Int J Adv Manuf Technol
DOI 10.1007/s00170-013-4762-y

ORIGINAL ARTICLE

Heuristic scheduling policies for a semiconductor wafer fabrication facility: minimizing variation of cycle times

Hyun Joong Yoon ·
Jin Gon Kim

Received: 27 March 2012 / Accepted: 13 November 2012

Abstract This paper presents heuristic scheduling policies for semiconductor wafer fabrication facilities. The proposed heuristic scheduling policies include the advanced operation due date (OPNDD) for a sequence control policy and the adaptive constant work-in-process (CONWIP) for an input release control policy. The objective of the proposed scheduling policies is to reduce the variation of cycle times in the wafer fab. The advanced OPNDD sets the higher priority to the front opening unified pod (FOUP) with the smallest operation due date that is computed using a generalized stochastic Petri net model, and at the same time regulates the queue lengths of the FOUPs in each stoker by preventing excessive queue lengths in bottleneck workstations. The adaptive CONWIP controls dynamically the input release time of FOUPs using the adaptive WIP level according to the current status of the wafer fab. The simulation experiments show that the proposed scheduling method is efficient in reducing the variation of cycle times.

Keywords Scheduling · Semiconductor · Wafer fab · Generalized stochastic Petri nets

1 Introduction

Semiconductor wafer fabrication, the process of building integrated circuits on a silicon wafer, is a complex production process composed of hundreds of sequential and reentrant operation steps. A wafer fab is a facility at which semiconductor wafers are fabricated. As shown in Fig. 1, a general wafer fab facility is composed of scores workstations (or also called as bays) and inter-workstation material handling systems such as overhead shuttles to transport front opening unified pods (FOUPs) from one workstation to another. The inter-workstation material handling system, which spans the central aisle of the wafer fab, connects to all workstations at their respective stokers. Each workstation consists of process machines, a stoker to store FOUPs, and an intra-workstation material handling system such as automated-guided vehicle, rail-guided vehicle, and intra-bay overhead hoist transport. The intra-workstation material handling system transfers FOUPs among the stoker and the process machines. The scheduling policies of the wafer fab can be classified into the sequence control and the input release control policies: the former is to select a FOUP and allocate it to an available process machine in a workstation, and the latter is to determine a FOUP and its release time into the wafer fab. The objectives of the scheduling policies are generally maximization of the throughput, minimization of the mean and variance of cycle time, minimization of tardiness, maximization of the utilities of the process machines, and minimization of work-in-process (WIP) in a wafer fab facility, and so on.

Generally, semiconductor chip makers have considered the maximization of the throughput as the most important performance index in the wafer fab. However, they are trying to reduce the mean cycle time and the variance of the cycle times for several reasons recently. First, the response time to their customers' needs can be faster and more effectively, since reduction of the mean and variance of cycle time makes the companies to predict and plan capabilities of their production lines much easier and precisely. Secondly, the mean cycle time reduction comes with the WIP level reduction, which makes a wafer fab more lean and efficient. For instance, delivery times of FOUPs in the wafer fab with the less WIP level become faster since the

H. J. Yoon (✉) · J. G. Kim
Faculty of Mechanical and Automotive Engineering,
Catholic University of Daegu, Hayang-Eup,
Gyeongsan-Si, Gyeongbuk 712-702, Republic of Korea
e-mail: yoon@cu.ac.kr

J. G. Kim
e-mail: kimjg1@cu.ac.kr

Published online: 23 January 2013

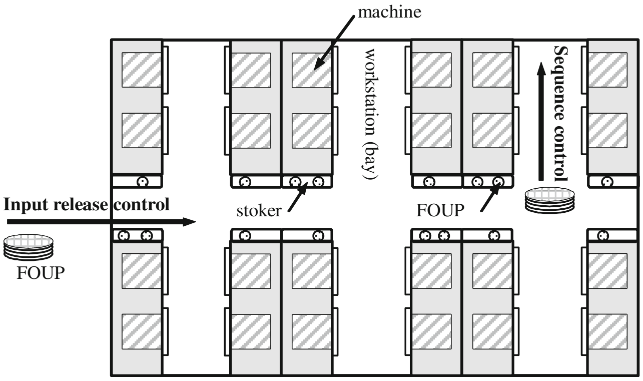

Fig. 1 GSPN model of a workstation with one stoker and *M* process machines

delivery loads of inter-/intra-workstation automated material handling system are lower by eliminating the meaningless transfers caused by excessive WIP. Although the lower cycle time can be easily achieved by reducing WIP level, the mean and the variance of the cycle time can also be realized through an effective scheduler.

In this paper, the scheduling policies in a wafer fab are classified into the sequence control policy to allocate FOUPs into process machines, and the input release control policy to determine the type of FOUPs and their releasing time into the wafer fab. Based on our previous research [1], in which the sequence control policy, named as operation due date (OPNDD) rule, is proposed, this paper presents the real-time heuristic scheduling policies that contain the advanced OPNDD, which is an advanced version of the OPNDD, for a sequence control policy, and the adaptive constant work-in-process (CONWIP) for an input release control policy. The objective of the proposed scheduling policies is to reduce the variation of cycle times.

The remainder of this paper is structured as follows. In Section 2, extensive literature reviews on related researches are provided, and Section 3 describes the proposed heuristic scheduling policies including the sequence control and the input release control policies. Section 4 provides the results of simulation experiments, and Section 5 concludes the paper.

