J. Soc. Korea Ind. Syst. Eng Vol. 38, No. 4 : 159-167, December 2015
http://dx.doi.org/10.11627/jkise.2015.38.4.159
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Two-Level Hierarchical Production Planning for a Semiconductor Probing Facility

June-Young Bang†

Sungkyul University, Gyeonggi-do, Korea

반도체 프로브 공정에서의 2단계 계층적 생산 계획 방법 연구

방 준 영†

성결대학교 산업경영학부

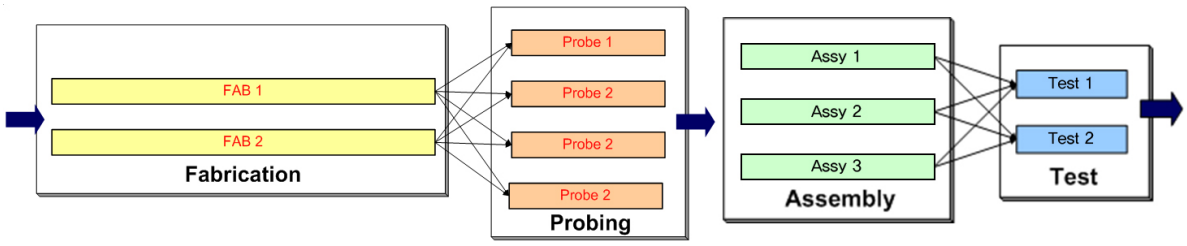
We consider a wafer lot transfer/release planning problem from semiconductor wafer fabrication facilities to probing facilities with the objective of minimizing the deviation of workload and total tardiness of customers' orders. Due to the complexity of the considered problem, we propose a two-level hierarchical production planning method for the lot transfer problem between two parallel facilities to obtain an executable production plan and schedule. In the higher level, the solution for the reduced mathematical model with Lagrangian relaxation method can be regarded as a coarse good lot transfer/release plan with daily time bucket, and discrete-event simulation is performed to obtain detailed lot processing schedules at the machines with a priority-rule-based scheduling method and the lot transfer/release plan is evaluated in the lower level. To evaluate the performance of the suggested planning method, we provide computational tests on the problems obtained from a set of real data and additional test scenarios in which the several levels of variations are added in the customers' demands. Results of computational tests showed that the proposed lot transfer/planning architecture generates executable plans within acceptable computational time in the real factories and the total tardiness of orders can be reduced more effectively by using more sophisticated lot transfer methods, such as considering the due date and ready times of lots associated the same order with the mathematical formulation. The proposed method may be implemented for the problem of job assignment in back-end process such as the assignment of chips to be tested from assembly facilities to final test facilities. Also, the proposed method can be improved by considering the sequence dependent setup in the probing facilities.

Keywords : Hierarchical Production Planning, Semiconductor Probing Facility, Lagrangian Relaxation Method

1. Introduction

As the wafer circuit design faces the limitation of technical obstacle such as the width of circuit lines, semiconductor manufacturing companies should find their competitiveness over other companies in the efficiency of wafer production and high service level. Without massive investment, the effectiveness of production can be obtained by selecting and constructing appropriate planning and scheduling architecture. Innovation and/or continuous improvement in production/operations management are also needed for the companies to increase their market share by meeting customers' demands which is denoted as wafer quantity and due date needed to be met. Furthermore, reducing the deviation of

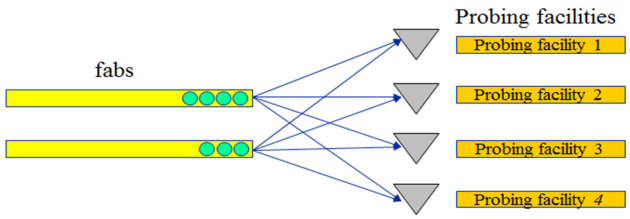
Received 15 October 2015; Finally Revised 7 December 2015; Accepted 8 December 2015
† Corresponding Author : jybang@sungkyul.ac.kr



<Figure 1> Semiconductor Manufacturing Process

workload in each production facilities can reduce the total production lead times and the required number of machines and operators. In general, the production control of non-memory-type semiconductors, such as system large-scale integrated-circuits (LSI), application-specific integrated circuit (ASIC) products and application processors (AP), CPU and GPU for mobile phones, are known to be more difficult than production control of memory-type products, such as DRAM, SSD, Flash memory. This is because non memory type semiconductors are relatively expensive and have many product types, and their orders are in low-volume and tough due-dates. However, in these days, major customers, such as APPLE, IBM, DELL, request special test and customized chip design even for the memory type semiconductor. This increases the difficulty of production control as hard as that of non-memory type semiconductor.

