Int J Adv Manuf Technol (2003) 22: 75–88
DOI 10.1007/s00170-002-1444-6

ORIGINAL ARTICLE

D. Y. Sha · Chao-Yang Liu

A simulated annealing algorithm for integration of shop floor control strategies in semiconductor wafer fabrication

Received: 15 March 2002 / Accepted: 7 June 2003 / Published online: 14 May 2003

Abstract The semiconductor manufacturing industry is one of the most important industries in Taiwan. Wafer fabrication is an essential process in semiconductor manufacturing. However, controlling the production system on the shop floor is extremely difficult owing to the complicated manufacturing process and reentrant characteristics. In this paper, the shop floor control (SFC) integration strategies (order review/release, dispatching, and rework strategies) in wafer fabrication are considered with using several performances. We reviewed the literature on SFC strategies in wafer fabrication. The proposed combination simulation and simulated annealing (SA) algorithm is presented for SFC strategies in wafer fabrication. The objective was to seek the near global optimum solution for the combination of SFC strategies for a specific performance indicator. From the results, the proposed methodology was found to perform well for combinations of SFC strategies using different performance indicators in wafer fabrication. However, no single combination of SFC strategies could satisfy all performance indicators. Hence, considering the trade-off among these production control strategies, a suitable strategy should be chosen based on the system control tactics. Considerable computational time was saved in this research.

Keywords Shop floor control · Dispatching · Order review and release · Rework · Wafer fabrication · Simulation · Simulated annealing

D. Y. Sha (✉) · C.-Y. Liu
Department of Industrial Engineering and Management,
National Chiao Tung University,
Hsin-Chu, 30050 Taiwan, R.O.C.
E-mail: yjsha@cc.nctu.edu.tw

1 Introduction

Semiconductor manufacturing is a very complex manufacturing process requiring several hundred process steps. A circuit is grown in layers requiring numerous visits to each of the processing areas. The production planning and scheduling problems encountered in the semiconductor manufacturing industry have several features that make them difficult and challenging: random yields, rework, complex product flows, rapidly changing products, and technologies [1]. Because of the above problems, it is more difficult to establish production control in semiconductor manufacturing. Finding a suitable shop floor control (SFC) strategy is very important.

There are four steps in semiconductor manufacturing: wafer fabrication, wafer probe, assembly or packaging, and final testing. A detailed description of the semiconductor manufacturing processes can be found in Uzsoy et al. [1]. This research focuses on the wafer fabrication process in which layers and patterns are built on wafers for the required circuits. Among these four steps, wafer fabrication is the most technologically complex and capital intensive. Because the required capital investment is extremely large, the implementation of an improved SFC strategy could result in a considerable increase in profits. The long flow time, ever-changing products yielded, re-entrant feature of the production sequence, and stochastic aspects of wafer fabrication, particularly machine failures, make developing a SFC strategy for wafer fabrication daunting. There are basically two types of SFC strategies for the wafer fabrication stage. The first and the most familiar strategy is the dispatch strategy. Each time a workstation is ready to commence processing another order, there is a queue of work-in-process (WIP) waiting to be processed at the workstation, at which point the dispatching strategy selects an order or orders to start processing. The second type of SFC strategy, referred to as the order review and release

(ORR) strategy, determines the type, amount, and time point of release for raw wafers to maximize wafer fabrication efficiency. The ORR has been the most important decision strategy in production control. The mean and variance of the workload depend directly upon the ORR strategy in the system, even though the dispatching strategy will impact some performance indicators, such as WIP, flow time, and the throughput rate. In practice, the ORR and dispatching strategies should be complementary to each other. Defective wafers often occur in wafer fabrication, wasting raw materials, and lowering the utilization rate of the manufacturing machines. Due to the high cost involved, finding the best strategy to reprocess defective wafers is critical to the goal of overall manufacturing efficiency. The rework strategy is affected by the photolithography dispatching strategy. Hence, this paper will consider the ORR, dispatching, and rework strategies simultaneously. The purpose of this paper is to find the best combination of SFC strategies for improving the system performance using a simulated annealing (SA) algorithm.

The SA is a step-by-step procedure that can be considered as an improvement of the local optimization algorithm. The algorithm starts with an initial solution and a relatively high temperature value to avoid being prematurely entrapped in a local optimum. This algorithm can search for a nearly global optimum solution. This paper tried to search for a near global optimum solution for a combination of SFC strategies for specific performance indicators such, as WIP, flow time, tardiness, delay cost, and throughput. This algorithm saves a great amount of computational time for this research.

The remainder of this paper is organized as follows. The second section summarizes the relevant literature on SFC strategies. The third section discusses SFC strategy classification and selection. The fourth section describes the simulation model and the proposed methodology. In the fifth section, the results from proposed methodology are presented and the proposed SFC strategies are discussed. Our conclusions and directions for future study are presented in the last section.

2 Literature review

A number of papers have been published on SFC strategies (order review/release, dispatching, and rework strategies) in wafer fabrication. In this section, the literature that addresses the SFC strategies in wafer fabrication and the SA algorithm are discussed.

2.1 SFC strategies in wafer fabrication

Uzsoy et al. [1] found the following factors that made production planning and scheduling in semiconductor industry particularly difficult.

1. Complex product flows
2. Random yield
3. Diverse equipment characteristics
4. Equipment downtime
5. Production and development in shared facilities
6. Data availability and maintenance

Because of the above factors, the many types of wafers that are processed, and the differing machine malfunctions, it is important to exercise reasonable control over the production environment.

Uzsoy et al. [2] pointed out that the researches on SFC in wafer fabrication are focused on order review/release (ORR) and dispatching strategies. There have been considerable researches done on ORR, dispatching, and their interaction. In this paper, three strategies are considered for SFC in wafer fabrication. They are order review/release, dispatching, and rework strategies. In the order review/release strategy it is necessary to determine the time and quantity of wafers for release into wafer fabrication. In the dispatching strategy it is necessary to determine an order or orders to start processing when a workstation is ready to commence processing another order. A queue of work-in-process (WIP) is waiting to be processed at certain workstations. In the rework strategy it is necessary to determine the best method to reprocess defective wafers when defective wafers occur in wafer fabrication. A literature review on order review/release, dispatching, rework strategies, and strategies of integration is presented as follows.

2.1.1 ORR strategy review

Melynk and Ragatz [3] proposed a framework for studying ORR systems. They suggested that release control should be done carefully to reduce the flow time. Furthermore, simulation studies indicated that the wafer release control mechanism has a stronger impact on the system performance than the dispatching strategy does [4-6]. Therefore, most researchers and practitioners in wafer fabrication focus primarily on wafer release control strategies. Lots of researches focused on developing suitable ORR strategies for production control in wafer fabrication, such as workload regulation (WR) [5], starvation avoidance (SA) [4], two-boundary (TB) [7], CONWIP [8], etc.

2.1.2 Dispatching strategy review

Although some researchers pointed out that the ORR strategy has a more significant impact on the system performance than the dispatching strategy, in practice the wafer releasing policy and the dispatching policy must complement each other [9]. Different dispatching strategies may yield different WIP distributions on the shop floor, thereby affecting the utilization of a critical machine and the throughput rate of fabrication. Hence, dispatching strategies have become another

important tool of production control in wafer fabrication. Numerous researches have focused on developing suitable dispatching strategies for production control in wafer fabrication. Glassey and Resende [10] developed a dispatching strategy, SA+, to complement the Starvation Avoidance input regulation policy. In Wein's [5] simulation model, fourteen dispatching strategies were included and altered based on the specific characteristics of wafer fabrication. In Lou and Kager's [7] ORR strategy, TB, the sequencing methodology in front of the bottleneck station is considered. The dispatching strategy, TB+, intends to control both output and WIP at each layer. In the research done by Kim et al. [11], workstations in the fabrication plant were divided into two groups according to their utilization level. These two groups were photo workstations and non-photo workstations such as workstations for deposition, etching, and ion implant. Different dispatching strategies were applied to different groups.