2 Literature review

There have been many research studies addressing the scheduling problems in wafer fabs since 1980s. Uzsoy et al. [2, 3] classify the scheduling problem in a wafer fab into dispatching rules and input regulation rules, deterministic scheduling algorithms, control theoretic approaches, and knowledge-based approaches. Wein [4] reports the results of the simulation experiments, in which six input regulation rules and 12 dispatching rules are evaluated in terms of the mean cycle time. He insists that the input regulation rules are more effective in reducing the mean cycle time compared with the dispatching rules. Lu et al. [5] show that a proper selection of dispatching rule is helpful in reducing the mean and variance of the cycle time, and then propose fluctuation smoothing policy to reduce the mean and variance of the cycle time. Chung et al. [6] address the scheduling problem considering both the throughput and the cycle time, in which they control the WIP level of the bottleneck process to meet the desired throughput and cycle time. Kim et al. [7, 8] propose dispatching rules to minimize mean tardiness of orders in a wafer fab producing multiple product types having different due dates and different process flows. Yoon and Lee [1] propose a real-time scheduling method based on a dispatching rule approach to reduce the standard deviation of the cycle times in a wafer fab and show that the proposed scheduling method is efficient in meeting due date and reducing tardiness. Gupta and Sivakumar [9] present a review on job shop scheduling techniques in semiconductor manufacturing, in which scheduling techniques are classified into dispatching heuristic rules, mathematical programming techniques, neighborhood search methods, and artificial intelligence techniques, and then they proposes a look-ahead batch scheduling method for the real-time control of due date objectives in semiconductor batch manufacturing [10]. Chung and Lai [11] address job releasing and throughput planning problem under demand fluctuation. They consider an environment where product mix changes periodically and present a production scheduling method to plan the wafer lot release and throughput. Upasani et al. [12] propose the problem reduction procedure that allows a work center-based global scheduling heuristic to be implemented in very low CPU times. They partition the work centers in a wafer fab into heavily loaded and lightly loaded classes and solve the global scheduling problem only for the heavily loaded work center. Pfund et al. [13] model a semiconductor wafer fabrication process as a complex job shop and adapt a modified shifting bottleneck heuristic (MSBH), which is proposed by Mason et al. [14], to facilitate the multi-criteria optimization of makespan, cycle time, and total weighted tardiness using a desirability function. They use the desirability approach at two different levels of the MSBH, which are the sub-problem solution procedure level and the machine criticality measure level, and then propose five approaches for scheduling complex job shop. Yoon and Shen [15] address decision making problems with hard interoperation temporal constraints in semiconductor wafer fabrication facilities. The objective is to allocate wafer lots into each workstation to satisfy both logical and temporal constraints. The proposed decision making system is developed based on a multi-agent architecture that is composed of scheduling agents, workstation (or workcell) agents, machine agents, and product agents. Zhang et al. [16] propose a simulation-based evaluation and optimization method, in which dispatching rules are selected by the performance evaluation and parameters are optimized by response surface methodology and the simulation. They also propose a dynamic bottleneck dispatching policy, where bottlenecks are detected in a timely way and adaptive dispatching decisions are made according to the real-time conditions. Baez-Senties et al. [17] proposes an artificial neural network approach coupled with a multi-objective genetic algorithm for multi-decision scheduling problems in a semiconductor wafer fabrication. The proposed scheduler selects decision variables in order to obtain the desired performance index at the end of a given production horizon. The authors insist that the proposed approach can be applied easily to the semiconductor manufacturing and significant benefits can be achieved in terms of cycle time distribution, facility average utilization, average waiting time, and storage. Guo et al. [18] propose the decomposition-based classified ant colony optimization method that is composed of decomposed and classified ant colony optimization algorithm. A large and complicated scheduling problem is decomposed into several smaller subproblems, and then ant colony optimization algorithm is used to find solutions. Chiang and Fu [19] propose a dispatching rule for wafer lot scheduling in semiconductor wafer fabs in order to improve the design of existing rules by the index function based on total degree of urgency and the due date extension procedure. The authors insist that the proposed rule is superior in terms of on-time delivery rate, mean tardiness, and maximum tardiness. Thus, although dispatching heuristic rule-based scheduling approach has the primary disadvantages in that it cannot hope for an optimal solution, it is still widely used for real applications in semiconductor wafer fabrication facilities. The major reason may be due to the high complexities and dynamic contingencies of the real scheduling problems in the semiconductor wafer fabrication facilities. However, local optimal or near-optimal schedulers for the bottleneck process, e.g., photolithography, or batch processes are used to enhance the utilities and throughputs of machines in real applications.

Regarding the Petri net-based scheduling approach for semiconductor wafer fabrication facilities, the systemic review before 1998 can be found in Zhou [20]. Lin and Fu [21] present a generalized stochastic colored timed Petri net to model a wafer fabrication. The model includes the dynamic behavior such as loading, reentrant processing, unloading, and machine failure, and modular and synthesis techniques are used to construct a large full system model. Odrey et al. [22] present a generalized Petri net-based modeling approach for semiconductor wafer fabrication, in which three Petri net models representing a reentrant flow line with three work centers and six machines are provided. Chen et al. [23] propose a genetic algorithm embedded search strategy over a colored timed Petri net model for wafer fabrication. The proposed Petri net model is composed of routing module and elementary module, and a genetic algorithm-based scheduler dynamically searches for the appropriate dispatching rules for each machine group. Jain et al. [24] present a generalized stochastic Petri net model that captures dynamic behaviors such as reentrant processing, machine failures, loading, and unloading and propose a simulated annealing-based scheduling strategy to minimize mean cycle time and tardiness. Huang and Sun [25] propose and evaluate two Petri net-based hybrid search strategies, algorithms H1 and H2, and their applications to flexible manufacturing system scheduling. The algorithm H1 guarantees an optimal solution at termination with an admissible heuristic function, and the algorithm H2 invokes faster termination conditions than H1 while guaranteeing the optimality of the solution. Liu et al. [26] propose an extended object-oriented Petri net approach for the effective modeling of semiconductor wafer fabrication systems. Hierarchical approach and a special type of transition named as main bus gate are used to cope with the complexity in terms of the reentrant process routes. Then the multiple-object scheduling and real-time dispatching problems are addressed in their following literature [27]. Lee et al. [28] propose a supervisory framework considering human-in-the-loop situations in semiconductor manufacturing systems. In the human-in-the-loop system, human operations may violate desired requirements and lead to destructive failure. Thus, the supervisory framework is developed to guarantee that manual operations meet required specifications so as to prevent human errors in operation using Petri nets. Liu et al. [29] propose a timed extended object-oriented Petri nets approach to performance modeling, real-timed dispatching, and simulation of semiconductor wafer fabrication systems. To implement dispatching policies, the concepts of an autonomy and coordination-based real-time dispatching mechanism and hybrid real-time dispatching control system are introduced. Wu and Hsieh [30] propose a real-time fuzzy Petri net diagnoser to replicate the plant and detect fault in discrete manufacturing systems. While monitoring events generated in the manufacturing system, the real-time Petri net model compares the outputs with pre-setting. If any difference is detected, the fuzzy Petri net diagnoser starts to locate fault. Visual Basic is used for the implementation. Qiao et al. [31] present a hierarchical colored timed Petri net approach to describe various states, behaviors, and substructures of a wafer fabrication system. For the scheduling problem, the authors propose an extended genetic algorithm to optimize the combination of scheduling policies.

3 Heuristic scheduling policies

3.1 Sequence control policy

In this section, a new sequence control policy, named the advanced OPNDD, is proposed to minimize the variance of the cycle times by setting the higher priority to the FOUP that is relatively delayed in its operation. The delayed time of an operation is obtained using the operation due date that is a new index computed from generalized stochastic Petri net (GSPN) model. The advanced OPNDD also regulates the queue lengths of the FOUPs in each stoker by preventing excessive queue lengths in the bottleneck workstations.