Most semiconductor manufacturing process is composed of four major stages; wafer fabrication, electric die sorting (EDS) or wafer probing, chip assembly, and chip test as depicted in <Figure 1>. A semiconductor wafer goes through a series of fabrication steps in a Fab to form a large number of ICs on its face (for 30~40 days) and stays in a Probe line to be probed for possible defects (for 2~5 days). Then the chips are put into an IC package in the back end lines (for 3~7 days). In the wafer fabrication, the equipment for photolithography is highly expensive (over than 100 billion dollar per one machine) and wafers should re-enter this equipment several times (around 20~30 times). There are hundreds of operations with complicated production characteristics such as main equipment constraints and auxiliary resources, reentrant flows, waiting time constraints and sequence-dependent setup times [5, 9]. For this reason, wafer fab should operate in push strategy to maximize the utilization of equipment. However the last two processes, chip assembly and chip test process should be operated in pull strategy to meet the customer's order. Therefore, the second process, wafer probing (or EDS), should process all the lots from the wafer fabrication facilities and meet the required amount of wafers in time the assembly needed. Usually, we call the wafer probing facilities as a decouple point of Push-Pull strategy. Therefore, as the transition point of push-pull strategy, the lot transfer/release plan plays one of the key roles to achieve effective and efficient production among the entire semiconductor manufacturing.



<Figure 2> Wafer Lot Transfer/Release Problem between Fabs and Probing Facilities

In this research, we consider a wafer lot transfer problem in which wafer lots are moved from fab facilities to the following probing facilities with the objective of minimizing the total tardiness of customers' orders as depicted in <Figure 2>. We propose a two-level hierarchical production planning method for a lot transfer problem between two parallel facilities to obtain executable production plan and schedule. In the higher level, the solution of the reduced mathematical model with Lagrangian relaxation method can be regarded as a coarse good lot transfer plans with a daily time bucket, and the discrete-event simulation is performed to obtain detailed lot processing schedules at the machines with a priority-rule-based scheduling method and the lot transfer (release) plan is evaluated in the lower level.

Many researchers have been focused on scheduling problems for wafer fabrication facilities. However, only a few researches are published for scheduling problem for the wafer probing or electric die shorting (EDS). Some groups of researchers including Chen et al. [5], Chen and Hsia [6], Pearn et al. [14, 15] and Yang et al. [20], and Ellis et al.

[7] deal with the problem of wafer probing facilities. Ovacik and Uzsoy [13] propose rolling horizon heuristics for an identical parallel machines scheduling problem with objective of minimizing maximum lateness with sequence dependent setup times and dynamic job arrivals. Liu and Chang [12] suggest an approach for production scheduling of flexible flow shops with significant sequence-dependent setup effects. Pearn et al. [14, 15] and Yang et al. [20] transform the scheduling problem with sequence-dependent setup into the vehicle routing problem with time windows. Pearn et al. [11] also present heuristic algorithms with job insertions and Pearn et al. [16] adopt saving vehicle-routing heuristics and addition heuristics with improvement. Ellis et al. [7] propose the model of the scheduling problem in wafer probing facility with objective of minimizing makespan and present heuristic algorithms, and Jang [8] considers the capacity of plant and generates the outsourcing production plan. Recently, Bang and Kim [2] propose the heuristic algorithms for probing facility scheduling for the objective of minimizing total tardiness of customer's orders, and Seo and Bang [18] proposed capacitated lot-order pegging strategies for semiconductor manufacturing.

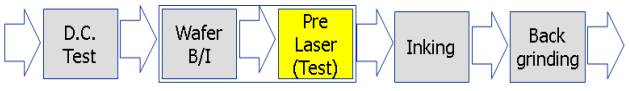
The lot transfer problem between multi-facilities is similar to the multi-level capacitated lot-sizing problem (MLCLP). In MLCLP, the lots-sizes must be determined for multi-level production inventory systems with capacity constraints on the production facilities. Brahimi et al. [3] thoroughly reviews the single time lot-sizing problems for uncapacitated and capacitated versions. Trigerio et al. [19] develops a promising heuristic to solve large-scale capacitated lot-sizing problems based on Lagrangian approach. Lin and Chen [11] propose a mathematical model of a multi-stage and multi-site production planning problem based on TFT-LCD industry. Aghezzaf [1] proposes a mixed integer programming model for mold transfer problem between plants, and develops a linear programming-based heuristic that combines Lagrangian relaxation and linear programming duality for solving the problems.