2.1.3 Rework strategy review

During the wafer manufacturing process, defects often occur due to the conditions of the physical environment, human errors, differences between machines, etc. In most production procedures, defective wafers are abandoned because they cannot be reprocessed. However, in the photolithography stage, wafer defects can be redressed. Although redressing defects by reprocessing wafers increases the cycle time and manufacturing cost, it can reduce costs associated with defects. Thus, the wafer rework during the photolithography stage is an important study topic [12]. In plants, wafers are usually processed in lots. The lot is the basic unit of measurement. At the photolithography stage, if defective wafers occur and need to be reworked, the quantity of wafers in each lot ready for processing changes. In each lot, the quantity of defective wafers requiring rework is defined as a child lot. The quantity of wafers that are normal is defined as a mother lot [13].

Zargar [13] mentioned four strategies for processing mother lots and reprocessing child lots. Zarger drew the conclusion that the fourth strategy is the best by comparing the wafer cycle time for each of the four strategies. In the fourth strategy, the mother lot continues to be processed in the next step and the reworked child lot joins the next mother lot to be processed in the next step. As Zargar also mentioned in his research, the third and fourth strategies have not been implemented in industry practice due to the tracing difficulties. Sha et al. [12] tried to find a better strategy for reworking child lots at the photolithography stage to minimize their influence on other normal lots and manage mother and child lots efficiently. They used a simulation to compare rework strategies, including the four strategies mentioned by Zargar [13] and the one proposed in their own research (Rendezvous strategy). From their simulation and statistical analysis results, the Rendezvous strategy was better than any of the four strategies reviewed by Zargar [13].

2.1.4 Strategies of integration review

Numerous researches focused on a single strategy, such as ORR, dispatching, and rework strategies in SFC of wafer fabrication. However, Wein [5] mentioned that the effect of a specific dispatching strategy is highly dependent on both the type of ORR strategy and the number of bottleneck stations in the fab. Sha et al. [12], who concentrated on dispatching and rework strategies, came to the conclusion that the choice of rework strategy should depend on the dispatching strategy and the main performance indicators. Hence, those control strategies have significant interactions. It is very important to consider the interaction between these strategies (ORR, dispatching, and rework strategies) in SFC decision-making. Researches that focused on the interaction between these strategies include Wein [5], Lee et al. [9], Glassey and Resende [10], Kim et al. [11, 18, 19], Lu et al. [14], Fowler [15], Hsieh et al. [16], Chung and Hung [17], and Sha et al. [20], etc. Most of these researches concentrated on the ORR and dispatching strategies and aimed to find either a better ORR strategy or dispatching strategy. Sha et al. [20] considered the rework condition and intended to find a dispatching and rework control strategy combination. This study attempted to consider ORR, dispatching, and rework strategies simultaneously in a simulation model.

2.2 SA algorithm

The origin of SA goes back to 1953, when it was used to simulate a crystal annealing process on a computer [21]. The idea was introduced by Kirkpatrick et al. [22] and applied to various difficult combinatorial optimization problems. SA is a point-by-point method that emulates the annealing process that attempts to force a system to its lowest energy through controlled cooling. Starting from an initial solution at high temperature, the SA generates a new solution x' in the neighborhood of the current solution x. The changes in the objective (energy) function, i.e. $\Delta E = f(x') - f(x)$ are then calculated. In minimization problems, if the $\Delta E < 0$ transition to the new solution is accepted, If $\Delta E \geqq 0$, then the transition to the new solution is accepted with a specified probability obtained by the function $e^{-\Delta E/T}$, where T is a control parameter called temperature and L is a control parameter called the epoch length. The T is gradually decreased by a cooling function.

The typical procedure for implementing a SA algorithm is as follows [23]:

- Step 1. Set an initial solution $x^0 \in S$ (S is the feasible solution spaces).
- Step 2. Set an initial temperature $T^0 > 0$.

- Step 3. Set $x = x^0$ and $T = T^0$.
- Step 4. Repeat steps 5∼10 until frozen.
- Step 5. Repeat steps 6∼9 L times (equilibrium).
- Step 6. Randomly generate solution x', a neighbor of x.
- Step 7. Calculate $\Delta E = f(x') - f(x)$.
- Step 8. Generate a random value r, $0 < r < 1$.
- Step 9. If $\Delta E < 0$, then set $x = x'$,
- Else
- If $r < e^{-\Delta E/T}$, then
- Set $x = x'$.
- Step 10. T = Cooling function (T).

In recent years, much attention and many papers have been devoted to SA, applicable in particular for solving combinational optimization problems. Ponnambalam et al. [24] mentioned many applications for the SA, such as VLSI design, pattern recognition, code generation, TSP problem, graph partitioning, scheduling manufacturing systems, layout design, etc. Many papers used or modified SA to solve job shop sequencing and scheduling problems for a variety of SFC conditions. However, there is no current research using the SA to solve integrated SFC strategy problems (order review/release, dispatching, and rework strategies) in wafer fabrication. This study used the SA to seek the near global optimum solution for a combination of SFC strategies for specific performance indicators in wafer fabrication. A huge amount of computational time was saved using this method.

3 Simulation model

The simulation model developed in this study was based on a wafer fabrication factory in Taiwan. The selected fab has three types of products with a product mix of 0.2, 0.35, and 0.45. The entire process requires 15, 18, and 17 loops. That is, a lot visits photolithographic exposure workstations 15, 18, and 17 times. The model consists of 53 single-server or multi-server workstations and 301 machines. The bottleneck workstation is the photolithographic exposure station. At this station, wafers on which photoresist is deposited are exposed to ultraviolet light through a mask using a glass plate holding a pattern for a single layer of circuits [11]. The transfer time between workstations is ignored in this model. The inter-arrival times for the orders are generated from an exponential distribution with a mean of 40 min. The arrival rate was determined using a preliminary experiment in which the bottleneck workstation utilization was nearly 100%. There are 24 wafers contained in a lot. The processing time for a lot was generated randomly from a uniform distribution between 0.95×MPT and 1.05×MPT, where MPT (mean processing time) is given for each workstation. The setup time is included in the processing time. The simulation model takes into account downtime, which includes unscheduled breakdowns. The MTBF and MTTR values for each workstation were randomly generated from exponential distributions with given mean values. There are four workstations that involve batch processing. The minimum batch size (MBS) rule and first come first served (FCFS) rule, widely used in practice, were used in this simulation model for batching and scheduling in these workstations. In the MBS rule, processing is started when the number of waiting lots at these workstations is greater than or equal to two lots.

This study considered the following SFC strategies, ORR, dispatching, and rework, simultaneously in a simulation model. A classified structure was defined for each kind of SFC strategy and the most feasible strategies were determined using a survey of the researches related to wafer fab SFC. Furthermore, the opinions of both practitioners and researchers were considered in selecting SFC strategies. The selected representative strategies were established in our simulation model.