A GSPN without priorities and inhibitor arcs is a six-tuple $G=\{P, T, I, O, W, M_0\}$, where $P=\{p_1, p_2, \ldots, p_m\}$ is a finite set of places, $T=\{t_1, t_2, \ldots, t_n\}$ is a finite set of transitions, $P\cup T\neq\varnothing$ and $P\cap T=\varnothing$, $I{:}(P\times T)\rightarrow N$ is a set of input functions, where N is a set of non-negative integers, $O{:}(P\times T)\rightarrow N$ is a set of output functions, $W{:}T\rightarrow\mathfrak{R}$ is a function defined on the set of transitions, and $M_0{:}P\rightarrow N$ is the initial marking. $I(p, t)=k$ implies that there are k directed arcs connecting from place p to transition t, whereas $O(p, t)=k$ implies that there are k directed arcs connecting from transition t to place p.

Figure 2 depicts an example of the GSPN model for a process machine. Since the aim of adopting the GSPN model in this paper is to compute a new performance index of a workstation, named as a *utilization index*, it is not necessary to comprise GSPN models for an entire wafer fab facility. The GSPN model of the process machine is composed of three places, i.e., $p_{\mathrm{PM\_PROC}}$, $p_{\mathrm{PM\_AVAIL}}$, and $p_{\mathrm{PM\_FAIL}}$, and four transitions, i.e., $t_{\mathrm{PM\_START}}$, $t_{\mathrm{PM\_FIN}}$, $t_{\mathrm{PM\_FAIL}}$, and $t_{\mathrm{PM\_REPAIR}}$. If mean processing time (MPT), mean time between failure (MTBF), and mean time to repair (MTTR) are assumed to be exponentially distributed, the rate of the four transitions, λ_1, λ_2, λ_3, and λ_4, are 2/MPT, 2/MPT, 1/MTBF, and 1/MTTR, respectively. Tables 1 and 2 show the specifications of the places and the transitions of the GSPN models.

The process machines involved in a workstation is assumed to be perform the same operation, for instance, oxidation, deposition, photolithography, etching, ion implant, etc., with identical MPT, MTBF, and MTTR. The utilization index (UI) for a workstation w is defined as follows:

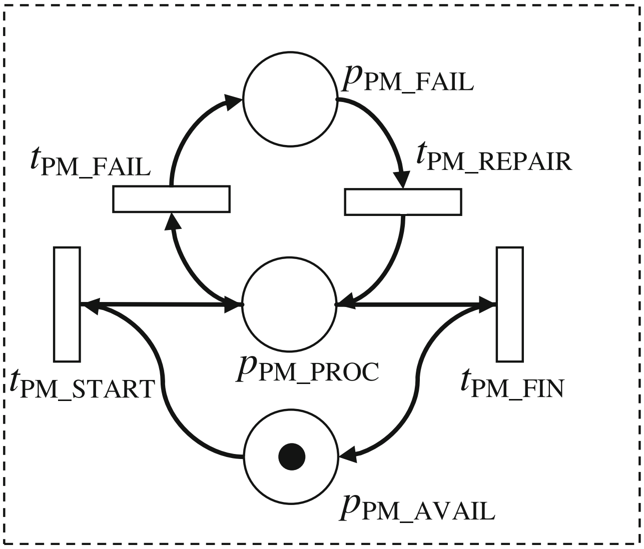

Fig. 2 GSPN model of a process machine [1]

Table 1 Specification and initial marking of places of the GSPN model

Place	Semantics	Initial marking
$p_{\mathrm{PM\_PROC}\_m}$	Process machine processing	0
$p_{\mathrm{PM\_AVAIL}\_m}$	Process machine ready to process	1
$p_{\mathrm{PM\_FAIL}\_m}$	Process machine breakdown	0

$$\mathrm{UI_w} = \mathrm{U_w}/\sum\nolimits_{\mathrm{i=1}}^{\mathrm{W}} \mathrm{U_i}, \tag{1}$$

where U_w is defined as follows:

$$U_w = \sum\nolimits_{\tau=1}^{\mathrm{T}} \mu_\tau \cdot \frac{\mathrm{NV}_{\mathrm{w},\tau}}{M_w \cdot \mathrm{TR_w}}. \tag{2}$$

The ratio of production amount for the FOUP type τ (τ=1, 2, …, T) to total productions is denoted by μ_τ ($\mu_1+\mu_2+\ldots+\mu_T=1$), the number of visits to the workstation w by the FOUP type τ is denoted by $NV_{w,\tau}$, the number of identical process machines in the workstation w is denoted by M_w, and the mean throughput rate of each process machine in the workstation w is denoted by TR_w. The mean throughput rate is computed as $\mathrm{TR}_w=\lambda_1\lambda_2\lambda_4/(\lambda_1\lambda_3+\lambda_1\lambda_4+\lambda_2\lambda_4)$ by analyzing the GSPN model of the process machine [32, 33].

The OPNDD rule [1] sets the higher priority to the FOUP with the smallest operation due date, the *opndd*. The opndd for the lth operation of the FOUP π of type τ is defined as follows:

$$\begin{gathered} opndd_{\pi,l} = opndd_{\pi,l-1} = \Delta opndd_{\pi,l}, opndd_{\pi,0} = \gamma_\pi \\ \Delta opndd_{\pi,l} = \frac{LT_\tau}{\sum_{i=1}^{L} UI_{\tau,i}} \cdot UI_{\tau,l}, \end{gathered} \tag{3}$$

where LT_τ is the lead time of the FOUP type τ, L is the number of operations of the FOUP type τ, $\mathrm{UI}_{\tau,l}$ is the utilization index of the workstation at which the lth operation of the FOUP type τ is performed, and γ_π is the released time of the FOUP π into the wafer fab.

Although, as reported in [1], the OPNDD rule shows the good performance in reducing the variation of cycle times, it

Table 2 Specification of transitions of the GSPN model

Transition	Semantics	Rate
$t_{\mathrm{PM\_START}}$	Start processing of process machine	λ_1
$t_{\mathrm{PM\_FIN}}$	Complete processing of process machine	λ_2
$t_{\mathrm{PM\_FAIL}}$	Process machine breakdown	λ_3
$t_{\mathrm{PM\_REPAIR}}$	Process machine repair	λ_4

has a drawback in that the queue length of the FOUPs waiting to be released is not taken into consideration. The bottleneck workstations have relatively longer queue length than ordinary workstations, since the FOUPs waiting to be released into the bottleneck workstations are delayed in their operation steps. However, the FOUPs waiting for the bottleneck workstations will have relatively smaller values of opndd and, therefore, the OPNDD rule repetitively sets higher priorities for them. This phenomenon makes the cycle times longer. Thus, the advanced OPNDD rule is designed to regulate the queue lengths of the FOUPs in each stoker by preventing excessive queue lengths in the bottleneck workstations.