In this research, we propose the hierarchical production planning method to determine which probing facilities wafer lots are to be processed in the wafer lot transfer problem. This decision is made based on the production information such as the production load of probing facilities, and the completion time of lots in the fabs. Computational experiments are done to evaluate performance of proposed planning method compared with the method which is used in the real production system.

2. Problem Statement

In the probing facilities, ready times of wafer lots are distributed in wide range since the wafer fabrication, the former process of wafer probing, is very complex and time-consuming process, and the lead time of wafer fabrication process is very long and wafer lots will be processed in more than 300 operations. Moreover, customer changes their orders even after their lots are released to wafer fabs. In this case, the wafer lots which are assigned to the canceled orders will be re-assigned to other orders, and their due date and priority will be also changed according to the newly assigned orders. For above reason, the expected completion times of lots at the fab and the ready times of lots at the probing facility assigned to the same order are widely distributed, although the operators release lots with one order to the fab at the same day. The difference of ready time of the first lot and the last lot assigned to the same order ranges less than a day to a week.

After finishing fabrication process, multiple types of wafers proceed to a set of test stations in a semiconductor wafer probing facility as depicted in <Figure 3>. During the wafer-level test operations, each die on a wafer is tested by performing the electrical and functional tests, and the dies are decided as good or bad. The wafer burn-in test is an operation to give electrical stress to wafers and test the electric performance of the die on the wafer being tested such as dielectric strength or surface resistivity. In probing test (called as pre-Laser in the semiconductor factory), the die is tested for the basic function of the product, and the recipes of the probing tests are different for the different product types. For dynamic random access memory (DRAM) product, for example, the dies on the wafer is tested by writing a specific number on the die, reading the value from the die after the specified time, and verifying if the die can store the value properly. The test is performed under the predetermined temperature for each product type. The temperature range is 60~125°C, and the temperature is closely related to the working temperature of the final product, i.e., the temperature the die is used.



<Figure 3> Sub-Processes of a Semiconductor Probing Facility

The test should be performed under the unique specification which can be different for each product type. If the specification of previous test is not same as current test, additional setup should be incurred. Additional setups are downloading the specified test software in the machine, changing temperature of the chamber, and/or changing auxiliary resources such as probe card and load board. This sequence dependent setup times between different product types can be from 10 minutes to 8 hours. For example, in the real probing facility, downloading the test program for different test conditions takes about ten minutes, replacing probe card and board takes about an half hour, and changing the temperature requires 1 to 8 hours which is depend on the previous test temperature. Note that more time is required to cool the temperature down than to heat up. Also, at the low temperature (about 40°C~10°C), probe card cannot be changed because of the fracture of probe pin at low temperature. In this case, before changing probe card and board, temperature should be changed to atmosphere temperature (around 20°C). That means probe card can be changed at the temperature higher than or equal to room temperature. This combination of setup operations causes sequence dependent setup time.

2.1 MIP Model for Lot Transfer Problem

In most hierarchical production planning methods, the higher-level decision problem does not include the detailed real production specifications, such as, sequence dependent setup times, unexpected machine breakdowns, reworks, chip quality downgrade, scraps, and yield rates of wafers. We reflect the workload of each facility and ready times of each lot in the mixed integer programming (MIP) model for lot transfer/release planning to probing facilities in the higher level decision. Since the fabrication facilities and probing facilities are located closely and planning horizon is only one or two days, we assume that the transportation cost and inventory holding cost are negligible. Due to the make-to-order policy, the material cost is constant. Therefore, the controllable cost is only back order costs, and this back order costs are directly related to the tardiness of customers' orders. For the simplicity of MIP model, we considered the cost from the tardiness as the objective function, and assume that the earliest due date rule is used in each probing facility as a scheduling rule.

Before we present the formulation of the mixed integer programming for transfer/release planning to probing facilities, we clarify the notation used in the model.

Notation

J_j Set of lots associated with order j.

E_j Set of lots associated with the orders whose due dates are earlier than the due date of order j.

n number of lots to be scheduled.

m_k number of machines in facility k.

r_i ready time of lot i to probing facilities (same as the completion time of lots in the wafer fabs).

p_{jk} remaining work (= TAT) of order j. That is the largest remaining work of lots assigned to order j at facility k.

x_{ik} = 1 if lot i is assigned to facility k.

R_j Release time of lots assigned to order j. Release time should be greater than the largest ready time of lots.