3.1 Selection of ORR strategy

Bergamaschi et al. [25] introduced a review and classification framework of the main research works on ORR in a job shop. An inherent characteristic that consisted of eight dimensions was identified. Eighteen ORR strategies were used for classification. These dimensions are independent. We attempted to develop a hierarchical classification structure for ORR strategies based on the first three dimensions and determine some representative strategies in each category. The three dimensions used in this study are explained in the following:

1. Order release mechanism: The ORR strategies can be classified into two major types, relating to the mechanism that triggers the release of one or more orders: load-limited methodologies and time-phased methodologies. Under the load-limited methodology, orders are released to the shop based upon their distinctive features and the existing workload in the shop. The time-phased order release approach is centered on computing a release time for each order and then letting orders enter the shop when that predetermined time is reached, regardless of the shop load at that time.
2. Timing convention: The timing convention of an ORR strategy determines when an order release can take place. According to the information provided by the reviewed literature, the timing convention may be either continuous or discrete. Under the continuous timing convention, a release may occur at any time during the system's operation. Under the discrete time convention, an order release procedure may occur only at periodic intervals (e.g. the beginning of each shift, day, or week).
3. Workload measure aggregation: Under load limited order release, the workload can be stated in various levels of aggregation. At one extreme is the total shop

load that gives no indication of the way the load is distributed among the different work centers in the shop. An alternative approach is to compute and control the workload for selected bottleneck work centers only. Another way is to compute the load separately at each work center.

The structure of these three layers and the eight categories used in this research are presented in Fig. 1. A review list of wafer fab ORR strategies is shown in Table 1. The eight ORR strategy categories include LDB, LDS, LDWc, LCS, LCB, LCWc, TD, and TC. The number of citations was used to select the representative strategies. The results in Table 1 show that the ORR strategies with a higher number of citations were selected for this research in every category. The TB, CONWIP, WR, SA, UNIF, and POISSON strategies were selected. No strategy was classified as LDS, LDWc, and LCWc. Through surveying other production systems, Melynk et al. [3] proposed that WCEDD is better than some simple releasing rules in WIP, mean tardiness, and average queue length. These were classified into the LCWc category and included for comparison with the other rules. Therefore, the TB, CONWIP, WR, SA, UNIF, POISSON, and WCEDD strategies were selected in this research.

3.2 Selection of dispatching strategy

The classification of dispatching strategies is based on the structure developed by Blackstone et al. [20]. In their research, dispatching strategies were divided into the following four classes:

1. Strategies involving processing time (PT oriented)
2. Strategies involving due dates (due date oriented)
3. Simple strategies involving neither processing time nor due dates (simple)
4. Strategies involving two or more of the first three classes (combined)

There are some load oriented dispatching strategies presented in SFC researches in wafer fabrication, such as NexQL, SA+, SA, WC, TB, WCEDD, etc. These rules cannot be explicitly classified into the previous categories. Hence, we added a new category to the classification structure to represent the rule manipulation characteristics.

5. Strategies involving the load status (load oriented)

In this research, dispatching strategies were classified into five classes. The classification result is presented in Table 2. Dispatching strategies were selected that had a higher total number of citations in every category. The SRPT, EDD, FIFO, CR, COVERT, SA, TB, and NexQL strategies were selected.

3.3 Selection of rework strategy

The photo station is a critical resource in wafer fabrication. Most of the rework operations are carried out from this workstation. In plants, each lot (one lot = 24 wafers) is subjected to a test after photo station completion. If the lot (mother lot) passes the test, it continues through its process. Otherwise, the defective wafer (child lot) goes to the resist-removal station and then returns to the photo station. Zargar [13] developed four strategies for the rework operation as follows:

1. The mother lot stops and waits until the child lot has been reworked, after which the child lot rejoins the mother lot for continuous joint processing in the next step. In this strategy, to avoid keeping the mother lot waiting too long, the child lot becomes a hot lot,

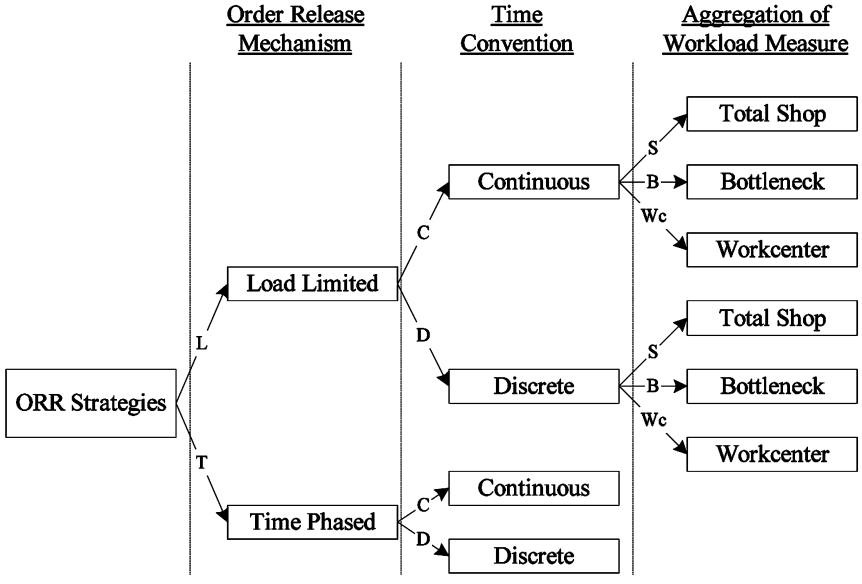

Fig. 1 The ORR strategy attributes

Table 1 The review list of ORR strategies in wafer fab

Type	ORR	Wein [5]	Chung [17]	Kim [11]	Glassey [4]	Lu [14]	Lee [9]	Heieh [16]	Yan [27]	Lou [7]	Glassey [10]	Fowler [15]	Total cites
L-D-B	TB								ˇ	ˇ			2
	WC		ˇ										1
L-D-S													
L-D-Wc													
L-C-S	CONWIP	ˇ	ˇ		ˇ	ˇ	ˇ	ˇ			ˇ	ˇ	8
L-C-B	WR	ˇ		ˇ	ˇ	ˇ		ˇ			ˇ		6
	SA		ˇ		ˇ						ˇ		3
	SA+		ˇ										1
L-C-Wc	WCEDD												
T-D	UNIF				ˇ		ˇ		ˇ	ˇ	ˇ	ˇ	6
	DETERMIN	ˇ				ˇ		ˇ					3
T-C	POISSON	ˇ				ˇ							2

Table 2 The review list of dispatching strategies in wafer fab

Category	authors Rules	Wein [5]	Chung [17]	Kim [11]	Glassey [4]	Lu [14]	Lee [9]	Hsieh [16]	Yan [27]	Lou [7]	Glassey [10]	Fowler [15]	Total cites
PT oriented	LTNV	ˇ				ˇ							2
	STNV	ˇ				ˇ							2
	M1-M2	ˇ				ˇ							2
	SRPT	ˇ			ˇ	ˇ	ˇ				ˇ		5
	SPT						ˇ				ˇ		2
	MOD			ˇ									1
	ATC			ˇ									1
Load oriented	NexQL					ˇ							1
	SA+		ˇ										1
	SA		ˇ								ˇ		2
	WC		ˇ										1
	TB								ˇ	ˇ			2
	W(a,b)	ˇ				ˇ							2
Combined	CR						ˇ					ˇ	2
	FSVCT					ˇ		ˇ					2
	FSMCT					ˇ		ˇ					2
	SLACK			ˇ			ˇ						2
	FIFO+	ˇ											1
	MODEP			ˇ									1
	COVERT			ˇ									1
	LDF							ˇ					1
	OSA							ˇ					1
	MSEC2			ˇ									1
	SRPT+	ˇ											1
Due date oriented	EDD		ˇ	ˇ		ˇ	ˇ					ˇ	5
Simple	FIFO	ˇ	ˇ			ˇ	ˇ		ˇ	ˇ	ˇ	ˇ	8
	CYCLIC	ˇ				ˇ							2
	FGCA	ˇ				ˇ							2
	FGCA/IMP	ˇ											1

which means that it is given processing priority (Lock-step strategy).

