Available online at www.sciencedirect.com
ScienceDirect
Int. J. Production Economics 112 (2008) 416–426
international journal of production economics
www.elsevier.com/locate/ijpe

Scheduling a hybrid MTO/MTS semiconductor fab with machine-dedication features

Muh-Cherng Wu[a,*], Jr-Hsiung Jiang[a], Wen-Jen Chang[b]

[a]Department of Industrial Engineering and Management, National Chiao Tung University, Hsin-Chu, Tawian, ROC
[b]Department of Industrial Engineering and Management, Ta-Hwa Institute of Technology, Hsin-Chu, Tawian, ROC

Received 31 August 2005; accepted 23 April 2007
Available online 16 May 2007

Abstract

A *semiconductor foundry* is essentially a make-to-order (MTO) factory. Yet, in a low-demand season, it may enter into a hybrid business model—producing make-to-stock (MTS) as well as MTO products to maintain high utilization of machines. This research proposes a scheduling method for such a hybrid MTO/MTS system with *machine-dedication* characteristics, a constraint imposed on the process route caused by the advance of manufacturing technology. The scheduling method aims to achieve a high on-time delivery rate for MTO products as well as a high throughput for MTS products. Simulation experiments show that the proposed scheduling method outperforms representative methods in the literature.

Keywords: Make-to-order; Make-to-stock; Machine dedication; Scheduling; Semiconductor manufacturing

1. Introduction

A *semiconductor foundry* is essentially a make-to-order (MTO) factory that manufactures semiconductor products designed by customers. Yet, in low-demand seasons, the MTO orders may be so low that a substantial amount of capacity becomes idle. This may lead to a higher production cost and result in undesirable loss in financial statements because semiconductor manufacturing is very capital-intensive. Some semiconductor foundries may thus include the production of make-to-stock (MTS) products to increase capacity utilization. Such a hybrid production system is called a hybrid MTO/MTS semiconductor fab.

Semiconductor manufacturing has two distinct points: a long process route with *re-entry* characteristics and an ongoing advance of manufacturing technology. The process route of a semiconductor product may involve over 500 operations; a workstation has to process several operations on the same wafer; therefore, a job (also called a lot) has to re-enter a workstation several times. The advance of semiconductor manufacturing technology is measured by the dimension of devices manufactured. The smaller the device dimension, the more advanced the manufacturing technology.

Steppers are types of machines in a fab, which essentially perform the *exposure operations*. An exposure operation is to “photo-print” a circuit

---

*Corresponding author. Fax: +886 3 5720 610.
E-mail address: mcwu@mail.nctu.edu.tw (M.-C. Wu).

0925-5273/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpe.2007.04.008

pattern onto a wafer by light projection through a mask that records the circuit pattern. The resolution of exposure operations determines the smallest dimension that could be manufactured on a device.

In an up-to-date fab, the exposure operations of a wafer have different resolution requirements. A high-resolution stepper that processes high-resolution exposure operations is generally imposed by a *machine-dedication* constraint. The constraint requests that once a wafer has been processed by a high-resolution stepper, its remaining exposure operations have to be processed by the same stepper. Other steppers, even with same engineering specification, cannot process the wafer. The purpose of imposing such machine-dedication constraints is to ensure good manufacturing quality, because any two machines even specified with identical specifications may still be slightly different.

This implies that machines in a *high-resolution* stepper workstation cannot mutually support in capacity. That is, a breakdown of a particular stepper would hold the production of all jobs that have been dedicated to the stepper; other steppers cannot be used to process the jobs. Akcalt et al. (2001) reported that a semiconductor fab, if imposed with machine-dedication constraints, would significantly increase its cycle time, both in mean and in variation.

Scheduling, which includes releasing and dispatching decisions, is an important technique for improving the shop floor performance of a semiconductor fab. Much research on semiconductor scheduling has been published; yet, most studies focused on fabs that run the business at either an MTO or an MTS model. In practice, the key performance measure for MTO products is *on-time delivery rate*, while that for MTS products is *throughput*. These two performance measures are typically against each other. Scheduling methods developed for a single MTO or MTS scenario may not perform as well in a hybrid MTO/MTS scenario.

Chang et al. (2003) proposed a scheduling method for a hybrid MTO/MTS semiconductor fab. By simulation, their method was claimed to outperform some representative methods in the literature. However, the hybrid MTO/MTS environment in their study does not involve any machine-dedication constraint. This constraint tends to reduce the *pooling capacity* of high-resolution steppers, which is the bottleneck of throughput, and cannot be ignored in fabs equipped with advanced manufacturing technology.

This paper presents a scheduling method for a hybrid MTO/MTS semiconductor fab imposed with machine-dedication constraints. The scheduling method involves two shop floor control decisions; releasing and dispatching. By simulation experiments, the proposed method outperforms that proposed by Chang et al. (2003) and some other scheduling methods under a hybrid MTO/MTS scenario with machine-dedication features.