T_j total tardiness of order j. (= $\max\{R_j+p_j-d_j, 0\}$)

O_k number of lots assigned to facility k more than average value $\sum_i p_{ik}\left\{m_k/\sum_k m_k\right\}$.

U_k number of lots assigned to facility k less than average value $\sum_i p_{ik}\left\{m_k/\sum_k m_k\right\}$

In the MIP model, we assume : wafer lots to be completed in the fabs are transferred to one of probing facilities directly; and yield rate, grade-down, reworks or scraps are not included. For the processing, unit of transportation is assumed to be one wafer. Sequence dependent setup times, unexpected machine breakdowns, and prescheduled preventive maintenance plans are neglected in the MIP model. Based on these assumptions, brief MIP model for lot transfer/release plan in probing facilities is described as following

$$\text{[P] Minimize } Z = c_1\sum_j T_j + c_2\sum_k \{O_k + U_k\}$$
$$= c_1\sum_j \max\{R_j + p_j - d_j, 0\} + c_2\sum_k \{O_k + U_k\} \quad (1)$$

subject to

$$\sum_k x_{ik} = 1, \qquad i = 1, \cdots, n, \quad (2)$$

$$R_j \geq \sum_{i\in E_j} p_{ik}x_{ik}, \qquad \forall j,\ k, \quad (3)$$

$$R_j \geq r_i, \qquad \forall j, \quad (4)$$

$$O_k \geq \sum_i p_{ik}x_{ik} - \sum_i p_i\left\{m_k/\sum_k m_k\right\}, \qquad \forall k, \quad (5)$$

$$U_k \geq \sum_i p_i\left\{m_k/\sum_k m_k\right\} - \sum_i p_{ik}x_{ik}, \qquad \forall k, \quad (6)$$

$$O_k,\ U_k \geq 0, \qquad \forall k, \qquad (7)$$

$$x_{ik} \in \{0,\ 1\}, \qquad \forall i,\ k. \qquad (8)$$

In equation (1), the objective function declares minimizing of the weighted sum of tardiness of customers' orders and deviation of the load which is defined as the sum of processing time in each facility. Constraints (2) ensure that lot i is assigned to only one of facilities. Constraint set (3) declares the release time of lots in facility k. Since we assumed that the lots are scheduled in EDD, the release times are greater than or equal to the time that all the lots with earlier due date are scheduled first. Constraints (4) ensure the release time greater than or equal to the time after the lot i is ready in the probing facilities. Constraints (5) to (7) represent the workload of each facility compared to number of machines. Constrains (8) ensure the decision variables x_{ik} are binary.

2.2 Discrete Event Simulation for Detailed Scheduling

After transfer/release plan of wafer lots is obtained from the mixed integer programming (MIP) model, schedules at the workstations are obtained with a dispatching rule based scheduling method which is built in discrete-event simulation for the low level decision in the hierarchical production planning. In the discrete event simulation for detailed lot release schedule, More than one wafer probing facilities and several operation steps of workstations are modeled in detail based on the real factory information. Also, dispatching-rule-based scheduling decisions are programed in the simulation model. In this model, sequence dependent setup times and the limitation of auxiliary resources such as the probe card and load board are considered. Therefore, the tardiness of customers' orders can be obtained from the simulation run.

We implement dispatching rule based scheduling rules proposed by Pfund et al. [17] called as *Apparent Tardiness Cost with Setups and Ready times* (ATCSR) rule for lot scheduling. This dispatching method based scheduling priority rule is an extended version of ATCS developed by Lee and Pinedo [10]. In ATCSR rule, not only the slack time of lots are considered in the priority rule, but also the ready times of lots are used for sorting of lots to be released. That is, separate exponential term of ready time is related to the slack of lots at the same level. The ATCSR index is given by equation (1).

$$I_{ATCSR}(t, l) = \frac{w_j}{p_j} exp\left(-\frac{\max\{d_j - p_j - \max(R_j, t), 0\}}{k_1 \bar{p}}\right) \times \exp\left(-\frac{q_{lj}}{k_2 \bar{q}}\right) \exp\left(-\frac{\max(R_j - t, 0)}{k_3 \bar{p}}\right)$$

The value of ATCSR is a priority value to sort the lots waiting for processing in the workstations. When one of the machines becomes available, a wafer lot with highest value of ARCSR is scheduled to the idle machine. The term, $d_j - p_j - \max(R_j, t)$ called as a slack term which means the time can be delayed to be processed until the job does not violate the due date of orders. If the lot assigned to an order already violates the due date of the order including remained work, the slack term of slack becomes 0, and the exponential term of slack equal to 1. If slack is still positive for a lot, the value of the exponential term is less than 1. This value depends on the amount of remaining slack and the value of the parameter k_1. A lower value of the weighting parameter k_1 amplifies slack effect. k_1, k_2, and k_3 were set as 2.4, 0.3, and 0.5 respectively, after tests on candidate values. See descriptions in Pfund et al. [17] for detailed information on ATCSR.