2. The mother lot continues processing in the next step and the child lot becomes independent. After the child lot is reworked, it goes on to the next processing step (Lot-split strategy).
3. The mother lot does not wait for the child lot and moves on to the next processing step while the reworked child lot waits in queue until the number of defective child lots becomes great enough for processing together in the next step (Lot-staging strategy).
4. The mother lot continues processing in the next step and the reworked child lot joins the next mother lot in the queue for processing in the next step (Step-mew strategy).

Sha et al. [12] developed a fifth strategy in which the mother lot does not wait for the child lot and instead goes ahead for processing in the next step while the child lot is being reworked. The child lot continues on for processing on its own in the following stage. After this is done, the mother lot and the child lot are rejoined for further processing. In this strategy, the child lot is

designated a hot lot and given priority to be reworked first. These five strategies were named lock-step, lot-split, lot-staging, step-mew, and rendezvous strategies by Sha et al. [22].

From the above detailed description, there are five rework strategies designed for wafer fabrication, i.e. the lock-step, lot-spit, lot-staging, step-mew, and rendezvous strategies. In the lot-staging and step-mew strategies, wafers are mixed, which makes it difficult to trace them. As a result, these two strategies have never been implemented in industry practice. Therefore, these two strategies will be not included. The lock-step, lot-spit, and rendezvous strategies were selected in this research.

4 Proposed methodology

To save considerable computational time, the best combination of SFC strategies was determined using the proposed methodology. This method will be presented in this section. This optimization methodology is a simulated annealing-based (named SA-based) methodology. The SA-based methodology concept, SFC strategy coding, and SA-based procedure are presented as follows.

4.1 SA-based methodology concept

The tools used in this concept are simulation and the SA algorithm. This research used fab data to establish a simulation module (virtual wafer fabrication). The data were then modified to establish the SA module. These two modules were linked to seek a near global solution for the SFC strategy combinations. The SA-based methodology concept is presented in Fig. 2. From Fig. 2 the SA module passed an initial solution for the combination of SFC strategies to the simulation module. The simulation module then executed virtual wafer fabrication. The various performance indicators from the combination of SFC strategies is obtained, measured against the various performance indicators, and passed on to the SA module. The various performance indicators are then evaluated as to whether or not the given SFC strategy is frozen by the SA module. If frozen, the SA module stops, and this combination of SFC strategies is a near optimum solution. Otherwise the SA module generates a new solution for the combination of SFC strategies and repeats the step until the algorithm is frozen. This method quickly seeks the near global optimum solution for a given combination of SFC strategies based on several specific wafer fabrication performance indicators. This process assures that huge computational time is saved.

4.2 Coding of SFC strategies

Three kinds of strategies selected in this research will be discussed. These strategies are order review/release, dispatching, and rework strategies. These strategies were established in our simulation module and must be given a code to link each strategy to the SA module. This code must be used when the SA module receives the combination of SFC strategies. The SA module can generate a combination code for new SFC strategy combination solutions and pass this combination code to the simulation module for executing virtual wafer fabrication. The codes for three kinds of strategies are presented in Tables 3, 4 and 5.

4.3 Parameter setting for the SA-based methodology

The simulation module compares the integrated strategies using a series of simulation experiments. The performance measures used for the comparison are work in process (WIP), flow time (FT), tardiness, percent of tardy jobs (tardy rate), delay cost, and throughput (by days). These parameters are used in the mass production environment. The due dates are determined internally as follows:

$$d_i = a_i + 4.4 * TPT \tag{1}$$

- d_i: *due date of lot i.* a_i: *arrival time*
- *TPT*: *total processing time*
- The function of the delay cost is defined as follows:

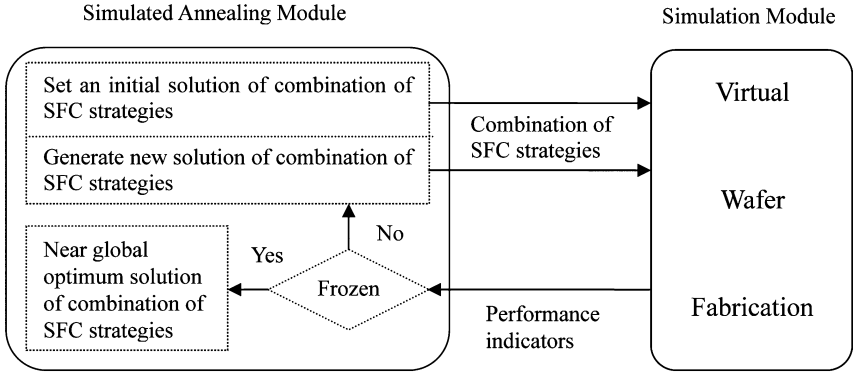

Fig. 2 SA-based methodology concept

Table 3 The coding of ORR strategy

ORR strategy	Description	Code
WR	Regulating new wafer releases to maintain a constant amount of expected work at a bottleneck station.	1
CONWIP	Regulating new wafer releases to maintain a constant number of lots in the production system.	2
SA	A new wafer lot is released to avoid starvation at a bottleneck workstation.	3
UNIF	Release a new lot into the fab at a constant rate, e.g. 16 lot/day.	4
POISSON	New wafer release time is randomly generated from a Poisson distribution.	5
TB	If the actual first layer output < the expected output and actual first layer inventory < predetermined inventory level, the job will be a candidate. The candidates that have the largest weight of product × difference in output value will be released.	6
WCEDD	Regulating new wafer releases to maintain their predetermined WIP level at the main workstations.	7

Table 4 The coding of dispatching strategy

Dispatching strategy	Description	Code
FIFO	First in first out.	1
EDD	Earliest due date.	2
CR	Smallest CR, CR = (due date-total remaining PT-present date)/ total remaining PT.	3
NexQL	The lot whose queue at the next station it will visit has the least amount of expected work per machine will be assigned high priority.	4
SRPT	Shortest total remaining processing time.	5
COVERT	Largest C, C = delaying cost/total remaining PT.	6
SA^+	Assign high priority to jobs that are close to the bottleneck station and/or that contribute a large amount of work content to the station.	7
TB^+	If the actual output of each layer < the expected output and actual inventory of each layer < predetermined inventory level, the job will be a candidate. The candidates that have the largest weight of product × weight of layers × difference in output values will be assigned high priority.	8

Table 5 The coding of rework strategy

Rework strategy	Description	Code
Lock-step	Mother lot stops and the mother lot waits until the child lot has been reworked, after which the child lot rejoins the mother lot for continuous joint processing in the next step.	1
Lot-spit	Mother lot continues to be processed in the next step and the child lot becomes independent. After the child lot is reworked, it goes on to the next processing step.	2
Rendezvous	Mother lot does not wait for the child lot, but instead goes ahead for processing in the next step while the child lot is being reworked. The child lot continues on for processing on its own in the following stage. After this is done, the mother lot and the child lot are rejoined for further processing.	3

$$x = finish\ date\ of\ lot\ i - due\ date\ of\ lot\ i$$

$$delay\ \cos t = \begin{cases} 0 & if\ -1<x<1 \\ -x & if\ x \le -1 \\ x^2 & if\ x \ge 1 \end{cases} \quad (2)$$