The remainder of this paper is organized as follows. Section 2 reviews the literature on job releasing and dispatching in semiconductor manufacturing. Section 3 describes the proposed method for job releasing. Section 4 presents the methods for job dispatching. Section 5 uses simulation to compare the proposed scheduling method with some other scheduling methods. Concluding remarks are described in Section 6.

2. Literature review

This section reviews the literature on job release and dispatching in semiconductor manufacturing.

2.1. Job release

Previous studies on job release to a semiconductor fab can be grouped into two categories: open-loop and closed-loop methods. Uniform releasing, a typical method of open-loop releasing, inputs a lot periodically and does not concern the dynamic status of the shop floor. In contrast, closed-loop releasing methods input a lot by perpetually reviewing the dynamic WIP status of the shop floor.

Typical examples of closed-loop releasing methods involve CONWIP (Spearman et al., 1990), starvation avoidance (Glassey and Resende, 1988), and workload regulation (Wein, 1988). These closed-loop techniques vary in defining the WIPs to be monitored. The CONWIP method monitors the whole WIPs in the shop floor, while the others monitor only the WIPs of bottleneck machines. Some other releasing methods based on the WIPs of bottleneck machines involve DBR (drum–buffer–rope) (Guide, 1996) and LOMC (load-oriented manufacturing control) (Bechte, 1994; Perona and Alberto, 1998).

In addition to the above methods, some releasing methods considered WIP levels between stages, production surplus, and lot sequencing logic. Examples of such studies include two-boundary, one-stage WIP control, and WIP to bottleneck (Lou and Kager, 1989; Yan et al., 1996, 2000).

The starvation avoidance (SA) technique by Glassey and Resende (1988) is designed to evaluate the *future* WIP status of the bottleneck workstation to see if the bottleneck might become "starved" and would have no WIP for processing. The decision time horizon is determined by the accumulated processing time from the releasing station to the bottleneck. A lot has to be released if a starvation tends to occur; otherwise, no lot is released.

2.2. Job dispatching

A semiconductor fab can be seen as a complicated job shop with characteristics of job re-entry and unexpected machine breakdown (Uzsoy et al., 1992, 1994). Its job scheduling therefore usually takes the form of a dispatching rule, which assigns currently available jobs in queue to a free machine based on the priority index. An extensive review of dispatching rules for job shop scheduling could be found in Panwalker and Iskandar (1977) and Blackstone et al. (1982).

In a semiconductor fab, there are two types of workstations: batch workstations and serial workstations. A batch machine could process several jobs simultaneously, while a series machine processes one job at a time. A dispatching decision for a series machine has to be made whenever a job is completed. For batch scheduling, one needs to group jobs waiting in front of a batch workstation into batches as well as to determine processing sequences and start times for these batches. Batch-scheduling problems have been dealt with in a few studies (Neuts, 1967; Glassey and Weng, 1991; Fowler et al., 1992, 2000).

Various dispatching algorithms for series machines have been proposed. Lu and Kumar (1991) proposed dispatching rules based on job due dates and buffer priorities. Dispatching algorithms aiming to reduce the mean and variance of cycle time for a semiconductor fab were also proposed (Lu et al., 1994). Li et al. (1996) developed a wafer fab-scheduling algorithm based on inventory variability. Kim et al. (1998, 2001) developed scheduling algorithms to meet the due dates of orders. Approaches that suggest the use of composite dispatching rules were also proposed (Lee et al., 2002; Dabbas and Fowler, 2003). Kim et al. (2003) proposed a simplification method for accelerating the speed of simulation while selecting dispatching rules in a real-time manner. Upasani et al. (2006) proposed a problem-reduction approach for the fab scheduling by using global information. Wu et al. (2006) proposed a line-balanced and SA dispatching algorithm for an MTO fab with machine-dedication features.

Previous studies have established significant milestones in the scheduling of wafer fabrication. However, most studies deal with either an MTO or an MTS fab. Very few studies investigate a hybrid scenario that manufactures MTO and MTS products simultaneously, where the key performance measure of MTO products is on-time delivery and that of MTS products is throughput. Chang et al. (2003) proposed a scheduling algorithm for such a hybrid environment. Yet, their work did not consider the machine-dedication feature that an up-to-date fab typically has.

This research aims to develop job releasing and dispatching methods for an MTO/MTS hybrid environment with machine-dedication features. In dispatching decisions, we focus on series machines and adopt the minimum batch size (MBS) method (Neuts, 1967) for batch machines.

3. Releasing method

This research develops a method that adapts SA (Glassey and Resende, 1988) to deal with two distinct fab features: *machine dedication* and *hybrid MTO/MTS production*. The proposed releasing method involves two decisions: (1) releasing time: determining when to release a job to a fab; and (2) releasing priority: determining which lot should be released at a releasing instant.

3.1. Decision of releasing time

With the machine-dedication feature, different machines (high-resolution steppers) in the bottleneck workstation cannot share the workload. Each bottleneck machine must therefore be regarded as an individual workstation in controlling job release. That is, for a fab with n bottleneck machines, it needs n job-releasing control systems.