3. Lagrangian Relaxation Approach for Higher Level Decision

To reduce the computational time to solve the problem [P] exactly, we propose the solution approach based on Lagrangian relaxation and sub-gradient optimization methods. In the algorithm, the problem is relaxed by dualizing the set of constraints with Lagrangian multipliers and then the relaxed problem is decomposed in two subproblems. Here, we show the following relaxed problem, [L.R.], by relaxing constraint equations (2) and (3) with Lagrangian multiplier $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are vectors with nonnegative elements, i.e. $\lambda_{jk} \geq 0$ and $\mu_i \geq 0$ for all i, j, and k.

[L.R.] Minimize Z =

$$\sum_j \left\{c_1 \max\{R_j + p_j - d_j, 0\} - \sum_k \lambda_{jk}\right\} + \sum_j \sum_k \lambda_{jk} \left\{R_j - \sum_{i \in E_j} p_{ik} x_{ik}\right\} + c_2 \sum_k \{O_k + U_k\} + \mu_i \left(1 - \sum_k x_{ik}\right)$$

subject to (4) and (8).

Then, [L.R.] is decomposed into two sub-problems as follows.

[S.P.1($\mathbf{X}$)] Minimize Z_1 =
$\sum_j \sum_k \lambda_{jk}\left\{\sum_{i\in E_j} p_{ik}x_{ik}\right\}+$
$c_2\sum_k\{O_k+U_k\}+\mu_i(1-\sum_k x_{ik})$
subject to (5) TO (8)

[S.P.2($\boldsymbol{\lambda}$)] Minimize Z2 =
$\sum_j\left\{\max\{R_j+P_j-d_j,\,0\}-\sum_k \lambda_{jk}R_j\right\}$
subject to (4)

Solution of [S.P.2] can be obtained with simple calculation and the feasible solution of [P] working as upper bound can be calculated from the solution of [S.P.1] with reasonably short time. Given x_{ik} from the solution of [S.P.1], let $R_j^{'}=\sum_{i\in E_j\cap B_j} p_{ik}x_{ik}$ and $R_j^{''}=\max_{i\in L_j} r_i$. Then, we can consider four cases to determine the R_j value for each j.

i) $R_j^{'} \le R_j^{''} \le d_j-P_j$
j-th objective function becomes $-\sum_k \lambda_{jk}R_j$. By selecting $R_j=R_j^{''}$, the objective value become minimized.

ii) $R_j^{'} \le d_j-P_j \le R_j^{''}$
j-th objective function becomes $-\sum_k \lambda_{jk}R_j$. By selecting $R_j=D_j-P_j$, the objective value become minimized.

iii) $R_j^{''} \ge d_j-P_j \ge R_j^{'}$
j-th objective function becomes $(1-\sum_k \lambda_{jk})R_j+P_j-d_j$. By selecting $R_j=D_j-P_j$, the objective value become minimized.

iv) $R_j^{''} \ge R_j^{'} \ge d_j-P_j$
j-th objective function becomes $(1-\sum_k \lambda_{jk})R_j+P_j-d_j$. By selecting $R_j=R_j^{'}$, the objective value become minimized.

In order to search the best Langangian multipliers, the subgradient method is adopted in general. After the multipliers, λ_{jk}, corresponding to ready times of lot i assigned to the order j, and μ_i corresponding to the assignment of lot i to facility k are obtained by iterative approach, the Lagrangian multiplier at the next iteration, can be calculated as

$$\mu_i^{t+1}=\mu_i^t+\beta^t\left(1-\sum_k x_{ik}\right) \text{ and } \lambda_{ik}^{t+1}=\lambda_{ik}^t+\beta^t\left(\sum_{i\in E_j} p_{ik}x_{ik}-R_j\right).$$

Here, β^t is a positive gap size calculated as $\beta^t=\rho^t(z^t-Z(LR(\theta^t)))\Big/\left\{\left\|1-\sum_k x_{ik}\right\|^2+\left\|\sum_{i\in E_j} p_{ik}x_{ik}-R_j\right\|\right\}^2$, where z^t is the best upper bound. The best upper bound means the best feasible solution calculated at the former iteration and the value ρ^t, which is positive, is set to 1 at the first iteration, and this number is reduced by a half if the lower bound UB of problem [P], call as LB, is not decreased for a predetermined number of iterations. In this research we set the maximum iteration number as 20 with regarding the computational performance. The overall procedure for solving [P] with Lagrangian Relaxation and subgradient method can be summarized as follows. For the stopping conditions, parameter U, ε, and B are used. Each sub problem is solved by iteratively updating Lagrangian multipliers. The stopping condition of the iteration is that any one of three termination conditions is satisfied.