The SA module begins by choosing an initial feasible solution (current solution), and the objective function is calculated. A new solution (neighbor solution) is then randomly generated, and the objective function is calculated again. The change in the objective function is calculated and evaluated. These procedures continue until the stopping criterions are met, which terminates the procedure. Four important factors that affect SA performance were addressed:

1. The initial temperature, T^0
2. The cooling rate, α, to determine how the temperature is to be changed

3. The number of iterations, L, to be performed at each temperature
4. A stopping criterion to terminate the algorithm

The initial temperature T^0 should be raised to a sufficient level to allow the algorithm to accept a high percentage of new solutions. Starting at a high T^0 allows the algorithm to avoid the local optimum. However, a high T^0 might increase the computational time. Hence, the T^0 should be selected so that the probability of acceptance of a new solution will be sufficient. Kirkpatrick et al. [24] stated that the T^0 value should be selected to ensure that at least 80% of the new solutions are accepted. That can be described mathematically in the form $e^{-\Delta E/T} \geqq 0.8$, where ΔE is the percentage of difference in the objective function and T is the initial temperature. Therefore, an initial temperature of 90 was chosen in this research. The SA module was designed to accept a portion of inferior solutions. Some pilot runs showed that the accepted inferior solutions were, on average, 20% relative to the original solution with an associated probability of acceptance of 0.8. The initial temperature setting is given below:

Probability of acceptance = exp (–ΔE/ T),
i.e. 0.8 = exp (–20/ T), or T = 90

The cooling rate α represents the rate at which the temperature is decreased. This parameter should be decreased slowly enough to allow the algorithm to search a wide area and accept a relatively large number of new solutions. The temperature is usually set at a range of 0.8~0.9 to ensure a gradual reduction in temperature. However, a high α might increase the computational time. Hence, an α of 0.8 was set in this research.

The number of iterations L at each temperature level, named the epoch length (L), was determined by multiplying the number of strategies (N_s) in the system by a constant called the size factor (c). As the number of iterations increases, more new solutions are generated. This will increase the possibility of finding the global solution. However, it also increases the run time. Therefore, a c of 1 was set in this research. L was determined using $N_s \times c$ is $3 \times 1 = 3$.

The stopping criterion refers to the parameters used to terminate the search process. Various stopping criteria have been suggested in the literature. However, these stopping criteria may not allow the annealing process to continue until the system is frozen. If the frozen level has been set high, the run time also increases. Therefore, SA module freezing was achieved using two criteria in this research. When the freeze counter reaches 5 times, i.e. when the performance indicators have not been improved after decreasing the temperature 5 times, the process freezes. When the temperature T falls down to 0.1114, the SA module will run for 30 'temperature stages', i.e. the final temperature will be $90 \times (0.8)^{30} = 0.1114$, which causes the SA module to freeze.

4.4 Procedure of SA-based methodology

This section presents SA-based methodology procedures for linking the simulation and SA modules directly applied to solving SFC wafer fabrication problems. The SA module can be used to obtain the optimal combination code for the control strategy from the possible solution spaces. Here, a vector of strategy codes represents the possible solution. For instance, a SFC system has three strategies A (ORR), B (dispatching), and C (rework). A vector (5, 7, 2) can represent the codes for the three strategies (A, B, and C), respectively. The procedure for generating a neighborhood solution is to select randomly any strategy code and then randomly assign another setting [23]. For instance, the neighborhood solution for vector (5, 7, 2) involves randomly selecting a strategy (say strategy B), and then randomly assigning another setting (say code 6) to replace code 7. In this instance, the neighbor of (5, 7, 2) is set as (5, 6, 2). The procedure for generating a neighborhood solution is different with the typical SA.

When the typical SA for generating a neighborhood solution generates an inferior solution, the procedure accepts this inferior solution with an associated probability for acceptance of $e^{-\Delta E/T}$. The procedure must accept a certain percentage of inferior solutions to advance toward the near global optimum solution. Then the procedure maybe missed the near global optimum solution, and the near global optimum solution may not be found again by the typical SA. Hence, this research used the variable x_g to record the near global optimum solution. This variable prevents the system from missing the near global optimum solution. Figure 3 depicts the SA-based methodology procedure. The detailed procedure is summarized as follows:

- Step 1. Set an initial temperature $T^0 > 0$.
- Step 2. Set the epoch length L, and the cooling factor α, $0 < \alpha < 1$.
- Step 3. Create an initial solution $x^0 \in S$ (S is the feasible solution spaces) by randomly selecting the combination code for the SFC strategy combination. Pass the combination code to the simulation module to execute virtual wafer fabrication. Six replications (runs) of the simulation are then executed, followed by obtaining the average of the performance indicators. This data is then passed to the SA module. These performance indicators are $f(x^o)$.
- Step 4. Set $x_l = x^0$, $x_g = x^0$ and $T = T^0$, where x_l is a near local optimum solution and x_g is a near global optimum solution.
- Step 5. Repeat steps 6~13 until frozen.
- Step 6. Repeat steps 7~12 L times (equilibrium).
- Step 7. Randomly generate solution x' using the procedure for generating a neighborhood solution, a neighbor of x_l.
- Step 8. Pass the combination code to the simulation module to execute virtual wafer fabrication. Execute 6 replications (runs) of the simulation. Obtain the