If we consider a workstation as a single server M/M/1 queueing system with the arrival rate of λ and the service rate of μ, the average queue length L_Q is given by [34]

$$L_Q = \lambda/(\mu - \lambda). \tag{4}$$

From the Little's law, we have

$$\lambda = L_Q/T_Q, \tag{5}$$

where T_Q is the average system time. From (4) and (5), the average queue length L_Q is obtained as follows:

$$L_Q = \left(\mu - \frac{1}{T_Q}\right) \cdot T_Q. \tag{6}$$

Thus, the upper bound of the queue length for the FOUP type τ in the workstation w, $W\max[w,\tau]$, is determined as

$$W\max[w,\tau] = \left[(\mu_\tau \cdot \mathrm{TR_w} - 1/\Delta\mathrm{opndd}[\mathrm{w},\tau]) \cdot \Delta\mathrm{opndd}[\mathrm{w},\tau]/\mathrm{NV}_{\mathrm{w},\tau}\right], \tag{7}$$

where Δopndd[w,τ], or equivalently the expected system time for the FOUP type τ in the workstation w, is given by

$$\Delta\mathrm{opndd}[\mathrm{w},\tau] = \frac{\mathrm{LT}_\tau}{\sum_{\mathrm{i}=1}^{\mathrm{L}} \mathrm{UI}_{\tau,\mathrm{i}}} \cdot \mathrm{UI_w}. \tag{8}$$

3.2 Input release control policy

In this section, the adaptive CONWIP is proposed for the input release control policy to control the WIP level dynamically in a wafer fab. Generally, input release control policies are classified into the release rate control approach and the WIP control approach. Comparing with the release rate control approach, the WIP control approach has some benefits [35]: first, whereas, for the release rate control approach, the throughput of near future should be predicted in a finite time horizon, a WIP level is directly observable, and it is, therefore, directly controllable. Secondly, the control of the WIP is more robust to error than that of the release rate. It is, therefore, well known that the WIP control approach achieves less WIP and less cycle time variability than the release rate control approach.

In regard to the WIP control approach, the CONWIP rule is different from the traditional pull control like the Kanban of the Toyota Production System. Under the CONWIP, a new job is introduced to a given production line each time a job departs, which results in the WIP level very constant [36]. The CONWIP system looks like a closed queueing network, in which jobs never leave the system but instead circulate around the network indefinitely [35]. However, the CONWIP rule has a limitation that it does not consider dynamic contingencies in a production line such as machine breakdowns. To cope with this problem, the adaptive CONWIP rule is newly presented for the input release control policy in the wafer fab. The major objective of the adaptive CONWIP rule is to adjust the WIP level dynamically according to the current status of the wafer fab. The adaptive CONWIP controls the WIP level in a wafer fab dynamically using the adaptive WIP level, $\mathrm{WIP_{adapt}}$, which is varying dynamically according to the current status of the wafer fab.

The adaptive WIP level is computed as follows. First, a total operation lateness is defined as the sum of the differences between the completion time and the operation due date for the all FOUPs in the wafer fab, that is,

$$total\ opnlat = \sum\nolimits_{FOUPs\ in\ fab} \left(C_{\pi,l-1} - opndd_{\pi,l-1}\right), \tag{9}$$

where, if the current operation of the FOUP π is l, $\mathrm{c}_{\pi,l-1}$ is the completion time and $opndd_{\pi,l-1}$is the operation due date of (l−1)th operation of the FOUP π, respectively. Then, the adaptive WIP level, $\mathrm{WIP_{adapt}}$, is computed as follows:

$$\mathrm{WIP_{adapt}} = \mathrm{WIP_{ref}} + \alpha \cdot (\mathrm{total\ opnlat}), \tag{10}$$

where $\mathrm{WIP_{ref}}$ is a predetermined reference WIP level, and α is a parameter to convert the total operation lateness to the WIP level.

The parameter α, for instance, can be obtained from the Little's law. If $\mathrm{CT_{ref}}$ is a reference cycle time and $\mathrm{TH_{capa}}$ is a reference throughput of a production line, then the reference WIP level is given by

$$\mathrm{WIP_{ref}} = \mathrm{CT_{ref}} \times \mathrm{TH_{capa}}. \tag{11}$$

Usually, $\mathrm{CT_{ref}}$ is unknown, but $\mathrm{TH_{capa}}$ is given as the production capacity of the production line. If there is a perturbation due to a certain dynamic contingency,

$$\begin{aligned}\mathrm{WIP_{adapt}} &= \mathrm{WIP_{ref}} + \Delta\mathrm{WIP} = (\mathrm{CT_{ref}} + \Delta\mathrm{CT}) \times \mathrm{TH_{capa}} \\ &= \left(\mathrm{CT_{ref}} = \tfrac{\mathrm{total\ opnlat}}{\mathrm{N}}\right) \times \mathrm{TH_{capa}} = \mathrm{WIP_{ref}} + \tfrac{\mathrm{TH_{capa}}}{\mathrm{N}} \times (\mathrm{total\ opnlat}),\end{aligned} \tag{12}$$

where N is the number of FOUPs in the wafer fab. From the Eqs. (10) and (12), the parameter α is obtained as follows:

$$\alpha = \mathrm{TH}_{\mathrm{capa}}/\mathrm{N}. \tag{13}$$

4 Simulation experiments

4.1 Simulation model

The proposed sequence control policy and input release control policy are evaluated through simulation experiments using the Hewlett-Packard Technology Research Center Silicon fab (the TRC fab) [4], which is a development facility in Palo Alto, California. Although the TRC fab is relatively smaller than ordinary production facilities, it uses essentially the same type of equipment for execution of essentially the same operation. In the simulation model, three versions of the wafer fab facilities are considered: FAB1, FAB2, and FAB3. Each wafer fab is composed of 24 workstations, and each workstation contains one or more parallel identical machines. The only difference between FAB1, FAB2, and FAB3 is the number of process machines in each workstation. Note that none of the three wafer fabs can exchange FOUP units. Table 3 shows the operation type, number of process machines, MPT, MTBF, and MTTR for each workstation. The numbers of bottleneck workstations are different for each wafer fab version since the number of machines in each workstation varies according to the versions of the wafer fabs. For instance, if the bottleneck workstations are defined as ones with utilization of 90 % or higher, there is one bottleneck workstation in the FAB1, and two bottleneck workstations in the FAB2, and four bottleneck workstations in the FAB3, where the utilization is defined as [((release rate) (number of visits per FOUP) (MPT)/(number of process machines)) + (MTTR)/(MTBF + MTTR)]×100 %. All of the time-related parameters are assumed to be exponentially distributed. All identical machines in a workstation have the same distributions of mean processing time, mean time between failure, and mean time to repair. The mean processing time includes setup time and rework time, and the machine failure includes unexpected machine failure, periodic maintenance, and machine tuning. The operation flow in the TRC fab is given in [4].