Stopping conditions for iteration

1) The iteration count reaches predetermined limit (defined as U)
2) The gap between an upper bound, UB, and a lower bound, LB, becomes less than a predetermined limit (defined as ε)
3) The lower bound has not been decreased for a predetermined number of iterations (defined as B).

Procedure 1. (*Solving* [P])
Step 0 : Set u = 0, b=0 and $\boldsymbol{X}$ = **0**.
Step 1 : If $u > U$ or $b > B$, stop; otherwise, go to step 2.
Step 2 : Find the optimal solution of [S.P.1($\boldsymbol{X}$)], and increase the number of u (i.e. u ← u+1). If the optimal solution of [S.P.1($\boldsymbol{X}$)] is feasible to [P], the obtained solution is optimal. Terminate the procedure. Otherwise, go to the next step.
Step 3 : Find a lower bound, LB, from the solution found in step 2, and update the best lower bound and set $b \leftarrow$ 0 if the newly found lower bound is less than current best lower bound. Otherwise, set $b \leftarrow b$+1 and update Lagrangian multipliers by the subgradient optimization method.

Step 4 : Find a feasible solution for [P]. If ratio of difference UB and LB, (UB-LB)/LB is less than ε, terminate. Otherwise, go the step 1.

4. Computational Experiments

The performance of the lot scheduling algorithms suggested in this study is evaluated through simulation experiments. For the experiments, we generated problem instances based on data of a real wafer probing facility in top ranking semiconductor manufacturing company in Korea. The following summarize information of the real fab as well as wafers and orders used in the simulation model.

1) There are two wafer fabrication facilities and four wafer probing facilities. Two wafer fabs are described as virtual ones which are generating the completion times of lots. The four wafer probing facilities are identical and each is modeled in detail.
2) Four workstations were included in the model of a probing facility : DC electrical Test, Wafer level Burn-in test, probing (call as pre-Laser in the production field), inking, and back grinding. Each workstation is composed of multiple identical parallel machines. Among these, wafer level Burn-in test, probing workstation is known as a bottleneck process, and there are 64 identical machines in this workstation.
3) The facilities produce 1,100 of product types. Each product type needs its own test specification such as the number of pins in probe card, the specified test temperature, the number of dies in one wafer, and its own test program.
4) The test time for a wafer lot on wafer burn-in and probing machine ranges from 30 minutes to one hour.
5) The setup time for downloading software for different test is 10 minutes, installing probe card and load board is 30 minutes. Sequence dependent setup time for changing the temperature for testing different product type ranges uniformly form 1 to 4 hours.

We use the test model provided by Bang and Kim [2]. The specification of test model is summarized as following. In the simulation model, it is assumed that orders for approximately 3,000 wafers are planned to arrive in each day from one wafer fabrication facility, as in the real probing facility. That is, the number of orders should be tested in each day (= order's due date) was generated from a discrete uniform distribution with range [60, 90], and the size of an order (in wafers) was generated from a discrete uniform distribution with range [25, 225]. Product types associated with the orders were randomly selected from the 1,100 product types.

The ready times of lots in probing facilities are to be the completion times of wafer lots in fabrication facilities. Therefore, the ready times of the lots associated with one order are distributed in some range, while the due date of an order is determined by customers or by the master plan of whole semiconductor manufacturing. Therefore, in this research, we assume that the due dates of orders are given and the ready times of lots associated with orders are normally distributed, that is, the ready time of lot k associated with order i was given as equation (10).

$$r_k = d_i - N(u \times W_i,\ RD^2) \tag{10}$$

where d_i is the due date of lot i, W_i is the sum of processing times of all operations for the order i, u is a parameter that defines tightness of ready time and due date, and $N(m,v)$ is a random number generated from a normal distribution with mean m, variance v. For each order, u is set to a random number generated from a discrete uniform distribution with range [1.5, 3.0].

