Computers & Industrial Engineering 96 (2016) 72–85

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Synchronized production planning and scheduling in semiconductor fabrication

Sun Hoon Kim *, Young Hoon Lee

Department of Information and Industrial Engineering, Yonsei University, 50 Yonsei-ro, Seodaemoon-gu, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form 12 March 2016
Accepted 19 March 2016
Available online 29 March 2016

Keywords:
Production control
Flexible manufacturing systems
Planning
Scheduling
Linear programming
Simulation

ABSTRACT

Semiconductor fabrication line runs with several hundreds of steps on several hundreds of equipment in the type of re-entrant flow process. The hierarchical planning decisions are made in a way that the production planning are determined on the integrated models at first, and the scheduling operations are then performed. The most complexity comes from the difficulties of synchronization of the production planning and scheduling. An iterative approach to achieve the synchronization is suggested to coordinate the input and output quantity of the production plan when generating a schedule. In the proposed algorithm, not only the input and output quantity of the production plan but also the production quantity is utilized as a coordination factor in the scheduling. Furthermore, the manufacturing lead time, the number of setup events, and the available work-in-process (WIP) level are updated through an iterative procedure of simulation and optimization for the synchronization of planning and scheduling decision. Computational experiments show that interaction framework yields good performance in terms of feasibility, demand satisfaction, and the manufacturing lead time. It is shown that a production planning model can be generated with practical parameters and can be implemented on the scheduling and dispatching levels.

1. Introduction

This study investigates the integration of production planning and scheduling for flexible fabrication systems in consideration of stochastic characteristics such as the moving time and the processing time. Wafer fabrication consists of the repeated sequential processes called a layer: the similar pattern of process steps composed of cleaning, etching, deposition/oxidation, ion implant and photolithography, where a wafer re-enters each steps several times. The sequence of steps may differ depending on the different technology and different device type (Johri, 1993). Semiconductor wafer fabrication is one of the most complex manufacturing systems due to reentrancy, a number of processing steps involved, and the long manufacturing lead times required. Furthermore, it is generally flexible with uncertainties; a certain type of device can be operated on one of several resource candidates with non-identical processing times, with a probabilistic processing time and moving time. It is still difficult, therefore, to develop an accurate production control decision to be implemented as given.

The most important production control decisions in wafer fabrication are known release planning, production planning and scheduling. Decision processes are heavily correlated, hence the coordination of these decisions should be considered for achieving the successful production control (Liao, Chang, Pei, & Chang, 1996). One of the ways to reach at the coordination is to generate a production planning and scheduling in one decision-making framework. This paper suggests an iterative approach to integrate production planning and scheduling in a coordinated manner. Simulation test is used to check the feasibility of the production plan.

There have been a great deal of studies on the planning and scheduling of semiconductor fabrication processes. In the literature, fabrication planning has been modeled in two ways: release planning and production planning. Release planning is an approach that serves to determine the input plan of a device in an effort to improve several performance metrics, such as the cycle time (Wein, 1988), the utilization of a bottleneck resource (Glassey & Resende, 1988a, 1988b), or the balance between the WIP and the throughput (Spearman, Woodruff, & Hopp, 1990). Leachman and Carmon (1992) suggested the use of "the step-separated formulation" for the determination of the amount of the input to the manufacturing line using alternative resources, in consideration of the lead time and capacity per process step. Furthermore, Leachman (2001) suggested the extended models of the step-separated formulation on a different manufacturing and market environments. It should be mentioned that these models assume that no

* Corresponding author. Tel.: +82 (2)2123 4813; fax: +82 (2)364 7807.
E-mail address: misia789@hanmail.net (S.H. Kim).

http://dx.doi.org/10.1016/j.cie.2016.03.019
0360-8352/

intermediate accumulation of work-in-process (WIP) is allowed. Production planning determines the production quantity of resources in each time period, which also covers release planning in general. Most production planning models were developed using linear programming or mixed integer programming (Lee & Kim, 2002; Lee & Lee, 2003; Habla, Monch, & Driebel, 2007; Kang & Lee, 2007; Lee, Lee, Yang, & Ignisio, 2008; Ponsignon & Mönch, 2012). In these models, some details at the operational level are assumed fixed, which may result in a gap between the production planning and scheduling, hence a difficulty in implementation. For example, the lead time is a well-known parameter depending on the workload of the resource, which is referred to as load-dependent lead time (LDLT). Asmundsson, Rardin, and Uzsoy (2006) suggested a production planning model using LDLT. It utilized a concave clearing function that shows the relationship between the available WIP level and the workload. Different dispatching rules and a clearing function are tested through the experiment, and it is found that the dispatching rule that seeks agreement between the plan and the schedule yield the best performance. Their model shows the superiority of reflecting the lead time compared to the other models (Kacar, Irdem, & Uzsoy, 2012); however, it cannot be applied to a flexible manufacturing system such as a semiconductor fabrication line.