A procedure *Determine_Releasing_Time* is developed for determining the releasing time for a particular bottleneck machine, say B. To carry out the procedure, it needs a pre-simulation (adopting uniform releasing and FIFO dispatching policies) to estimate L_i and L, where L_i denotes the mean flow time of product i from wafer release to machine B

and $L = \max(L_i)$. Also a workload threshold S for machine B is manually determined, where $S = L/\alpha$, $0 < \alpha < 1$.

Procedure: Determine_Releasing_Time

Step 1: Estimate future arrival of WIP to machine B

- Compute Q, the set of WIPs that are expected to arrive at machine B before time $t+L$, where t is the current time.

Step 2: Estimate the future workload of machine B

$$= T_Q + T_R,$$

where W is the future workload (in time units), T_Q is the aggregate workload of Q on machine B and T_R is the estimated down time of machine B, from t to $t+L$.

Step 3: Releasing decision

If $W \leqslant S$, a lot should be released at time t
If $W > S$, no lot can be released at time t

In the above procedure, if more than one bottleneck is simultaneously "starved" ($W \leqslant S$), the most starved one (i.e., the largest one in $S-W$) has the highest priority to activate the wafer release. Surely, the released lot must be dedicatedly assigned to the particular bottleneck that activates the release signal.

3.2. Decision of releasing priority

In the hybrid MTO/MTS environment, we assume that the due dates and planned release dates of MTO products are known. Define $O(t)$ as the set of MTO jobs that are planned to be released on day t. MTS jobs are not assigned any due dates because the production of MTS jobs occurs only when scheduled MTO products cannot fully utilize the fab capacity. That is, customers of MTS products at the time of production are still unknown and due dates cannot be given.

Two criteria are used to determine the releasing priority of jobs at a releasing instant. First, MTO jobs should be released in a *just-in-time* manner. That is, at day t, the MTO jobs in $O(t)$ must all be released before releasing any MTS job. Moreover, when the jobs in $O(t)$ have been exhaustively released at day t, only MTS jobs are allowed for subsequent release on that day. This policy is intended to increase the throughput of MTS as much as possible while MTO products appear to be on time.

Second, the critical ratio (CR) dispatching rule (Blackstone et al., 1982) is used to prioritize the MTO lots that are available to release. The CR value of a lot is denoted by $\mathrm{CR} = d - t/\sum pt_i$, where d denotes the due date of the lot, t is the present time and $\sum pt_i$ denotes the remaining processing time of the lot. The lower the CR value, the higher dispatching priority is the lot.

3.3. Comparison between proposed releasing algorithm and SA

The aforementioned job-releasing algorithm that deals with both releasing time and releasing priorities is called SA*. Compared to the original SA technique (Glassey and Resende, 1988), the SA* algorithm is distinct in four points due to the inclusion of *machine-dedication* and *MTO/MTS hybrid production features*.

Firstly, in the decision of job priority at a releasing time, we proposed that *MTO jobs* should be released in a *just-in-time* manner, and MTS jobs should be released as much as possible while the just-in-time policy of MTO jobs could be maintained. The algorithmic design attempts to optimize both the objectives of hybrid MTO/MTS production. Such a release control between MTO and MTS products is not addressed in traditional SA.

Secondly, the WIP level of each bottleneck machine is *individually evaluated* rather than evaluated by their *aggregation*. Due to the constraint of machine dedication, different bottleneck machines cannot mutually support in capacity. Therefore, the WIPs waiting before each bottleneck machine should be *individually* considered in the decision for releasing a lot.

Thirdly, *mean cycle time* (L_i), rather than *processing time*, of a product is used to predict the arrival of WIP to bottleneck. This modification is intended to improve the accuracy of estimating future WIP levels. Define *X-factor* as the ratio of cycle time over processing time. In a fab with machine-dedication features, the *X*-factors in different segments of a process route greatly differ. Therefore, the use of cycle time in predicting future WIP arrivals would be much more accurate.

Fourthly, at a releasing instant, the lot to release has to be dedicatedly assigned to the high-resolution stepper that activates the job release.

Such a machine assignment is not required in the traditional SA algorithm.

4. Dispatching methods

This section presents the dispatching methods for two types of series machines: *dedicated* and *non-dedicated*. Dedicated machines (i.e., high-resolution steppers) are the *bottleneck* of a fab because they are relatively expensive and limited in number. The two dispatching procedures and their notation are presented below, where the statistics of cycle times are estimated by the aforementioned pre-simulation program.