The heuristic algorithm was coded in C language and the optimal solution of [S.P.1] is obtained by CPLEX 10.1, and the series of computational tests were performed on a personal computer with a Pentium IV operating at 3.2 GHz. We carefully set the values of the parameters for the stopping conditions of iterations by result of some trial tests with several candidate values of the parameters. We omit the detailed results of these tests, and these values are comparably working well : U = 300, ε = 0.01, B = 100 and U' = 600 for Lagrangian relaxation approach.

We tested the proposed algorithm for five scenarios in which the standard deviation value is varied for the distribution of the arriving time of lots assigned to the same order. That is, in scenario $RD2$, the ready time of lots assigned to order i is randomly generated according to equation (10) and the standard deviation is set as 2 hours. In the same way, the standard deviation is set as 4, 6, 8, and 10 hours for scenario $RD4$, $RD6$, $RD8$, and $RD10$, respectively. For each instance, orders of two days are generated and tested.

<Table 1> Performance of the Proposed Algorithm

Scenarios	Percentage gap†	Percentage reduction‡	Average CPU time (sec.)
R2	3.8	32.0	2309
R4	4.9	44.6	2830
R6	3.4	52.4	2043
R8	4.4	58.6	3277
R10	4.8	70.4	2332
Average	4.3	52.4	2558

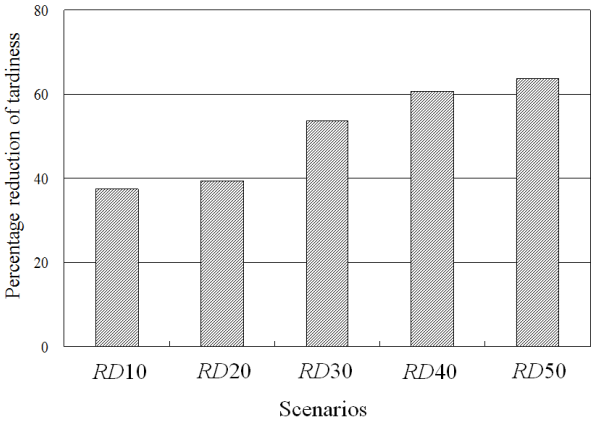
† average of the percentage gap from the best lower bound.
‡ average of the percentage of cost reduction from the resulting cost of REAL.

The performance of the suggested algorithm was compared with planning algorithm (named as REAL) which is used in the real probing facilities in the top rank semiconductor company in Korea. The performance index is shown as percentage of cost reduction from the resulting cost of REAL in the simulation test. In the REAL algorithm, wafer lots completed in fabs are transferred to the probing facilities with minimum load without considering due dates of orders which is the lots are associated. Furthermore, the other performance index is shown as percentage of gap from the lower bounds that are obtained from the Lagrangian relaxation method.

Results of the computational test are shown in <Table 1>. The average percentage gap of the total tardiness and deviation of workload of the suggested algorithm from lower bound was less than 5%. The suggested algorithm reduces about 70% of tardiness and deviation of workload compared with the REAL algorithm. This improvement may come from consideration of the due date and ready times of orders associated with lots completed in the fabs as well as the production load of each probing facility.

<Figure 4> shows that the performance of the proposed method according to the different scenarios. The reduction rate increases as the *RD* value become larger. The wide distribution of ready time of lots can be overcome by considering the due date of orders and ready times of lots associated with the same order in the proposed method. Note that only the load of each facility is considered in the REAL algorithm.

As the results given above, the proposed planning and scheduling methods outperforms the simple algorithm used in the real system. It is shown that the scheduling performance of the probing facilities is improved significantly by considering the due date of orders and the workload of facilities in the higher level decision.



<Figure 4> Performance According to Scenarios

5. Conclusion

In this paper, we suggest a two-level hierarchical production planning method for a lot transfer problem between wafer fabrication facilities and wafer probing facilities to reduce the total tardiness of orders. In the higher level, a lot transfer plan is obtained by solving mathematical model, and schedules at the machines are obtained with a priority-rule-based scheduling method and the lot transfer/release plan is evaluated with the discrete-event simulation in the lower level. The mixed integer programming model for the lot transfer is solved by Lagrangian relaxation method to reduce the computational time.

Results of computational tests showed that the total tardiness of orders can be reduced more effectively by using more sophisticated lot transfer methods, such as considering the due date and ready times of lots associated the same order with the mathematical formulation. The proposed method may be implemented for the problem of job assignment in back-end process such as the assignment of chips to be tested from assembly facilities to final test facilities. Also, the proposed method can be improved by considering the sequence dependent setup in the probing facilities. However, it may be more complicated and difficult to implement such a method in real situations.

References

[1] Aghezzaf, E., Production planning and warehouse management in supply networks with inter-facility mold transfers. *European Journal of Operational Research*, 2007, Vol. 182, No. 3, pp. 1122-1139.