Table 3 Plant data of the TRC fab [4]

Workstation no.	Type of operation	Number of machines			MPT (hour)	MTBF (hour)	MTTR (hour)
		FAB1	FAB2	FAB3			
1	Cleaning	2	2	1	1.55	42.18	2.22
2	Oxidation	2	2	1	4.98	101.11	10.00
3	Oxidation	2	2	1	5.45	113.25	5.21
4	Oxidation	1	1	1	4.68	103.74	12.56
5	Deposition	1	1	1	6.14	100.55	6.99
6	Deposition	1	1	1	7.76	113.25	5.21
7	Deposition	1	1	1	6.23	16.78	4.38
8	Deposition	1	1	1	4.35	13.22	3.43
9	Deposition	1	1	1	4.71	10.59	3.74
10	Deposition	1	1	1	4.05	47.53	12.71
11	Deposition	1	1	1	7.86	52.67	19.78
12	Deposition	1	1	1	6.10	72.57	9.43
13	Photolithography	4	4	2	4.23	22.37	1.15
14	Photolithography	3	3	3	7.82	21.76	4.81
15	Photolithography	1	1	1	0.87	387.20	12.8
16	Photolithography	2	2	1	2.96		No failure
17	Photolithography	1	1	1	1.56	119.20	1.57
18	Photolithography	1	1	1	3.59		No failure
19	Etching	2	2	1	13.88	46.38	17.42
20	Etching	1	1	1	5.41	36.58	9.49
21	Etching	2	2	1	7.58	36.58	9.49
22	Etching	2	2	1	1.04	118.92	1.08
23	Resist strip	2	2	1	1.09		No failure
24	Ion implant	1	1	1	3.86	55.18	12.86

4.2 Simulation results

The proposed sequence control policy, named as the advanced OPNDD, and the input release control policy, named as the adaptive CONWIP, are evaluated using the TRC fab model. For the simulation experiments, we use AutoMod version 12.1 (Applied Materials, Inc.) that is a commercial discrete event simulation software widely used in semiconductor industry. With 10,000 FOUPs, five simulations are carried out for each simulation case. When a machine fails during its operation, the remaining operation is completed after the machine is repaired. The performances of the proposed scheduling method are compared with seven sequence control policies, i.e., FIFO, FIFO+, SRPT, SRPT+, LWNQ/M, FSVCT, and OPNDD under the deterministic input release control policy. We use FIFO+, SRPT+, and LWNQ/M rules same as described in [4], FSVCT as in [5], and OPNDD as in [1]. The deterministic input release control policy releases FOUPs with constant interval times. The description of these sequence control policies are as follows.

FIFO : Select the FOUP that arrives at the queue at the earliest time.

FIFO+ : If there are FOUPs going next to a workstation which has a queue of size four or smaller, select among these using FIFO. If not, use FIFO [4].

SRPT : Select the FOUP that has the shortest expected remaining processing time until it exits the wafer fab.

SRPT+ : If there are FOUPs going next to a workstation which has a queue of size four or smaller, select among these using SRPT. If not, use FIFO [4].

LWNQ/M : Select the FOUP going to the workstation with the least amount of expected work per process machine [4].

FSVCT : Select the FOUP with smallest $\alpha(\pi)-\zeta$, where $\alpha(\pi)$ is the release time of the FOUP, and ζ is the estimate of the remaining delay [5]

OPNDD : Select the FOUP with the smallest operation due date [1].

The three graphs of Fig. 3 show the simulation results with the single wafer type in the FAB1, FAB2, and FAB3, respectively. The abscissa is the applied sequence control policies and the ordinates are the standard deviation of the cycle times and the mean cycle time in hours. The proposed scheduling method, advanced OPNDD with adaptive CONWIP, shows the best performance among the applied scheduling policies in the three FABs. As for the FAB1, the results show that the proposed scheduling method reduces the standard deviation of the cycle times by 4.5, 6.9, and 44.7 % compared with the OPNDD, the FSVCT, and the FIFO, respectively. As for the FAB2, the results show that the proposed scheduling method reduces the standard deviation of cycle times by 7.5, 28.8, and 58.9 % compared with the OPNDD, the FSVCT, and the FIFO, respectively. As for the FAB3, the results show that the proposed scheduling method reduces the standard deviation of cycle times by 7.9, 30.6, and 58.5 % compared with the OPNDD, the FSVCT, and the FIFO, respectively.

To investigate the performances of the adaptive CONWIP, additional simulation experiments are performed with the single FOUP type. Note that the aim of the adaptive CONWIP is to make mean cycle time uniformly by adjusting dynamically the