Notation

$\sigma_{i,e}$	standard deviation of the cycle time for operation e in lot i
β	a manually determined parameter
$\delta(k)$	the MTO lot that has dispatching rank k, i.e., $r(\delta(k)) = k$
B	the dedicated machine addressed in the dispatching decision
$\sum_{k=1}^{f}\text{CT}_{i,k}$	mean cycle time from wafer release to wafer out, for lot i
$\sum_{k=1}^{e}\text{CT}_{i,k}$	mean cycle time from wafer release to completion of operation e, for lot i
D_i	predefined due date of lot i
e	presently waiting-to-be processed operation of a lot
E_i	release time of lot i
$\text{ES}_{i,e}$	target time for starting operation e of MTO lot i
i*	the one with highest dispatching priority among MTO lots
j*	the one with highest dispatching priority among MTS lots
$\text{OPDD}_{i,e} = E_i + (D_i - E_i)\cdot[\sum_{k=1}^{e}\text{CT}_{i,k}/\sum_{k=1}^{f}\text{CT}_{i,k}]$	due time for operation e in lot i
$\text{PS}_{i,e}$	time for starting operation e in MTO lot i while MTS lot j* is processed first
$\text{PT}_{i,e}$	processing time of operation e in MTO lot i
$\text{PT}_{j^*,e}$	processing time of operation e in MTS lot j*
$\text{PT}_{\delta(k),e}$	processing time of operation e in lot $\delta(k)$
$r(i)$	dispatching rank of MTO lot i
RPT_j	remaining processing time of MTS lot j
$\text{RPT}_{j,b}$	remaining processing time from present operation to next bottleneck for MTS lot j; if there is no bottleneck operation left, $\text{RPT}_{j,b} = \text{RPT}_j$
T_{now}	present time

4.1. Dispatching for dedicated machines

The basic idea for dispatching a dedicated machine is to justify whether prioritizing the processing of an MTS lot will cause any MTO lot to become *operation-delayed*. If so, MTO lots have higher dispatching priority; otherwise, MTS lots do.

Procedure: Dispatching_Dedicated_Machine

Step 1: Compute $\text{ES}_{i,e}$

$$\text{ES}_{i,e} = \text{OPDD}_{i,e} - \text{PT}_{i,e} - \beta\sigma_{i,e}.$$

Step 2: Determine i* and $r(i)$ for MTO lots

- Sort MTO lots based on $\text{ES}_{i,e}$, in ascending order, and compute $r(i)$
- $* = \text{Arg min}(\text{ES}_{i,e})$. That is, lot i* has top dispatching priority and $r(i^*) = 1$

Step 3: Determine j* for MTS lots

$$* = \text{Arg min}(RPT_j).$$

Step 4: Evaluate the start time of each MTO lot i if MTS lot j* is dispatched first

$$\text{PS}_{i,e} = T_{\text{now}} + \text{PT}_{j^*,e} + \sum_{k=1}^{r(i)-1}\text{PT}_{\delta(k),e}.$$

Step 5: Dispatch lot i* or j*

If $\text{PS}_{i,e} < \text{ES}_{i,e}$ for each MTO lot i, then process MTS lot j* .
Otherwise, process MTO lot i*.

The above procedure is summarized as follows. Firstly, MTS lots are prioritized based on SRPT (shortest remaining processing time), which is used here because it has been justified to be effective in increasing throughput (Blackstone et al., 1982). Secondly, MTO lots are prioritized based on $\text{ES}_{i,e}$, the target time to start operation e. The earlier is $\text{ES}_{i,e}$, the more urgent is lot i. Therefore, the use of $\text{ES}_{i,e}$ tends to increase the on-time delivery of MTO lots. Thirdly, the dispatching decision in Step 5 implies that keeping on-time delivery of MTO lots is much more important than increasing the throughput of MTS lots.

4.2. Dispatching for a non-dedicated workstation

The basic idea for dispatching a non-dedicated workstation is to justify whether prioritizing the processing of MTS lot j* will cause MTO lot i* to become *operation-delayed*.

```
**Procedure**: Dispatching_Non-Dedicated_Machine
  *Step 1*: Find MTO lot $i^*$
$\mathrm{ES}_{i,e} = \mathrm{OPDD}_{i,e} - \mathrm{PT}_{i,e} - \beta\sigma_{i,e}$
      $\bullet * = \text{Arg min}(\mathrm{ES}_{i,e})$
  *Step 2*: Find MTS lot $j^*$
$j^* = \text{Arg min}(\mathrm{RPT}_{j,b})$
  *Step 3*: If lot $j^*$ is processed first, estimate its completion time of operation $e$
$C_{j^*,e} = T_{\mathrm{now}} + \mathrm{PT}_{j^*,e}$
  *Step 4*: Dispatching lot $i^*$ or $j^*$
    If $j^*, e < \mathrm{ES}_{i^*,e}$, then dispatch MTS lot $j^*$
    Otherwise, dispatch MTO lot $i^*$
```

In the above procedure, MTO lots are prioritized based on $\mathrm{ES}_{i,e}$ to increase the MTO on-time delivery rate. In contrast, MTS lots are prioritized based on $\mathrm{RPT}_{i,b}$, the remaining processing time to the next bottleneck operation, which is intended to prevent the next bottleneck from being "WIP-starved" and lead to higher throughput.

4.3. Comparison between the two dispatching methods

With some commonalities in basic ideas, the above two dispatching procedures have some distinctions in detail due to consideration of the machine-dedication feature.