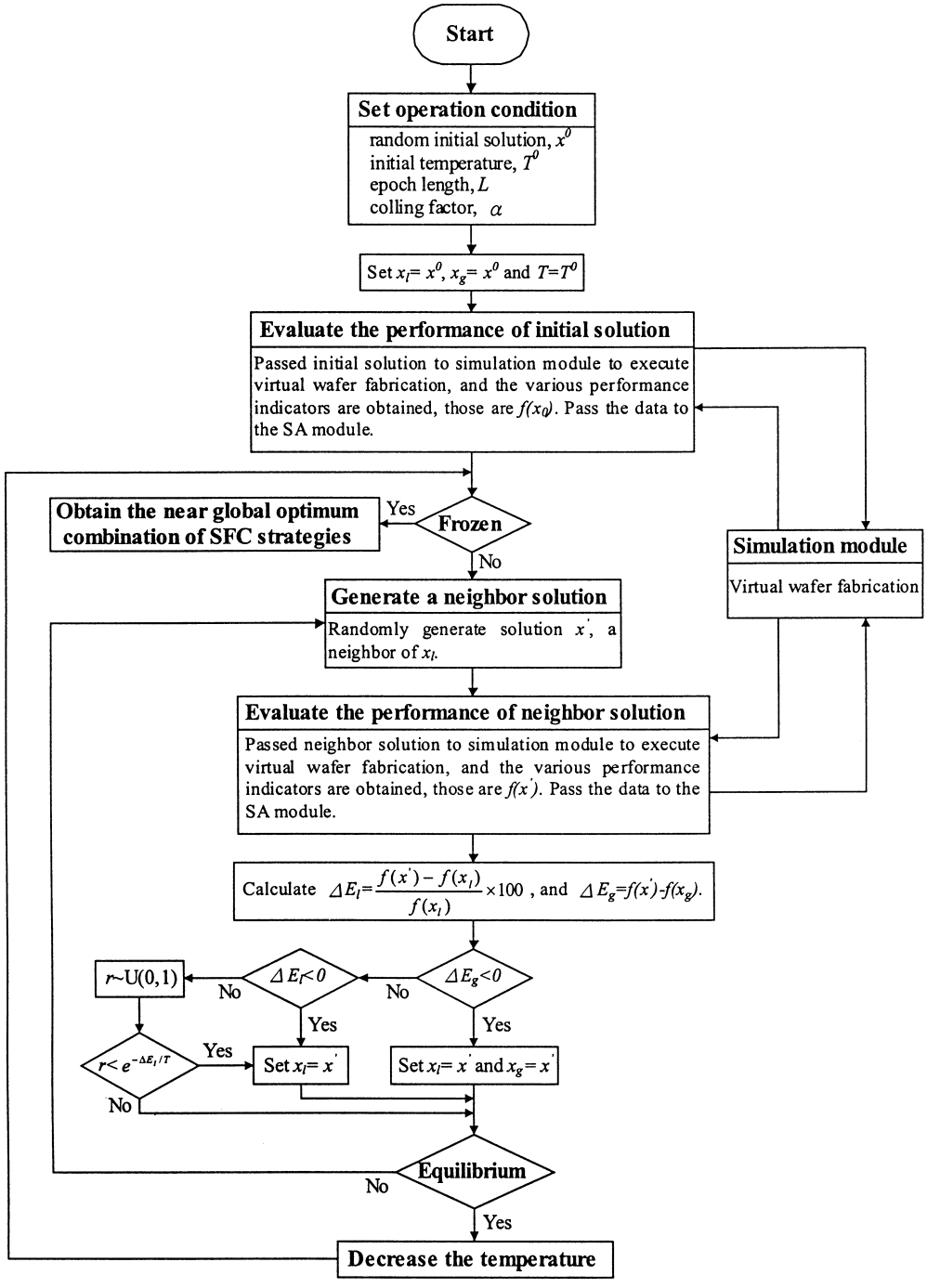

Fig. 3 Procedure for the SA-based methodology

average of the performance indicators. Pass the data to the SA module. These performance indicators are $f(x')$.

- Step 9. Calculate $\Delta E_l = \frac{f(x') - f(x_l)}{f(x_l)} \times 100$, and $\Delta E_g = f(x') - f(x_g)$.
- Step 10. Generate a random value r, $0 < r < 1$.
- Step 11. If $\Delta E_g < 0$, then set $x_l = x'$ and $x_g = x'$,
- else
- if $\Delta E_l < 0$, then set $x_l = x'$
- else
- if $r < e^{-\Delta E_l / T}$, then
- set $x_l = x'$.
- Step 12. Call the current combination code settings the optimal condition.
- Step 13. Set $T = \alpha T$.
- Step 14. Obtained the near global optimum solution for the combination of SFC strategies using the optimal combination code, x_g.

5 Implementation and numerical illustrations

In the simulation experiments, 6 simulation replications (runs) in a steady state were executed for each and every

combination of seven ORR strategies (WR, CONWIP, SA, UNIF, POISSON, TB, WCEDD), eight dispatching strategies (FIFO, EDD, CR, NexQL, SRPT, COVERT, SA+, TB+), and three rework strategies (Lock-step, Lot-spit, Rendezvous). The average values for the performance indicators can be obtained using the simulation. Each simulation run was designed for a simulation time period of 24 hours a day. The simulation would end when 3000 lots were finished after 150 warm days. Different random number seeds were used for the six runs. Each run started with the production area empty.

The values of the required parameters for each strategy were determined using a series of preliminary simulation tests on several candidate values. The selected parameter values are those that give the smallest WIP inventory when (almost) the highest throughput rate and bottleneck utility are obtained. The SA-based methodology was performed on personal computers with Pentium III 1G processors using eM-plant, a simulation package developed by Tecnomatix Technologies Corp. The numerical illustration procedure used the SA-based methodology to solve the WIP indicator example presented in the following:

- Step 1. Set $T^0 = 90$.
- Step 2. Set $L = 3$, and $\alpha = 0.8$.
- Step 3. Create $x^0 = 111$ (the combination of SFC strategies is WR*FIFO*Lock-step) by randomly selecting the combination code for the combination of SFC strategies. Pass the 111 combination code to the simulation module to execute 6 virtual wafer fabrication runs. The average WIP is obtained. Pass the data to the SA module. The $f(x^o) = 675.47$.
- Step 4. Set $x_l = x^0$, $x_g = x^0$ and $T = T^0$, then $x_l = 111$, $x_g = 111$ and $T = 90$, $f(x_l) = 675.47$, $f(x_g) = 675.47$.
- Step 5. Repeat steps 6∼13 until the simulation is frozen (when the performance indicators have not been improved after decreasing the temperature 5 times or when the temperature T falls down to 0.1114).
- Step 6. Repeat steps 7∼12 3 times (equilibrium).
- Step 7. Randomly generate solution x', a neighbor of x_l. The top six combinations of SFC strategies for this neighborhood are shown in Table 6.
- Step 8. Pass the combination code to the simulation module to execute 6 virtual wafer fabrication runs. Obtain the average WIP. Pass the data to the SA module. These $f(x')$ are also shown in Table 6.
- Step 9. Calculate ΔE_l and ΔE_g. The first
- $\Delta E_l = \frac{844.58 - 675.47}{675.47} \times 100 = 25.04$, and
- $\Delta E_g = 844.58 - 675.47 = 169.11$
- Step 10. Generate a random value r, shown in Table 6.
- Step 11. If $\Delta E_g < 0$, then set $x_l = x'$ and $x_g = x'$,
- else
- if $\Delta E_l < 0$, then set $x_l = x'$
- else
- if $r < e^{-\Delta E_l / T}$, then
- set $x_l = x'$, as shown in Table 6.
- Step 12. Call the current combination code settings the optimal condition. The convergence process for the near local optimum solution for WIP is presented in Fig. 4.
- Step 13. Set $T = \alpha T$, as shown in Table 6.
- Step 14. Obtained the near global optimum solution of combination SFC strategies by optimal combination code that is x_g, as shown in Fig. 5.

Table 6 The annealing process numerical changes

T	x_l	$f(x_l)$	x_g	$f(x_g)$	r	$e^{-\Delta E_l / T}$
90.00	111	675.47	111	675.47	0.00	2.82
90.00	511	844.58	111	675.47	0.64	0.76
90.00	571	814.91	111	675.47	0.00	1.04
72.00	271	740.52	111	675.47	0.00	1.14
72.00	771	693.01	111	675.47	0.00	1.09
72.00	671	410.58	671	410.58	0.00	1.76

The numerical illustration shows that the WIP near local optimum solution is 622, but the WIP near global optimum solution is 652. The near global optimum solution for the combination of SFC strategies for the WIP performance indicator is TB, SRPT, Lot-spit

