Computers & Operations Research 39 (2012) 2820–2835

Contents lists available at SciVerse ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Rule-based scheduling in wafer fabrication with due date-based objectives

Tsung-Che Chiang [a], Li-Chen Fu [b,c,*]

[a] Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC
[c] Department of Electrical Engineering, National Taiwan University, Taiwan, ROC

ARTICLE INFO

Available online 24 February 2012

Keywords:
Semiconductor manufacturing
Wafer fabrication
Scheduling
Dispatching rules
On-time delivery
Tardiness

ABSTRACT

Wafer fabrication is a capital-intensive and highly complex manufacturing process. In the wafer fabrication facility (fab), wafers are grouped as a lot to go through repeated sequences of operations to build circuitry. Lot scheduling is an important task for manufacturers to improve production efficiency and meet customers' requirements of on-time delivery. In this research we propose a dispatching rule for lot scheduling in wafer fabs, focusing on three due date-based objectives: on-time delivery rate, mean tardiness, and maximum tardiness. Although many dispatching rules have been proposed in the literature, they usually perform well in some objectives and bad in others. Our rule implements good principles in existing rules by means of (1) an urgency function for a single lot, (2) a priority index function considering total urgency of multiple waiting lots, (3) a due date extension procedure for dealing with tardy lots, and (4) a lot filtering procedure for selecting urgent lots. Simulation experiments are conducted using nine data sets of fabs. Six scenarios formed by two levels of load and three levels of due date tightness are tested for each fab. Performance verification of the proposed rule is achieved by comparing with nine benchmark rules. The experimental results show that the proposed rule outperforms the benchmark rules in terms of all concerned objective functions.

1. Introduction

In recent years, the number of applications and demand for integrated circuits has increased dramatically. Microprocessors, memory chips, and other semiconductor devices are now a part of our daily lives, appearing everywhere in computer, communication, and consumer products. The semiconductor manufacturing process consists of four phases: wafer fabrication, wafer probe, packaging, and final testing. Among them, wafer fabrication is the most complex and costly one. In a wafer fabrication facility (fab), wafers are grouped and put into a container, usually called a *lot*. Each lot goes through repeated sequences of operations including diffusion, photolithography, etching, ion implanting, etc., to build up layers of circuitry on wafers. These operations are complicated and need high technology, and thus the equipment is usually very expensive. Gupta et al. [1] and Pfund et al. [2] mentioned that a large portion of capital cost in the wafer fab is due to the cost of manufacturing equipment. The high cost of equipment prohibits manufacturers from buying equipment and increasing manufacturing capacity unlimitedly, and it also forces production managers to utilize the equipment effectively in order to achieve high production efficiency (e.g. short mean cycle time) and meet customers' requirements (e.g. high on-time delivery rate). Hence, scheduling, which refers to the allocation of equipment over time to lots to optimize the concerned performance measures, becomes an important task in wafer fabs.

A wafer fab is usually viewed as a job shop with the following extensions:

(1) *Reentrant process flow*: Lots may visit the same station more than once;
(2) *Dynamic job arrival*: Customer orders may arrive at different times;
(3) *Parallel machines*: More than one station is able to process one operation;
(4) *Batch processes*: In some stages (e.g. diffusion) more than one lot can be processed simultaneously;
(5) *Sequence dependent setup*: In some stages (e.g. ion implanting) a setup time is required between two operations with different recipes;
(6) *Machine failure*: A station might not be available for an uncertain duration.

Since scheduling in a classical job shop is already known to be NP-hard for many performance measures like mean tardiness [3],

* Corresponding author at: Department of Computer Science and Information Engineering, National Taiwan University, No. 1 Roosevelt Road, Sec. 4, Taipei, Taiwan 106, ROC.
E-mail addresses: tcchiang@ieee.org (T.-C. Chiang), lichen@ntu.edu.tw (L.-C. Fu).

0305-0548/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2012.02.014

Nomenclature

A_i	time at which lot i is released into the fab
d_i	due date of lot i
e_i	number of due date extension of lot i
p_i	processing time of the imminent operation of lot i
P_i	sum of processing time of all operations of lot i
Q	queue of the station that becomes available
R_i	sum of processing time of the unfinished operations of lot i
r_i	time at which lot i arrives at the current station
s_{ij}	sequence dependent setup time required for processing lot j right after lot i
t	system time, the time at which the dispatching decision is to be made
Z_i	index value of lot i

scheduling in a wafer fab is much more difficult. The challenging complexity and practical value of fab scheduling attracted researchers in the academia and practitioners in the industry, and several kinds of scheduling approaches have been developed in the last decades [4]. Dispatching rules are one of the most popular approaches in the industry due to its ease of implementation, computational efficiency, convenience to deal with dynamic environments, and flexibility to incorporate domain knowledge and expertise. It has been applied for fab scheduling successfully by many real-world companies, including Siemens [5], Motorola [6], Samsung [7], IBM [8], and Agere Systems (now in LSI corporation) [9].

