Reading, MA).

Ham, M. and Dillard, F., Dynamic photo stepper dispatching/scheduling in wafer fabrication, in *Proceedings of International Symposium on Semiconductor Manufacturing*, 2005, 75–79.

Hsieh, B.W., Chang, S.C. and Chen, C.H., Scheduling semiconductor wafer fabrication by using ordinal optimization-based simulation. *IEEE Trans. Robot. Autom.*, 2001, **17**, 599–608.

Huang, S.-C. and Lin, J.T., An interactive scheduler for a wafer probe center in semiconductor manufacturing. *Int. J. Prod. Res.*, 1998, **36**, 1883–1900.

Hung, Y.-F. and Chen, I.-R., A simulation study of dispatch rules for reducing flow times in semiconductor wafer fabrication. *Prod. Plan. Control*, 1998, **9**, 714–722.

Johri, P.K., Practical issues in scheduling and dispatching in semiconductor wafer fabrication. *J. Manuf. Syst.*, 1992, **2**, 474–485.

Kim, Y.-D., Kim, J.G., Choi, B. and Kim, H.-U., Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. *IEEE Trans. Robot. Autom.*, 2001, **17**, 589–598.

Kim, Y.-D., Kim, J.-U., Lim, S.-K. and Jun, H.-B., Due-date based scheduling and control policies in a multiproduct semiconductor wafer fabrication facility. *IEEE Trans. Semicond. Manuf.*, 1998a, **11**, 155–164.

Kim, Y.-D., Lee, D.-H. and Kim, J.-U., A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities. *J. Manuf. Syst.*, 1998b, **17**, 107–117.

Lee, L.H., Tang, L.C. and Chan, S.C., Dispatching heuristic for wafer fabrication, in *Proceedings of the 2001 Winter Simulation Conference*, 2001, pp. 1215–1219.

Lee, Y.H., Park, J.W. and Kim, S.Y., Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. *IIE Trans.*, 2002, **34**, 179–190.

Li, S., Tang, T. and Collins, D.W., Minimum inventory variability schedule with applications in semiconductor fabrication. *IEEE Trans. Semicond. Manuf.*, 1996, **9**, 145–149.

Lin, J.T., Wang, F.K. and Kuo, P.C., A parameterized-dispatching rule for a logic IC sort in a wafer fabrication. *Prod. Plan. Control*, 2005, **16**, 426–436.

Lu, S.C.H., Ramaswamy, D. and Kumar, P.R., Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. *IEEE Trans. Semicond. Manuf.*, 1994, **7**, 374–388.

Min, H.-S. and Yih, Y., Selection of dispatching rules on multiple dispatching decision points in real-time scheduling of a semiconductor wafer fabrication system. *Int. J. Prod. Res.*, 2003, **41**, 3921–3941.

Ovacik, I.M. and Uzsoy, R., Decomposition methods for scheduling semiconductor testing facilities. *Int. J. Flex. Manuf. Syst.*, 1996, **8**, 357–388.

Pearn, W.L., Chung, S.H. and Yang, M.H., The wafer probing scheduling problem (WPSP). *J. Oper. Res. Soc.*, 2002, **53**, 864–874.

Pearn, W.L., Chung, S.H., Chen, A.Y. and Yang, M.H., A case study on the multistage IC final testing scheduling problem with reentry. *Int. J. Prod. Econ.*, 2004, **88**, 257–267.

Rose, O., Some issues of the critical ratio dispatch rule in semiconductor manufacturing, in *Proceedings of Winter Simulation Conference*, 2002, pp. 1401–1405.

Rose, O., Comparison of due-date oriented dispatch rules in semiconductor manufacturing, in *Proceedings of Industrial Engineering Research Conference*, 2003.

Sethi, S.P., Sriskandarajah, C., Chu, K.-F. and Yan, H., Efficient setup/dispatching policies in a semiconductor manufacturing facility, in *Proceedings of the 38th Conference on Decision & Control*, 1999, pp. 1368–1373.

Uzsoy, R., Lee, C.-Y. and Martin-Vega, L.A., A review of production planning and scheduling models in the semiconductor industry. Part I: system characteristics, performance evaluation and production planning. *IIE Trans.*, 1992, **24**, 47–58.

Uzsoy, R., Lee, C.-Y. and Martin-Vega, L.A., A review of production planning and scheduling models in the semiconductor industry part II: shop-floor control. *IIE Trans*., 1994, **26**, 44–55.

Yang, J. and Chang, T.-S., Multiobjective scheduling for IC sort and test with a simulation testbed. *IEEE Trans*. *Semicond*. *Manuf*., 1998, **11**, 304–315.

Yoon, H.J. and Lee, D.Y., A control method to reduce the standard deviation of flow time in wafer fabrication. *IEEE Trans*. *Semicond*. *Manuf*., 2000, **13**, 389–392.