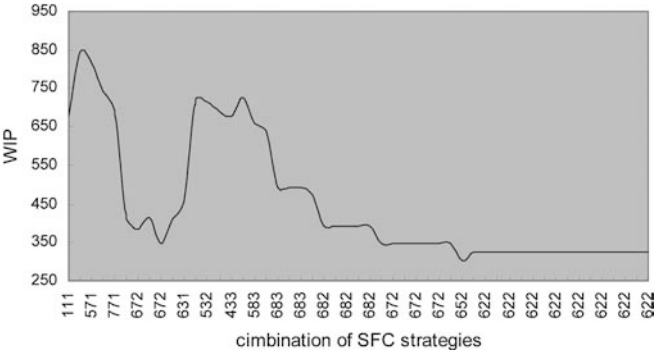

Fig. 4 The convergence process for the WIP near local optimum solution

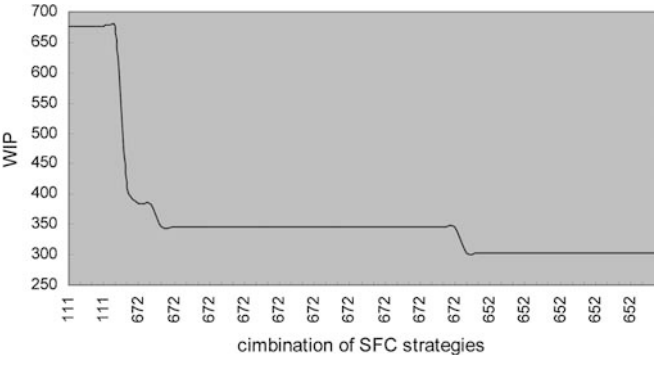

Fig. 5 The WIP near global optimum solution process evolution

(652). The WIP is 303.29. The SA-based methodology prevented missing the near global optimal solution (652). The convergence process for the near local optimal solution for FT, tardiness, tardy rate, delay cost, and throughput are presented in Figs. 6, 7, 8, 9, and 10, respectively.

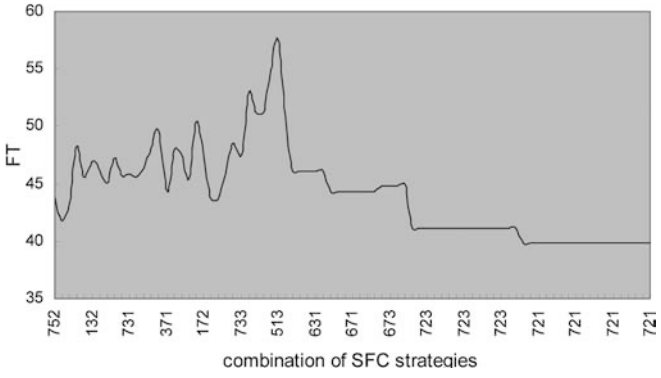

Fig. 6 The convergence process for the FT near local optimum solution

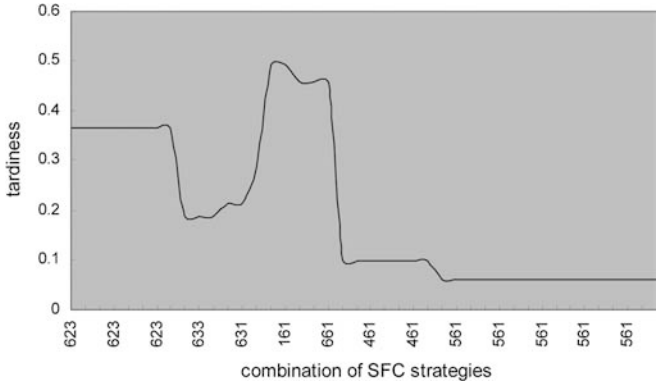

Fig. 7 The convergence process for the tardiness near local optimum solution

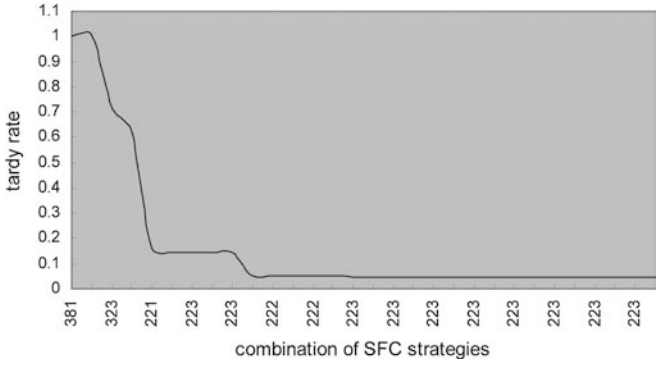

Fig. 8 The convergence process for the tardy rate near local optimum solution

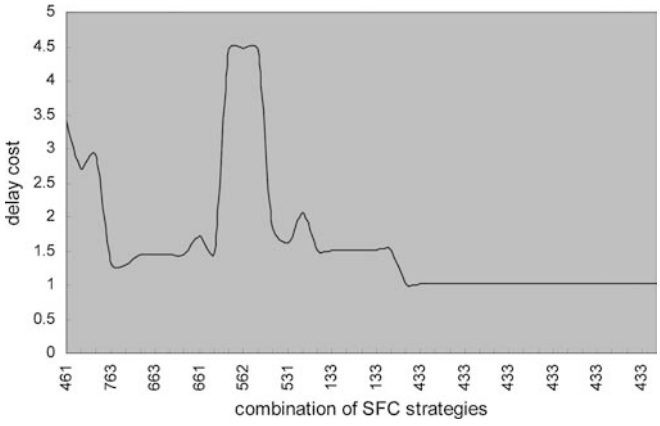

Fig. 9 The convergence process for the delay cost near local optimum solution

As Figs. 6, 7, 8, 9, and 10 have shown, every performance indicator can converge to a near local optimum solution (combination of SFC strategies) by SA-based methodology. Furthermore, if the local optimum solution is the best in the convergence process, the local optimum solution can be as good as the global optimum solution. The SA-based methodology therefore seeks a well-performed combination of SFC strategies for every performance indicator. From the observations of Figs. 6, 7, 8, 9, and 10, we can see that the local optimum solutions are best in the convergence processes, except Fig. 10. The local optimum solution is not a global optimum solution for the throughput indicator. Besides, the objects of FT, tardiness, tardy rate, and delay cost are minimal, while throughput is maximal.

5.1 SA-based methodology results and discussion

The SFC strategy combination results from using the SA-based methodology are summarized in Table 7. The performance indicators for WIP, FT, tardiness, tardy

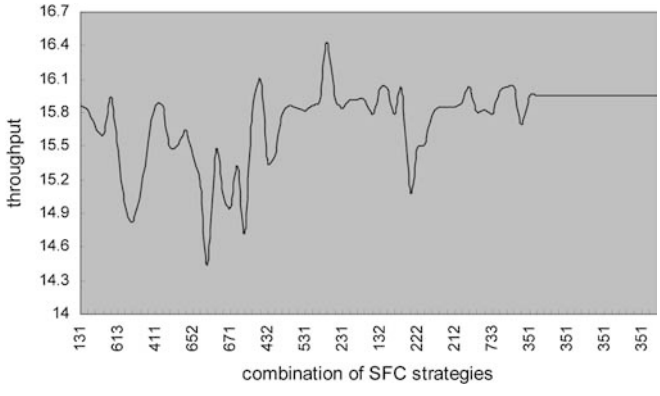

Fig. 10 The convergence process for the throughput near local optimum solution

Table 7 Results from SFC strategy combinations using SA-based methodology

Performance indicators	Near local optimum solution	Optimum value	Near global optimum solution	Optimum value	Total annealing times
WIP	622	326.027	652	303.291	51
FT	721	39.891	721	39.891	81
tardiness	561	0.061	561	0.061	42
tardy rate	223	0.046	223	0.046	30
delay cost	433	1.029	433	1.029	42
throughput	351	15.952	333	16.426	78

rate, delay cost, and throughput are included. As Table 7 shows, the SA-based methodology that was found to perform well for combinations of SFC strategies varies according to the different performance indicators. TB*SRPT*lot-spit, WCEDD*EDD*lock-step, POISSON*COVERT*lock-step, CONWIP*EDD*rendezvous, UNIF*CR*rendezvous, and SA*CR*rendezvous are the performance combinations for the WIP, FT, tardiness, tardy rate, delay cost, and throughput SFC strategies, respectively. Table 7 shows that not any one SFC strategy combination can satisfy all performance indicators. Hence, considering the trade-off among these production control strategies, we should choose the strategy based on system control. Besides, Table 7 shows that the objects of WIP, FT, tardiness, tardy rate, and delay cost are minimal, while throughput is maximal. The local optimum solution did not equal the global optimum solution for WIP and throughput performance indicators. This means that the two local optimum solutions are not the best in the convergence process. The near global optimum solutions are missed for WIP and throughput performance indicators in the convergence process. Therefore, the SA-based methodology can prevent missing the near global optimum solution.