On the other hand, scheduling upon fabrication also has received a great deal of attention in areas such as dispatching and shift scheduling. Dispatching is a decision-making process that is used to determine which job should be done next depending on idle resources by creating priority ratings of jobs. Several dispatching rules have been suggested for semiconductor wafer fabrication with several objectives. These include cycle time reduction (Wein, 1988; Lu, Ramaswamy, & Kumar, 1994), demand satisfaction (Kim, Kim, Lim, & Jun, 1998; Kim, Kim, Choi, & Kim, 2001; Kim, Shim, Choi, & Hwang, 2003; Rose, 2003), profit maximization (Hsieh & Hou, 2006; Pierce & Yurtsever, 2000), and rules pertaining to multiple objectives (Hsieh, Chen, & Chang, 2001; Lee, Park, & Kim, 2002; Lee, Jiang, & Liu, 2009). These dispatching rules determine the short-term scheduling decision, while the shift scheduling models work for the relatively long-term scheduling decision. Shift scheduling determines simultaneously the production volume and its sequence in each time shift under a device change setup. Kim, Yea, and Kim (2002) suggested a mixed integer programming model for the shift scheduling of steppers, where a rolling horizon framework is employed due to the complexity of the problem. Lee and Kim (2011) proposed a rule-based scheduling algorithm using a proper WIP.

In summary, production planning for fabrication has been modeled as a decision-making framework to allocate demand requirements depending on resource capacity, which is usually focused on the optimal uses of resources over a medium-term horizon. It has the strength of allowing overall decisions for medium-term or long-term periods; however, several details at the operational level may not be considered. In the scheduling procedure, jobs are to be sequenced for processing on each resource over a short-term time horizon considering the arrival time, the processing time, and the setup. In contrast to planning, scheduling level has to include the decisions in details at the operational level; however, the scope of decision-making at the global level is not precise enough to accomplish optimality. Therefore, the interconnection between two decision processes has a major effect on the quality of fabrication management, as they can complement each other with regard to their different strengths and weaknesses.

The integration of planning and scheduling for fabrication has been studied in hierarchical or iterative structures, as summarized in Table 1. From a hierarchical approach, production planning has generally played a role at the higher level decisions, while scheduling is carried out at a lower level. Adl, Rodriguez, and Tsakalis (1996) suggested a hierarchical model for semiconductor fabrication in which a linear flow model is utilized for decisions at higher level, while a tracking controller is for lower level decisions. Vargas-Villamil, Rivera, and Kempf (2003) proposed a three-layer hierarchical approach; an adaptive layer, an optimization layer, and a direct control layer. The adaptive layer provides parameters of production planning, while the optimization layer works for production planning. At the bottom layer, a distributed control policy is implemented on discrete-event simulation to pursue the target of the optimization layer, with performance measure such as the utilization of the bottleneck resource.

A critical issue pertaining to these approaches is how implementable the plan is at the scheduling level, which is referred to the feasibility of the production plan. For example, when the feasibility of the production plan is high, the information gaps between the plan and the schedule are minor with a higher probability to fulfill the plan at the operation level. Otherwise, the plan and the schedule generated based on it may have great differences in their values, which implies that the plan is meaningless for scheduling. In an iterative approach, when the differences in the values between the plan and its schedule are large, the plan is regenerated, absorbing the scheduling results to improve the feasibility of the plan such that the gaps between them can be reduced. Hung and Leachman (1996) suggested an iterative framework employing linear programming and simulation for release planning as part of the semiconductor fabrication process. The optimization stage of their model creates a release schedule with a given lead time, and the simulation is performed to estimate the lead time with the input plan issued by the optimization. Kim and Kim (2001) also proposed an iterative framework similar to that proposed by Hung and Leachman (1996) for release planning, where one of the remarkable differences is the addition of utilization data to the feedback parameter from the simulation to the optimization process. These two models focus on capturing LDLT in release planning, while Bang and Kim (2010) suggested an iterative production planning model that sought to create an efficient product group, in which the devices in the same product group use the same type of mask. In the model proposed by Bang and Kim (2010), the release schedule and the output plan resulting from the planning stage are employed as dispatch target such that the release schedule is converted to the input time of the wafers and the output plan is transformed to the due date of the wafers. Moreover, in their model, the simulation model with dispatching rules proposed by Kim et al. (2001), ES/RW2 and MDBH, is run to evaluate feasibility of the plan and to obtain data that is harnessed for regrouping the product type. These approaches exploit First In First Out (FIFO) rule or Earliest Due Dates (EDD) rule incorporating only the release and output schedule in the production plan when a schedule is made, and they do not use the production quantity that shows resource utilization for processing a certain type of device in a specific time period. This makes it difficult to apply these models to a flexible manufacturing system with non-identical processing times. For example, when the FIFO rule is used at the dispatching level of the iterative model, the selection of the wafer that is going to be operated on a specific idle resource is determined based on the input time and not on productivity. Thus, when the input time of a specific wafer occurs earlier than that in other wafers, the specific device would be selected even if its productivity is worse than the others. This example is also applicable in regard to the EDD rule. If resources in a specific group have identical processing times, it is not important to consider their productivity at the dispatching level. However, in other occasions, productivity should be considered. One of the ways adopted to handle this issue is to determine a wafer to be operated on an idle resource, based on the production quantity issued by the optimization model. This requires a new dispatching rule to generate priority based on the production quantity.