In the two algorithms, MTO lots are both prioritized based on the value of $\mathrm{ES}_{i,e} = \mathrm{OPDD}_{i,e} - \mathrm{PT}_{i,e} - \beta\sigma_{i,e}$. As stated, dedicated machines cannot share capacity. A breakdown of a particular dedicated machine would hold production of all lots that have been assigned to it. Therefore, $\sigma_{i,e}$ (standard deviation of cycle time) of dedicated machines tend to be larger than those of non-dedicated machines. This higher variation in cycle time leads us to use a "*highly conservative*" policy in dispatching a dedicated machine to ensure on-time delivery of MTO lots. That is, if any MTO lot may delay production at present operation e, then we cannot process MTS lots.

In contrast, the lower variation in cycle time leads us to use a "*medium-conservative*" policy in dispatching a non-dedicated machine. That is, we can always process MTS lots, except the MTO lot with highest priority (i^*), which may delay production at present operation e. The use of the "*medium-conservative*" policy aims to maximize the supply of MTS lots to the next bottleneck as long as the most urgent MTO lot is on time.

5. Simulation experiments

A discrete-event simulation program is established to compare the performance between the proposed scheduling method and some representative ones.

5.1. Experiment design

The simulation program models a hypothetical semiconductor fab, where the process routes and number of machines are provided by a semiconductor company in industry. The hypothetical fab includes 60 workstations, nine of which are batch type and 51 are series type. High-resolution steppers—fab bottlenecks—are with machine-dedication features. Machine breakdown and repair data are also available, with exponential distributions.

The fab produces two MTS products and nine MTO products, whose processing times are shown in Table 1. We assume that each order of MTO lots involves only one lot and the planned daily release number of MTO lots for each product is known. The due date of MTO products is defined as follows: $i = R_i + U(a,b) \cdot \mathrm{TPT}_i$, where D_i denotes the due date of lot i, R_i denotes its planned release date, TPT_i denotes its total processing time, and $U(a,b)$ represents a uniform distribution where

Table 1
Total processing time (TPT) (hours) of each product

Product	MTS type		MTO type								
	A	B	C	D	E	F	G	H	I	J	K
TPT	399.8	479.9	380.0	408.0	439.7	290.6	322.2	511.5	543.0	574.6	606.2

parameters a and b are manually determined by referring to the mean cycle time provided by the pre-simulation program. In our simulation experiments, we set $a = 1.66$ and $b = 1.71$ for MTO lots.

The proposed scheduling method is compared with eight scheduling methods. These eight benchmarks involve the Chang et al. (2003) method, SA*-FIFO, SA*-SRPT, SA*-CR, SA*-EDD, SA*-LNQ, SA*-FLNQ, and SA*-LTNV. Of these SA*-X methods, SA* denotes the proposed releasing method and X denotes a particular dispatching method.

The Chang et al. (2003) method is adopted because its algorithm was developed particularly for a hybrid MTO/MTS production environment. First-in–first-out (FIFO) is adopted for its ease of implementation and wide application in the industry. The shortest remaining processing time (SRPT) is adopted because Glassey and Resende (1988) found that SA/SRPT performs quite well in their experiments. CR and EDD (earliest due date) are adopted because they are both due-date-based heuristics that attempt to improve on-time delivery. As stated, MTS lots have no due dates in the context of this research. However, to carry out the experiments that adopt CR and EDD rules, we define a *virtual due date* for each MTS lot as follows: $i = r_i + U(2.50, 2.55) \cdot \mathrm{TPT}_i$, where r_i denotes the release time of the MTS lot.

The largest number in queue (LNQ) denotes a dispatching rule that selects the lot with the largest WIP level in queue, which is adopted because it tends to remove "traffic jam" and balance the production line (Wu et al., 2006). Fewest lots at the next queue (FLNQ) gives highest dispatching priority to the lot whose next queue is fewest in WIP, which is adopted because Dabbas et al. (2003) found that FLNQ performs quite well in on-time delivery in their experiments. LTNV denotes a dispatching rule that at a bottleneck station selects the lot which has the longest expected processing time until its next visit to the same station and FIFO is used at all other stations (Wein, 1988; Lu et al., 1994).

Of these SA*-X heuristics, FIFO is arrival-time related; SRPT is processing-time related, CR and EDD are due-date related, LNQ is line-balance related, FLNQ is SA related, and LTNV is bottleneck-control related.

These eight scheduling methods as well as the proposed one are compared in three simulation scenarios. Scenario I produces one MTS product (A) and three MTO products (C–E), where the product mix for MTO is 1:1:1. Scenario II produces the same four products as in Scenario I, but the product mix of MTO becomes 1:1:2. Scenario III produces two MTS products (A, B) and nine MTO products (C–K) as shown in Table 1.

Twenty-seven simulation cases (9 dispatching methods × 3 scenarios) are tested, and each test case runs 20 replicates. The time horizon for a simulation case is 270 days and only data of the last 180 days are collected. For parameters in the releasing and dispatching decisions, we set $\alpha = 0.98$ and $\beta = 0.5$. Notice that the value of β denotes how conservative we are in ensuring the on-time delivery of MTO lots. The higher the value of β, the more conservative we are. That is, a higher value of β tends to increase the on-time delivery of MTO lots at the price of decreasing the throughput of MTS lots.