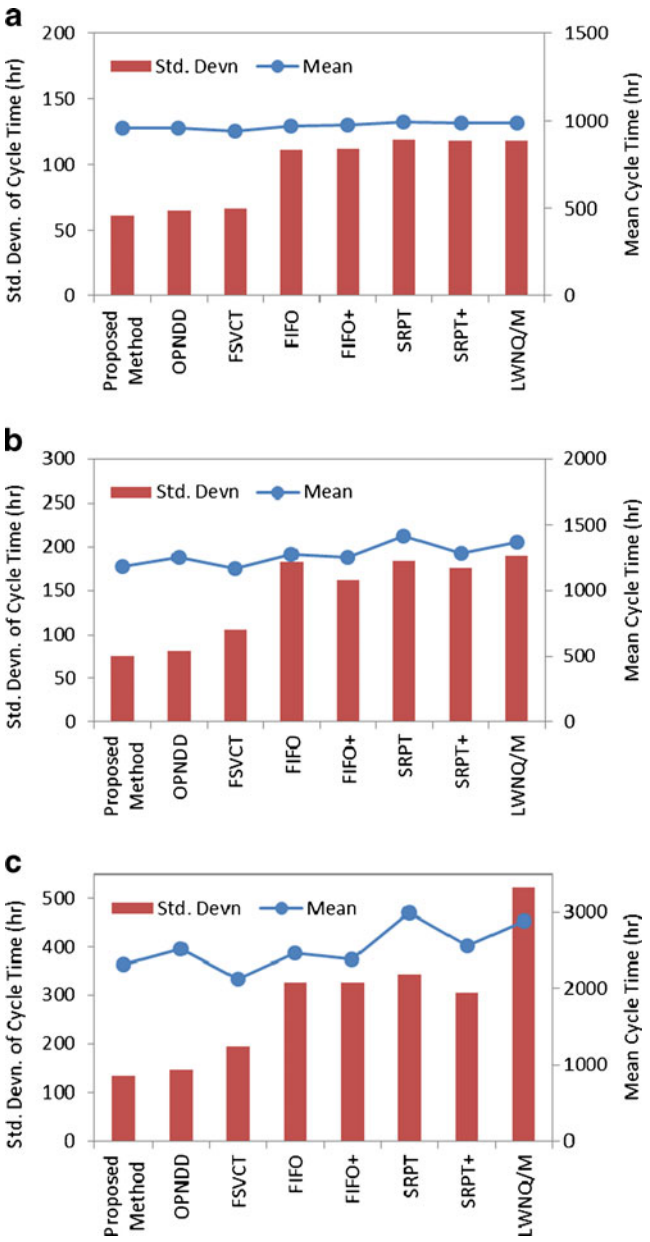

Fig. 3 The simulation results with the single FOUP type **a** in the FAB1, **b** in the FAB2, and **c** in the FAB3. They show the standard deviation of cycle times and the mean cycle time under the proposed scheduling method and other seven sequence control policies

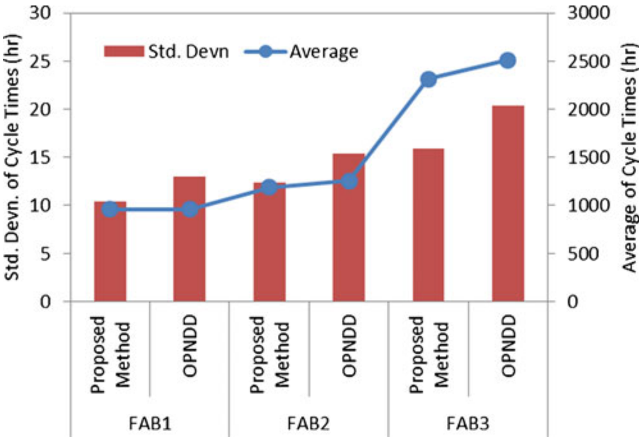

Fig. 4 The standard deviation and the average of mean cycle times with single FOUP type under the proposed scheduling method and the OPNDD with the deterministic input release control policy

WIP level in a wafer fab. The proposed scheduling method is compared with the OPNDD with the deterministic input release control policy. One hundred simulations are carried out for each scheduling case, and single-type 10,000 FOUPs are used for each simulation experiment. A mean cycle time is obtained with 10,000 FOUPs in each simulation experiment, and then the average and the standard deviation of 100 mean cycle times are computed. Figure 4 shows the simulation result, which indicates that the standard deviation of the mean cycle times is reduced by 20.0 % in the FAB1, 19.6 % in the FAB2, and 21.8 % in the FAB3, when the proposed scheduling method is applied.

For the next simulation experiments, three FOUP types A, B, and C are considered to evaluate the performances of the proposed scheduling method with the multiple FOUP types. The ratio of production amount for each FOUP type is assumed to be 2:3:5, which implies that the production amount of FOUP A, B, and C is 20, 30, and 50 %, respectively. The operation flow of the FOUP A is identical with one given in [4], and those of the FOUP B and C are given in Fig. 5. Each number in Fig. 5 indicates the corresponding workstation number from 1 to 24 shown in Table 3. With total 10,000 FOUPs, i.e., the production amounts of the FOUP A, B, and C are 2,000, 3,000, and 5,000, respectively, five simulations are carried out for each case. The proposed scheduling method is compared with the six sequence control policies of OPNDD, FIFO, FIFO+, SRPT, SRPT+, and LWNQ/M under the deterministic input release control policy. In the multiple FOUP simulation experiments, FSVCT rule is not considered since it is designed only for the single FOUP type.

Figure 6 shows the simulation results with the multiple FOUP types in the FAB1, FAB2, and FAB3. In the FAB 1, with the proposed scheduling method, the standard deviations of cycle times are reduced by 8.1, 10.2, and 8.0 % compared with the OPNDD for the FOUP A, B, and C, respectively; the standard deviations of cycle times are reduced by 48.8, 51.4, and 47.3 % compared with the FIFO for the FOUP A, B, and C, respectively. In the FAB 2, with the proposed scheduling method, the standard deviations of cycle times are reduced by 9.8, 9.0, and 7.4 % compared with the OPNDD for the FOUP A, B, and C,

Fig. 5 Operation flow of (**a**) FOUP B and (**b**) FOUP C. Each *number* indicates the workstation number

ENTER → 1 → 3 → 13 → 14 → 23 → 15 → 16 → 23 → 15 → 16 → 24 → 23 → 22 → 17 → 1 → 3 → 10 → 22
→ 12 → 6 → 22 → 6 → 1 → 1 → 4 → 10 → 19 → 23 → 1 → 1 → 13 → 14 → 23 → 15 → 16 → 24 → 23 → 22
→ 17 → 1 → 2 → 8 → 9 → 2 → 22 → 1 → 4 → 22 → 22 → 1 → 2 → 13 → 14 → 23 → 15 → 16 → 24 → 24 →
23 → 22 → 17 → 24 → 1 → 2 → 7 → 1 → 17 → 1 → 1 → 3 → 13 → 14 → 16 → 24 → 23 → 22 → 17 → 24 → 1
→ 2 → 7 → 1 → 17 → 1 → 1 → 3 → 13 → 14 → 16 → 24 → 23 → 22 → 17 → 9 → 21 → 1 → 3 → 13 → 14 →
15 → 23 → 15 → 16 → 24 → 23 → 22 → 17 → 1 → 2 → 13 → 14 → 23 → 15 → 20 → 22 → 23 → 22 → 17 →
13 → 14 → 15 → 23 → 16 → 24 → 23 → 22 → 17 → 1 → 8 → 4 → 22 → 22 → 1 → 2 → 8 → 13 → 14 → 18 →
23 → 15 → 16 → 23 → 18 → 22 → 1 → 10 → 13 → 14 → 16 → 21 → 12 → 13 → 14 → 18 → 23 → 15 → 15 →
15 → 16 → 19 → 23 → 22 → 17 → 11 → 13 → 14 → 15 → 21 → 23 → 5 → **EXIT**