Sha et al. [12] mentioned that, while redressing defects by reprocessing wafers increases the cycle time and manufacturing cost, it can reduce costs associated with defects. Thus, the wafer rework during the photolithography stage is an important study topic. Moreover, the research is to find a better strategy, that is, the Rendezvous strategy (coding 3 for rework strategy). However, Table 7 shows that the rework strategy of every global optimum solution is not all rendezvous strategy. Therefore, when considering integrating the ORR, dispatching, and rework strategies simultaneously on SFC, the rendezvous strategy may not be a suitable strategy for every performance indicator. For the same reason as with rework strategy, ORR and dispatching strategies have the same feature.

On the other hand, from the results in Table 7, 51, 81, 42, 30, 42, and 78 annealing times were required to converge to a specific target. However, if DOE was used to solve this problem, $7*8*3=168$ combinations run would be required. The SA-based methodology requires fewer simulation runs than DOE. A great amount of computational time was saved using the method proposed in this research.

6 Conclusions and future research

This paper has made an attempt to integrate the SFC strategies for several performance indicators in wafer fabrication. The research included the rework strategy while considering the ORR, dispatching, and rework strategies simultaneously in a simulation model. The combination of ORR, dispatching, and rework strategies will improve the performance of production control, and the model is more complete than those in previous research.

This paper presents SA-based methodology procedure for linking the simulation and SA modules directly applied to solving the integration of SFC strategies problem in wafer fabrication using several performance indicators. Particularly, the SA-based methodology procedure of generate neighborhood solution is different than the typical SA. And it can prevent missing the near global optimum solution. The results showed that the SA-based methodology was found to have outstanding SFC strategy combinations using different performance indicators. But there is not any one SFC strategy combination can satisfy all performance indicators. Hence, considering the trade-off among these production control strategies, we should choose a suitable SFC strategy combination based on system control targets. Besides, as previous research has only considered one type of SFC strategy, we can see that one specific strategy is significantly better than other strategies. But for SFC integration strategies, the specific strategy may not be a suitable strategy. On the other hand, great computational time was saved using the method proposed in this research. The near global optimum SFC strategy combination can be found by SA-based methodology quickly. If the problem scale becomes greater (i.e. more types and greater numbers of SFC strategies), significant computational time will be saved.

Several topics can be discussed in future research. An integral strategy for SFC can be developed. ORR or dispatching strategies considering the demand for rework will improve the system performance. Batching, due date assignment strategies, etc., can be added to this problem to produce a more complete SFC strategy. Furthermore, we have considered several performance indicators simultaneously in order to find a suitable combination of SFC strategies using multi-criteria decision-making.

References

1. Uzsoy R, Lee CY, Martin-Vega LA (1992) A review of production planning and scheduling models in the semiconductor industry Part I: System characteristics, performance evaluation and production planning. IIE Trans 24(4):47–60
2. Uzsoy R, Lee CY, Martin-Vega LA (1994) A review of production planning and scheduling models in the semiconductor industry Part II: Shop-floor control. IIE Trans 6(5):47–60
3. Melynk SA, Ragatz GL (1988) An evaluation of order release mechanisms in a job-shop environment. Decis Sci 19(1):167–189
4. Glassey CR, Resende MGC (1988) A scheduling rule for job release in semiconductor fabrication. Oper Res Lett 7(5):213–217
5. Wein LM (1988) Scheduling semiconductor wafer fabrication. IEEE Trans Semicond Manuf 1(3):115–128
6. Miller DJ (1990) Simulation of a semiconductor manufacturing line. Commun ACM 33(10):99–108
7. Lou SXC, Kager PW (1989) A robust production control policy for VLSI wafer fabrication. IEEE Trans Semicond Manuf 2(4):159–164
8. Spearman ML, Woodruff DL, Hopp WJ (1990) CONWIP: A pull alternative to kanban. Int J Prod Res 28(5):879–894
9. Lee CE, Chen CW (1997) A dispatching scheme involving move control and weighted due date for wafer foundries. IEEE Trans Compon, Pack, and Manuf Technol C 20(4):268–277
10. Glassey CR, Resende MGC (1988) Closed-loop job release control for vlsi circuit manufacturing. IEEE Trans Semicond Manuf 1(1):36–46
11. Kim YD, Kim JU, Lim SK, Jun HB (1998) Due-date based scheduling and control policies in a multiproduct semiconductor wafer fabrication facility. IEEE Trans Semicond Manuf 11(1):155–164
12. Sha DY, Hsieh LF, Chen KJ (2001) Wafer rework strategies at the photolithography stage. Int J Ind Eng 8(2):122–130
13. Zarger A (1995) Effect of rework strategies on cycle time. 17th International Conference on Computers and Industrial Engineering 29(1–4):239–243
14. Lu SCH, Ramaswamy D, Kumar PR (1994) Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants IEEE Trans Semicond Manuf 7(3):374–388
15. Fowler JW, Brown S, Gold H, Schoemig A (1997) Measurable improvements in cycle-time-constrained capacity. IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings pp. A21–A24
16. Hsieh BW, Chen CH, Chang SC (1999) Fast fab scheduling rule selection by ordinal comparison-based simulation. IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings, pp. 53–56
17. Chung SH, Huang HW (1999) The design of production activity control policy. J Chin Inst Ind Eng 16(1):93–113
18. Kim J, Leachman RC, Suh B (1996) Dynamic release control policy for the semiconductor wafer fabrication lines. J Operat Res Soc 47(12):1516–1525
19. Kim YD, Lee DH, Kim JU, Roh HK (1998) A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities. J Manuf Syst 17(2):107–117
20. Sha DY, Hsieh LF, Lin SH (2001) A study on wafer rework strategies and dispatching rules at the photolithography stage. ProdInventory Manage J (Rev)
21. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A (1953) Equation of state calculation by fast computing machines. J Chem Phys 21:1087–1092
22. Kirkpatrick S, Gelatt CD Jr., Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
23. Chang HH (1999) Optimal parameter design via soft computing. Dissertation, National Chao Tung University
24. Ponnambalam SG, Jawahar N, Aravindan P (1999) A simulated annealing for job shop scheduling. Prod Plann Control 10(8):767–777
25. Beragamaschi D, Cigolini R, Perona M, Portioli A (1997) Order review and release strategies in a job shop environment: A review and a classification. Int J Prod Res 35(2):399–420
26. Blackstone JH, Phillips DT, Hogg GL (1982) A state-of-the-art survey of dispatching rules for manufacturing job shop operations. Int J Prod Res 20(1):27–45
27. Yan H, Lou S, Sethi S, Gardel A, Deosthail P (1996) Testing the robustness of two-boundary control policies in semiconductor manufacturing. IEEE Trans Semicond Manuf 9(2):285–288