5.2. *Experimental results*

With the experimental results available, an ANOVA test for each performance metric in each scenario has been carried out. The test results indicate that the scheduling method indeed has a significant effect on each performance metric in each scenario. Table 2 shows some of the test results for the three scenarios in the case of comparing on-time delivery. The results indicate that the scheduling method indeed has a significant effect on on-time delivery.

Table 3 shows the mean and standard deviation of *on-time delivery rate* of the MTO products for the nine scheduling methods. The tables reveal that the proposed method outperforms the other methods in terms of on-time delivery rate in each scenario. Moreover, with a small standard deviation in on-time delivery rate, the proposed method appears quite reliable in providing on-time delivery service.

Tables 4 and 5 compare the dispatching methods in terms of tardiness and cycle time of MTO products, where their means and standard deviations are both revealed. The table indicates that the proposed dispatching method also outperforms the other eight benchmarks in each test scenario for the two performance metrics. Moreover, with a small standard deviation in the two performance metrics, the proposed method appears quite reliable in terms of tardiness and cycle time of MTO products.

Table 2
Results of ANOVA tests for the effect of scheduling on on-time delivery in three scenarios

Source of variation	Sum of square	Degree of freedom	Mean square	*F*	*P*-value
(a) *Scenario I*					
Scheduling	13.769	8	1.721	228.6024	4.16E−87
Error	1.287	171	0.0075		
Total	15.057	179			
(b) *Scenario II*					
Scheduling	12.718	8	1.5897	193.669	1.54E−81
Error	1.403	171	0.0082		
Total	14.121	179			
(c) *Scenario III*					
Scheduling	13.031	8	1.629	265.58	3.25E−92
Error	1.0489	171	0.0061		
Total	14.080	179			

Table 3
Comparing on-time-delivery rate of MTO products

Scheduling methods	Scenario I			Scenario II			Scenario III		
	Mean (%)	Std. dev. (%)	Duncan test	Mean (%)	Std. dev. (%)	Duncan test	Mean (%)	Std. dev. (%)	Duncan test
Proposed	99.80	0.23	A	100.00	0.01	A	99.23	0.55	A
Chang	82.35	5.95	B	55.70	11.16	B	65.38	6.93	B
SA*-CR	79.09	8.34	B	60.26	18.29	B	71.22	17.67	B
SA*-EDD	43.09	19.15	D	18.39	6.07	D	30.15	4.18	D
SA*-SRPT	31.41	6.74	E	37.15	9.65	C	36.31	10.04	D
SA*-FIFO	2.68	1.19	F	3.15	0.97	E	1.13	0.59	E
SA*-LNQ	45.62	7.04	D	40.39	5.26	C	37.89	13.08	D
SA*-FLNQ	56.32	5.54	C	57.26	7.25	B	52.65	4.9	C
SA*-LTNV	60.23	6.69	C	65.29	15.63	B	67.23	12.88	B

Table 4
Comparing tardiness (hours) of MTO products

Scheduling methods	Scenario I			Scenario II			Scenario III		
	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test
Proposed	0.15	0.07	A	0.04	0.03	A	2.25	1.02	A
Chang	4.65	1.90	A	43.82	26.95	AB	58.58	11.02	B
SA*-CR	5.66	2.84	A	50.67	30.80	B	60.77	20.12	B
SA*-EDD	47.69	28.17	B	147.73	26.96	C	138.08	20.48	C
SA*-FIFO	190.27	39.52	D	275.45	36.71	D	311.15	45.38	D
SA*-SRPT	320.97	101.03	E	502.40	200.04	E	509.31	150.28	E
SA*-LNQ	101.66	24.41	C	142.36	27.12	C	155.63	35.34	C
SA*-FLNQ	67.38	16.26	B	78.69	18.44	B	79.23	24.68	B
SA*-LTNV	59.66	13.09	B	50.35	19.04	B	61.25	20.40	B

Table 5
Comparing cycle time (hours) of MTO products

Scheduling methods	Scenario I			Scenario II			Scenario III		
	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test
Proposed	645.06	3.02	A	643.54	2.77	A	729.79	4.18	A
Chang	657.87	12.13	AB	724.66	40.82	BC	778.35	35.44	B
SA*-CR	676.74	10.24	B	708.53	42.61	B	740.21	18.50	A
SA*-EDD	749.22	25.26	C	850.52	37.76	D	786.57	31.87	BC
SA*-FIFO	876.90	34.61	D	897.96	41.33	E	1004.59	44.94	E
SA*-SRPT	916.50	61.11	E	1026.66	109.01	F	1139.74	110.52	F
SA*-LNQ	772.57	32.01	C	825.35	27.85	D	885.91	32.56	D
SA*-FLNQ	749.46	24.83	C	742.63	28.76	C	812.15	21.71	C
SA*-LTNV	742.36	27.68	C	739.24	21.36	C	790.22	30.72	C