ENTER → 1 → 3 → 13 → 14 → 15 → 23 → 15 → 16 → 24 → 23 → 22 → 17 → 1 → 3 → 10 → 22 → 12 → 6 →
22 → 6 → 1 → 1 → 4 → 10 → 19 → 23 → 1 → 4 → 22 → 22 → 1 → 2 → 7 → 1 → 3 → 22 → 13 → 15 → 23 →
22 → 22 → 22 → 17 → 13 → 14 → 18 → 23 → 15 → 16 → 20 → 23 → 1 → 17 → 1 → 1 → 3 → 13 → 14 → 16
→ 24 → 23 → 22 → 17 → 9 → 21 → 1 → 2 → 8 → 13 → 14 → 18 → 23 → 15 → 16 → 23 → 18 → 22 → 1 → 1
→ 13 → 14 → 23 → 15 → 16 → 24 → 23 → 22 → 17 → 1 → 2 → 13 → 14 → 23 → 15 → 20 → 22 → 23 → 22
→ 17 → 13 → 14 → 15 → 23 → 16 → 24 → 23 → 22 → 17 → 1 → 8 → 4 → 22 → 22 → 1 → 10 → 13 → 14 →
16 → 21 → 12 → 13 → 14 → 18 → 23 → 15 → 15 → 15 → 16 → 19 → 23 → 22 → 17 → 11 → 13 → 14 → 15 →
21 → 23 → 5 → **EXIT**

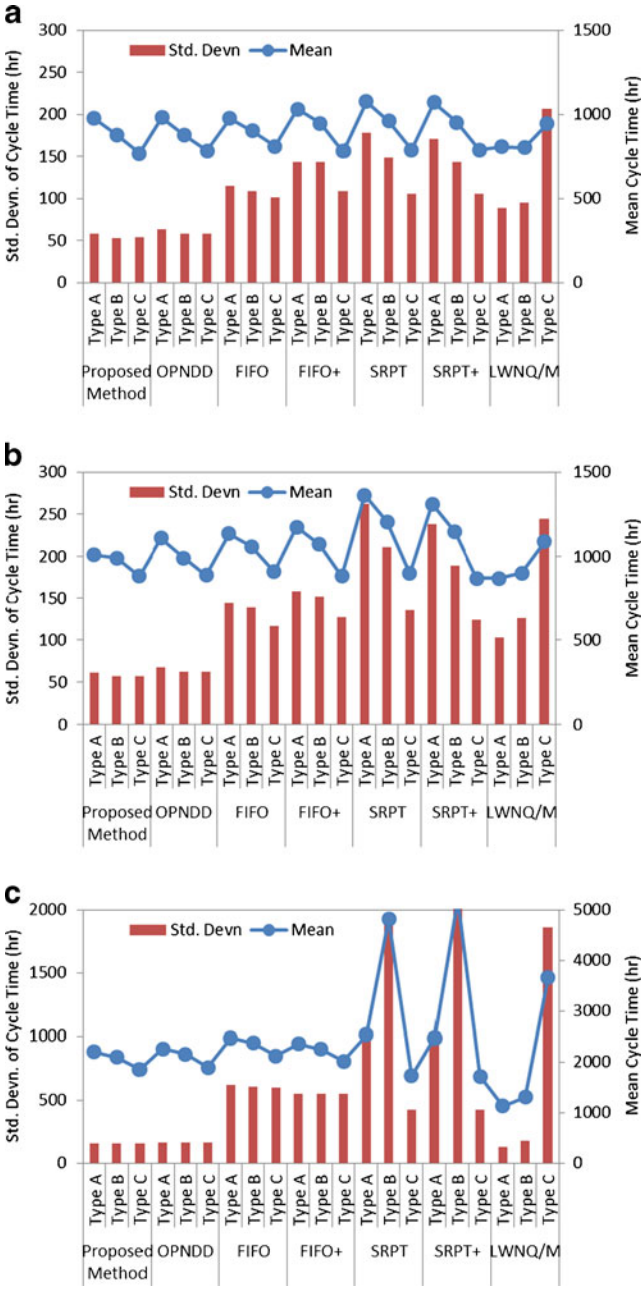

Fig. 6 The simulation results with the multiple FOUP types **a** in the FAB1, **b** in the FAB2, and **c** in the FAB3. They show the standard deviation of cycle times and the mean cycle time under the proposed scheduling method and other six sequence control policies

respectively; the standard deviations of cycle times are reduced by 57.6, 58.7, and 50.7 % compared with the FIFO for the FOUP A, B, and C, respectively. In the FAB 3, with the proposed scheduling method, the standard deviations of cycle times are reduced by 6.5, 6.1, and 4.0 % compared with the OPNDD for the FOUP A, B, and C, respectively; the standard deviations of cycle times are reduced by 75.2, 74.1, and 73.5 % compared with the FIFO for the FOUP A, B, and C, respectively.

To investigate the performances of the adaptive CONWIP, as the simulations for the single FOUP type, 100 simulation experiments are performed with 10,000 multiple FOUP types under both the proposed scheduling method and the OPNDD with the deterministic input release control policy. Figure 7 shows the standard deviations and the averages of 100 mean cycle times under the two scheduling cases. In the FAB1, the standard deviations of the mean cycle times are reduced by 20.5, 32.3, and 24.4 % for the FOUP A, B, and C, respectively; in the FAB2, the standard deviations of the mean cycle times are reduced by 23.8, 22.9, and 20.7 % for the FOUP A, B, and C, respectively; in the FAB3, the standard deviations of the mean cycle times are reduced by 17.7, 20.8, and 16.4 % for the FOUP A, B, and C, respectively.

5 Conclusion

This paper proposes heuristic scheduling policies for a semiconductor wafer fabrication facility. The scheduling policies are composed of the advanced OPNDD for the sequence control policy and the adaptive CONWIP for the input release control policy. Although the OPNDD shows good performances in reducing the variation of cycle times in our previous study, it has a limitation that it may cause excessive queue lengths in the bottleneck workstations. To cope with this problem, the advanced OPNDD is proposed to regulate the queue lengths of the FOUPs in each stoker by preventing excessive queue lengths in the bottleneck workstations. In regard to the input release control policy, the adaptive CONWIP is proposed to adjust the WIP level dynamically according to the current status of the wafer fab. The simulation experiments are carried out for both single and multiple types of FOUPs in the TRC fabs. The proposed scheduling method shows the better performances in reducing the standard deviation of cycle times compared

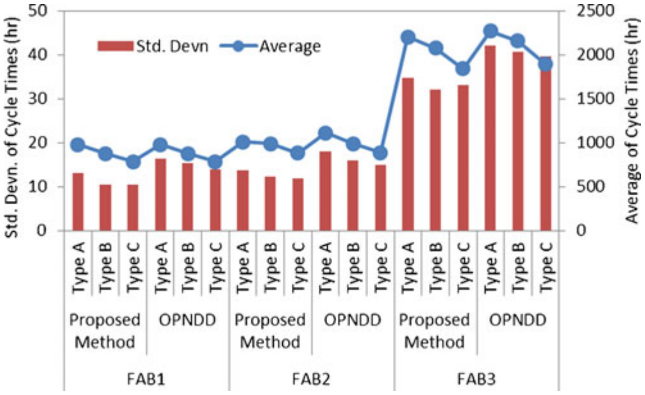

Fig. 7 The standard deviation and the average of mean cycle times with multiple FOUP types under the proposed scheduling method and the OPNDD with the deterministic input release control policy

with the OPNDD, the FSVCT, and other well-known dispatching rules. Also, the simulation results show that the proposed scheduling method is very efficient in reducing the variation of mean cycle times. Compared with the OPNDD under the deterministic input release control policy, the proposed scheduling method reduces the standard deviations of mean cycle times by 19.6~21.8 % in the single-type FOUP simulation experiments, and by 16.4~32.3 % in the multiple-type FOUP simulation experiments.