With Application Specific Integrated Circuit (ASIC) and specialty processors gaining more and more market share, the capability of meeting due dates is becoming a critical factor in the low-volume, high-variety, and make-to-order wafer fabs. Several studies indicated that today's wafer fabs have been forced to become increasingly conscious of their due date delivery performance [1,2]. In this study, we aim at developing a dispatching rule for the scheduling of wafer fabs with respect to due date-based measures. The proposed rule is distinguished from the existing ones in the use of group information of competing lots and a due date extension procedure. Its performance is verified by simulation experiments using nine fab models and nine benchmark rules in the literature. The concerned performance measures include on-time delivery (OTD) rate, mean tardiness, and maximum tardiness. They are defined by

$$OTD\ \text{rate} = \frac{1}{|N|}\sum_{i\in N} U(i), \quad U(i) = \begin{cases} 0 & \text{if } C_i > d_i, \\ 1 & \text{if } C_i \le d_i, \end{cases} \tag{1}$$

$$\text{mean tardiness} = \frac{1}{|N|}\sum_{i\in N} \max\{C_i - d_i, 0\}, \tag{2}$$

$$\text{maximum tardiness} = \max_{i\in N} \max\{C_i - d_i, 0\}, \tag{3}$$

where N denotes the set of finished lots, and C_i and d_i denote the completion time and due date of a job i, respectively. The rest of this paper is organized as follows. Section 2 gives a review of related work, and Section 3 details the proposed rule. The simulation model and experimental setting are presented in Section 4. Experimental results and discussion are provided in Section 5. Finally, Section 6 gives the conclusion and future research directions.

2. Literature review

A dispatching rule is usually a simple mathematical equation or a short algorithm for calculating priority indices for jobs. It is often invoked when a station finishes a job and becomes available to process the next job. The rule assigns priority indices for waiting jobs, and the job with the highest priority (sometimes the highest index value and sometimes the lowest index value, depending on the rule) is taken as the next processing target. Some reviews of dispatching rules can be found in Panwalker and Iskander [10], Blackstone et al. [11], Rajendran and Holthaus [12], Jayamohan and Rajendran [13], and Sarin et al. [14]. The key points in designing a dispatching rule is what attributes are included and how they are combined if there are more than one attribute.

Early research studies on fab scheduling were usually concerned about cycle time-based performance measures. The first way to design rules for fab scheduling is to borrow ideas from rules for classic flow shop or job shop scheduling. Inspired by the classic least slack (SLACK) rule, Lu et al. [15] proposed the FSVCT and FSMCT rules. The FSVCT rule is like the SLACK rule but replaces the term of due date with the lot arrival time. In this way, the cycle time is equal to the lateness, and hence the FSVCT rule can reduce the variation of cycle time just like the SLACK rule can reduce the variation of lateness. The FSMCT rule is a little more complex. It aims at reducing the burstiness of arrivals of lots to each buffer and consequently reducing the mean cycle time. Based on the idea of the operation due date (ODD) rule, Yoon and Lee [16] developed a rule by allocating the desired cycle time to operations according to the utilization rate of the corresponding stations and then assigning due dates to operations. The lot with the earliest due date of the imminent operation is the next one to be processed. Their rule outperformed the rule proposed by Lu et al. in terms of standard deviation of cycle times. Bahaji and Kuhl [17] proposed four rules based on two rules proposed by Rajendran and Holthaus [12]. They introduced the X factor, the ratio of the accumulated flow time to the sum of the processing time of completed operations, in the new rules. Considering mean cycle time and OTD rate, their recommended rule is the one which combines shortest-processing-time rule (SPT), lowest-work-in-next-queue rule (WINQ), and the X factor.

With Little's Law in queueing theory [18], several researchers noticed the relationship between cycle time and inventory (or work-in-process (WIP)) and devised rules based on WIP information. Li et al. [19] proposed the MIVS rule by introducing correlation between inter-arrivals and services to reduce variability. The rule separates lots into four classes according to the deviation of current inventory from the average inventory. The basic principle is to expedite the lots with excessive inventory at the current stage and to postpone those with excessive inventory at the downstream stage. Lee et al. [20] addressed lot release and lot scheduling by two push-type and two pull-type rules. The push-type rules calculate the deviation of current WIP level from the planned WIP level for each layer of each device and process the lots of layers with more excess WIP level at higher priority. The pull-type rules calculate a due date for each layer of each device by averaging the sum of due dates of lots at the layer over the current WIP level. Then, this layer due date is used in the SLACK rule and ATC rule [21]. By comparing the two pull-type rules with two push-type rules, the pull-type rules showed better performance on several performance criteria including cycle time and machine utilization. Duwayri et al. [9] proposed a rule for