Table 6
Comparing throughput for scenario I

Scheduling methods	Aggregated throughput			MTO throughput			MTS throughput		
	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test
Proposed	5335	31.13	A	3949	17.56	A	1386	41.14	B
Chang	5110	39.67	C	3857	22.78	B	1253	38.95	C
SA*-CR	5198	70.34	B	3717	32.20	C	1481	69.62	A
SA*-EDD	5054	56.55	CD	3674	58.18	C	1380	41.72	B
SA*-FIFO	4952	81.18	E	3597	64.37	D	1345	40.05	B
SA*-SRPT	4772	118.97	G	3652	144.39	C	1120	215.94	D
SA*-LNQ	4893	55.70	F	3496	28.69	E	1397	57.91	B
SA*-FLNQ	4899	37.37	F	3500	30.70	E	1399	33.78	B
SA*-LTNV	5058	51.13	D	3668	38.45	C	1390	30.83	B

Tables 6–8 compare the throughputs of dispatching methods for each test scenario, where the throughput of MTO products, that of MTS products, and the aggregated throughput of MTO/MTS products are, respectively, displayed. The tables reveal that the proposed method outperforms the other dispatching methods both in the aggregated throughput and in the throughput of MTO products. Moreover, with small standard deviations in MTO throughput, the proposed dispatching method is also good in providing a stable throughput rate.

However, in terms of the throughput of MTS products, the proposed method also performs well, but is not the best, which by Duncan tests are categorized either in the first group or in the second group. This finding coincides with the idea of the proposed algorithm—paying more attention to MTO products than to MTS products.

From the tables, one might wonder why SA*-CR would perform so well in terms of MTS throughput, which by Duncan tests is categorized in the first group in each scenario. As stated, in the SA*-CR algorithm, we define a *virtual due date* for each MTS lot released to the shop. With the due-date assignment, MTS lots now compete with MTO lots for the utilization of capacity. Compared with the proposed approach, SA*-CR would then tend to decrease the MTO throughput and increase the MTS throughput. As a result, the aggregated throughput of SA*-CR might exceed that of the proposed algorithm if the value of β is set in a highly conservative manner.

6. Concluding remarks

This paper presents a scheduling method for a semiconductor fab that simultaneously produces MTO/MTS products. The fab is distinct in having

Table 7
Comparing throughput for scenario II

Scheduling methods	Aggregated throughput			MTO throughput			MTS throughput		
	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test
Proposed	5200	39.3	A	3670	18.31	A	1530	31.30	A
Chang	4868	64.89	BC	3569	32.44	B	1299	47.34	D
SA*-CR	4978	81.15	B	3477	71.86	C	1501	45.34	AB
SA*-EDD	4820	63.03	C	3289	44.09	D	1531	26.06	A
SA*-FIFO	4767	89.51	D	3239	66.24	E	1528	43.56	A
SA*-SRPT	4201	202.19	F	2798	132.57	F	1403	200.17	C
SA*-LNQ	4693	59.28	E	3205	39.53	E	1433	29.12	C
SA*-FLNQ	4804	70.36	C	3353	41.87	C	1451	32.85	BC
SA*-LTNV	4773	64.02	C	3288	45.94	D	1485	37.19	B

Table 8
Comparing throughput for scenario III

Scheduling methods	Aggregated throughput			MTO throughput			MTS throughput		
	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test	Mean	Std. dev.	Duncan test
Proposed	5103	31.26	A	3698	16.91	A	1405	24.15	BC
Chang	4843	56.7	B	3478	31.06	B	1365	44.11	B
SA*-CR	4823	64.65	B	3337	65.51	C	1486	43.42	A
SA*-EDD	4675	49.34	C	3291	67.48	C	1384	38.54	CD
SA*-FIFO	4507	70.36	D	3200	91.88	D	1307	52.94	E
SA*-SRPT	4001	131.43	E	2709	61.69	E	1292	140.69	E
SA*-LNQ	4699	52.29	C	3215	27.26	CD	1484	30.74	A
SA*-FLNQ	4895	61.08	B	3469	39.46	B	1426	34.23	B
SA*-LTNV	4931	75.80	B	3499	67.07	B	1432	24.14	B

a *machine-dedication* feature, which is caused by the advance of manufacturing technology. The performance measure for scheduling MTO products is on-time delivery rate and that for MTS products is throughput. Scheduling such a hybrid MTO/MTS fab has been studied by Chang et al. (2003); yet, the machine-dedication feature was considered in their study.

The proposed scheduling method involves the decisions of releasing and dispatching. The releasing technique (called SA*) is a modification of the starvation-avoidance algorithm (Glassey and Resende, 1988) by considering the characteristics of the machine-dedication feature and MTO/MTS hybrid production. The basic ideas of the dispatching techniques are to produce MTS products as much as possible subject that MTO lots can be on time in each operation step. This idea is deployed in the dispatching methods for dedicated and non-dedicated machines.

Simulation experiments involving three scenarios and eight benchmarking dispatching methods are performed. Results show that the proposed scheduling method outperforms the eight benchmarking dispatching methods in on-time delivery rate, the throughput of MTO products, and the aggregated throughput of MTO/MTS products.