Acknowledgments This work was supported by research grants from the Catholic University of Daegu in 2011.

References

1. Yoon HJ, Lee DY (2000) A control method to reduce the standard deviation of flow time in wafer fabrication. IEEE Trans Semicond Manuf 13(3):389–392
2. 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):47–61
3. 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 26(5):44–55
4. Wein LM (1988) Scheduling semiconductor wafer fabrication. IEEE Trans Semicond Manuf 1(3):115–130
5. Lu SH, 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
6. Chung SH, Yang MH, Cheng CM (1997) The design of due date assignment model and the determination of flow time control parameters for the wafer fabrication factories. IEEE Trans on Components, Packaging, and Manuf Technol-Part C 20(4):278–287
7. 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
8. Kim YD, Kim JG, Choi B, Kim HU (2001) Production scheduling in a semiconductor wafer fabrication facility producing multiple product type with distinct due dates. IEEE Trans Robot Autom 17 (5):589–598
9. Gupta AK, Sivakumar AI (2006) Job shop scheduling techniques in semiconductor manufacturing. Int J Adv Manuf Technol 27(11–12):1163–1169
10. Gupta AK, Sivakumar AI (2006) Optimization of due-date objectives in scheduling semiconductor batch manufacturing. Int J Mach Tool Manuf 46(12–13):1671–1679
11. Chung SH, Lai CM (2006) Job releasing and throughput planning for wafer fabrication under demand fluctuating make-to-stock environment. Int J Adv Manuf Technol 31(3–4):316–327
12. Upasani AA, Uzsoy R, Sourirajan K (2006) A problem reduction approach for scheduling semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(2):216–225
13. Pfund ME, Balasubramanian H, Fowler JW, Mason SJ, Rose O (2008) A multi-criteria approach for scheduling semiconductor wafer fabrication facilities. J Sched 11(1):29–47
14. Mason SJ, Fowler JW, Carlyle WM (2002) A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shop. J Sched 5(3):247–262
15. Yoon HJ, Shen W (2008) A multiagent-based decision-making system for semiconductor wafer fabrication with hard temporal constraints. IEEE Trans Semicond Manuf 21(1):83–91
16. Zhang H, Jiang Z, Guo C (2009) Simulation-based optimization of dispatching rules for semiconductor wafer fabrication system scheduling by the response surface methodology. Int.J of Manuf Technol 41(1–2):110–121
17. Baez-Senties O, Azzaro-Pantel C, Pibouleau L, Domenech S (2010) Multi-objective scheduling for semiconductor manufacturing plants. Computes and Chem Eng 34(4):555–566
18. Guo C, Jiang Z, Zhang H, Li N (2012) Decomposition-based classified ant colony optimization algorithm for scheduling semiconductor wafer fabrication system. Comput Ind Eng 62(1):141–151
19. Chiang TC, Fu LC (2012) Rule-based scheduling in wafer fabrication with due date-based objectives. Comput Oper Res 39 (11):2820–2835
20. Zhou M (1998) Modeling, analysis, simulation, scheduling, and control of semiconductor manufacturing systems: a Petri net approach. IEEE Trans Semicond Manuf 11(3):333–357
21. Lin MH, Fu LC (2000) Modeling, control and simulation of an IC wafer fabrication system: a generalized stochastic coloured timed Petri net approach. Int J Prod Res 38(14):3305–3341
22. Odrey NG, Green JD, Appello A (2001) A generalized Petri net modeling approach for the control of re-entrant flow semiconductor wafer fabrication. Robotics and Comput Integr Manuf 17(1):5–11
23. Chen JH, Fu LC, Lin MH, Huang AC (2001) Petri-net and GA-based approach to modeling, scheduling, and performance evaluation for wafer fabrication. IEEE Trans Robot Autom 17(5):619–636
24. Jain V, Swarnkar R, Tiwari MK (2003) Modeling and analysis of wafer fabrication scheduling via generalized stochastic Petri net and simulated annealing. Int J Prod Res 41(5):3501–3527
25. Huang B, Sun Y (2005) Improved methods for scheduling flexible manufacturing systems based on Petri nets and heuristic search. J of Control Theory and App 3(2):139–144
26. Liu H, Fung YK, Jiang Z (2005) Modeling of semiconductor wafer fabrication systems by extended object-oriented Petri nets. Int J Prod Res 43(3):471–495
27. Lee YF, Jiang ZB, Liu HR (2009) Multiple-objective scheduling and real-time dispatching for the semiconductor manufacturing system. Comput Oper Res 36(3):866–884
28. Lee JS, Zhou M, Hsu PL (2007) A Petri-net approach to modular supervision with conflict resolution for semiconductor manufacturing systems. IEEE Trans Autom Sci Eng 4(4):584–588
29. Liu H, Jiang Z, Fung YK (2009) Performance modeling, real-time dispatching and simulation of wafer fabrication system using timed extended object-oriented Petri nets. Comput Ind Eng 56(1):121–137
30. Wu Z, Hsieh SJ (2012) A realtime fuzzy Petri net diagnose for detecting progressive faults in PLC based discrete manufacturing system. Int J Adv Manuf Technol 61(1–4):405–421
31. Qiao F, Ma Y, Li L, Yu H (2013) A Petri net and extended genetic algorithm combined scheduling method for wafer fabrication. IEEE Transactions on Automation Science and Engineering (in press)
32. Al-Jaar RY, Desrochers AA (1990) Performance evaluation of automated manufacturing systems using generalized stochastic Petri nets. IEEE Trans Robot Autom 6(6):621–639
33. Marsan MA, Balbo G, Conte G, Donatelli S, Franceschinis G (1995) Modeling with generalized stochastic Petri nets. Wiley, Chichester
34. Cassandras CG (1993) Discrete event systems: modeling and performance analysis. IRWIN, Boston
35. Hopp WJ, Spearman ML (2008) Factory physics. McGraw-Hill, Boston
36. Spearman ML, Woodruff DL, Hopp WJ (1990) CONWIP: a pull alternative to kanban. Int J Prod Res 28(5):879–894