[2] Bang, J.-Y. and Kim, Y.-D., Scheduling algorithms for a semiconductor probing facility. *Computers and Operations Research*, 2011, Vol. 38, pp. 666-673.

[3] Brahimi, N., Dauzere-Peres, S., Najid, N.M., and Nordli, A., Single item lot sizing problems. *European Journal of Operational Research*, 2006, Vol. 168, No. 1, pp. 1-16.

[4] Cha, M.-S. and Jang, J.-S., Effective Operation of SPC System in Semiconductor Manufacturing. *Journal of the Korean Institute of Plant Engineering*, 2009, Vol. 14, No. 4, pp. 95-103.

[5] Chen, T.-R. and Hsia, T.C., Scheduling for IC sort and test facilities with precedence constraints via Lagrangian Relaxation. *Journal of Manufacturing Systems*, 1997, Vol. 16, pp. 117-128.

[6] Chen, T.-R., Chang, T.-S., Chen, C., and Kao, J., Scheduling for IC sort and test with preemptiveness via Lagrangian Relaxation. *IEEE Transactions on Systems, Man, and Cybernetics*, 1995, Vol. 25, pp. 1249-1256.

[7] Ellis, K.P., Lu, Y., and Bish, E.K., Scheduling of wafer test processes in semiconductor manufacturing. *International Journal of Production Research*, 2004, Vol. 42, pp. 215-242.

[8] Jang, S.-H., Optimal Production Capacity and Outsourcing Production Planning for Production Facility Producing Multi-Products. *Journal of the Society of Korea Industrial and Systems Engineering*, 2012, Vol. 35, No. 4, pp. 110-117.

[9] Lee, E.-Y., Assessment Criteria for Process FMEA in Assembly Process of Semiconductor. *Journal of the Korean Institute of Plant Engineering*, 2013, Vol. 18, No. 1, pp. 5-18.

[10] Lee, Y.H. and Pinedo, M., Scheduling jobs on parallel machines with sequence dependent setup times. *European Journal of Operational Research*, 1997, Vol. 100, pp. 464-474.

[11] Lin, J.T. and Chen, Y.-Y., A multi-site supply network planning problem considering variable time buckets-A TFT-LCD industry case. *International Journal of Advanced Manufacturing Technology*, 2007, Vol. 3, pp. 1031-1044.

[12] Liu, C.-Y. and Chang, S.-C., Scheduling flexible flow shops with sequence-dependent setup effects. *IEEE Transactions on Robotics and Automation*, 2000, Vol. 16, No. 4, pp. 408-419.

[13] Ovacik, I.M. and Uzsoy, R., Rolling horizon procedures for dynamic parallel machine scheduling with sequence-dependent set-up times. *International Journal of Production Research*, 1995, Vol. 33, pp. 3173-3192.

[14] Pearn, W.L., Chung, S.H., and Yang, M.H., A case study on the wafer probing scheduling problem. *Production Planning and Control*, 2002a, Vol. 13, pp. 66-75.

[15] Pearn, W.L., Chung, S.H., and Yang, M.H., Minimizing the total machine workload for the wafer probing scheduling problem. *IIE Transactions*, 2002b, Vol. 34, pp. 211-220.

[16] Pearn, W.L., Chung, S.H., Yang, M.H., and Chen, A. Y., Algorithm for the wafer probing scheduling problem with sequence-dependent set-up time and due date restrictions. *Journal of the Operational Research Society*, 2004, Vol. 55, pp. 1194-1207.

[17] Pfund, M., Fowler, J.W., Gadkari, A., and Chen, Y., Scheduling jobs on parallel machines with setup times and ready times. *Computers and Industrial Engineering*, 2008, Vol. 54, pp. 764-782.

[18] Seo, J.-C. and Bang, J.-Y., On-time Production and Delivery Improvements through the Demand-Lot Pegging Framework for a Semiconductor Business. *Journal of the Society of Korea Industrial and Systems Engineering*, 2014, Vol. 37, No. 4, pp. 126-133.

[19] Trigerio, W.W., Thomas, L.J., and McClain, J.O., Capacitated lot-sizing with setup times. *Management Science*, 1989, Vol. 35, pp. 353-366.

[20] Yang, M.H., Pearn, W.L., and Chung, S.H., A case study on the wafer probing scheduling problem, *Production Planning and Control*, 2002, Vol. 13, pp. 66-75.

ORCID

June-Young Bang | http://orcid.org/0000-0003-4999-9161