Acknowledgement

This research was supported by National Science Council, Taiwan, under research contracts NSC90-2622-E009-001 and NSC94-2213-E-009-083.

References

Akcalt, E., Nemoto, K., Uzsoy, R., 2001. Cycle-time improvements for photolithography process in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 14, 48–56.

Bechte, W., 1994. Load-oriented manufacturing control just-in-time production for job shop. Journal of Production Planning and Control 5 (3), 292–307.

Blackstone, J.H., Phillips, D.T., Hogg, G.L., 1982. A state-of-the-art survey of dispatching rules for manufacturing job shop operations. International Journal of Production Research 20, 27–45.

Chang, S.H., Pai, P.F., Yuan, K.J., Wang, B.C., Li, R.K., 2003. Heuristic PAC model for hybrid MTO and MTS production environment. International Journal of Production Economics 85, 347–358.

Dabbas, R.M., Fowler, J.W., 2003. A new scheduling approach using combined dispatching criteria in wafer fabs. IEEE Transactions on Semiconductor Manufacturing 16, 501–510.

Dabbas, R.M., Fowler, J.W., Rollier, D.A., Mccarville, D., 2003. Multiple response optimization using mixture-designed experiments and desirability functions in semiconductor scheduling. International Journal of Production Research 41 (5), 939–961.

Glassey, C.R., Resende, M.G.C., 1988. Closed-loop job release control for VLSI circuit manufacturing. IEEE Transactions on Semiconductor Manufacturing 1, 36–46.

Glassey, C.R., Weng, W.W., 1991. Dynamic batching heuristic for simultaneous processing. IEEE Transactions on Semiconductor Manufacturing 4, 77–82.

Guide, V.D.R., 1996. Scheduling using drum–buffer–rope in remanufacturing environment. International Journal of Production Research 34 (4), 1081–1091.

Fowler, J.W., Hogg, G.L., Philips, D.T., 1992. Control of multi-product bulk services diffusion/oxidation processes. IIE Transactions 24, 84–96.

Fowler, J.W., Hogg, G.L., Philips, D.T., 2000. Control of multi-product bulk server diffusion/oxidation processes: part 2: multiple servers. IIE Transactions 32, 167–176.

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

Kim, Y.D., Kim, J.G., Choi, B., Kim, H.U., 2001. Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. IEEE Transactions on Robotics and Automation 17, 589–598.

Kim, Y.D., Shim, S.O., Choi, B., Hwang, H., 2003. Simplification methods for accelerating simulation-based real-time scheduling in a semiconductor wafer fabrication facilities. IEEE Transactions on Semiconductor Manufacturing 16 (2), 290–298.

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

Li, S., Tang, T., Collins, D.W., 1996. Minimum inventory variability scheduler with applications in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 9, 1–5.

Lou, S.X.C., Kager, P.W., 1989. A robust production control policy for VLSI wafer fabrication. IEEE Transactions on Semiconductor Manufacturing 2, 159–164.

Lu, S.C.H., Kumar, P.R., 1991. Distributed scheduling based on due date and buffer priorities. IEEE Transactions on Automatic Control 36, 1406–1416.

Lu, S.C.H., Ramaswamy, D., Kumar, P.R., 1994. Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. IEEE Transactions on Semiconductor Manufacturing 7 (3), 374–388.

Neuts, M.F., 1967. A general class of bulk queue with Poisson input. Annals of Mathematical Statistics 38, 759–770.

Panwalker, S.S., Iskandar, W.W., 1977. A survey of scheduling rules. Operations Research 25 (1), 45–61.

Perona, M., Alberto, P., 1998. The impact of parameters setting in load oriented manufacturing. International Journal of Production Economics 55, 133–142.

Spearman, M.L., Woodruff, D.L., Hopp, W.J., 1990. CONWIP: a pull alternative to Kanban. International Journal of Production Research 19, 481–490.

Upasani, A.A., Uzsoy, R., Sourirajan, K., 2006. A problem reduction approach for scheduling semiconductor wafer fabrication facilities. IEEE Transactions on Semiconductor Manufacturing 19 (2), 216–225.

Uzsoy, R., Lee, C.Y., Martin-Vega, L.A., 1992. A review of production planning and scheduling models in the semiconductor industry, part I: system characteristics, performance evaluation and production planning. IIE Transactions on Scheduling and Logistics 24, 47–61.

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

Wein, L.M., 1988. Scheduling semiconductor wafer fabrication. IEEE Transactions on Semiconductor Manufacturing 1, 115–130.

Wu, M.C., Huang, Y.L., Chang, Y.C., 2006. Dispatching in semiconductor fabs with machine-dedication features. International Journal of Advanced Manufacturing Technology 28 (9), 978–984.

Yan, H., Lou, S., Sethi, S.P., Gardel, A., Deosthali, P., 1996. Testing the robustness of two-boundary control policies in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 9 (2), 285–288.

Yan, H., Lou, Sethi, S.P., 2000. Robustness of various production control policies in semiconductor manufacturing. Production and Operations Management 9 (2), 171–183.