Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

ON THE CONSEQUENCES OF UN-MODELED DYNAMICS TO THE OPTIMALITY OF SCHEDULES IN CLUSTERED PHOTOLITHOGRAPHY TOOLS

Hyeong-Ook Kim
Se-Hyeon Park
Jung Yeon Park
James R. Morrison

Department of Industrial and System Engineering
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu
Daejeon, KS015, SOUTH KOREA

ABSTRACT

Clustered photography tools (CPTs) are very complex and can substantially influence the throughput of wafer fabrication facilities. Therefore, efficient lot scheduling for CPTs can directly improve fab performance. In this paper, we develop mixed integer linear programs for linear, affine, exit recursion, and flow line models of CPTs to optimize schedules with respect to mean cycle time, makespan, and tardiness. We simulate a true CPT using a flow line and solve the MILPs for other above mentioned, reduced models. Schedules from reduced models are then input into the flow line optimization model in order to evaluate the loss. Using numerical experiments, we show that exit recursion models outperform other models. Under time limits, exit recursion models exhibit at least 6% better performance than flow lines for large problems on cycle time.

1 INTRODUCTION

Clustered photolithography tools (CPTs) are one of the most expensive and complex tools in semiconductor manufacturing, and significantly impact the throughput of wafer fabrication facilities, cf. Morrison and Martin (2007). As such, CPTs should be operated effectively and wafers should be scheduled optimally in order to decrease inefficiencies at the tool.

There are several ways to model a CPT, ranging from simple to complex. Simpler models are easier to construct and require less computation time, while complex models can more accurately describe the dynamics of the system. We consider several equipment models of CPTs and evaluate the impact of the un-modeled dynamics on the optimality of lot schedules.

1.1 Lot Sequence Scheduling in Fab

Semiconductor fabs process multiple types of wafers, depending on the product. As every class of wafer has its own process flows, even within a CPT, it is important to efficiently schedule lots so as to improve the overall throughput. Traditionally, rule based and mathematical approaches have been used for sequence scheduling and are still often applied in the industry. Monch et al. (2011) provides an overview of scheduling and example problems of the scheduling in the fab. Wein (1988) focuses on lot sequence rules, and varies the inputs to compare the results. Cakici and Mason (2007) proposes heuristics based on a network mathematical model for scheduling photolithography machines. Recently, learning approaches have been employed for lot sequence scheduling. Riedmiller and Riedmiller (1999) uses reinforcement learning techniques to learn dispatching policies under uncertainty in semiconductor manufacturing

production. Ramirez-Hernandez and Fernandez (2009) uses an actor-critic architecture for the dispatching of lots in Intel mini-fab benchmark models.

1.2 Equipment Models for CPT

Many different types of equipment models have been applied to CPTs for optimization and simulation, and simpler models such as linear or affine are still often used in the industry. Linear models assume that the wafer throughput for a wafer class is constant; the production rate may be different for each class. Kabak et al. (2013) and Yan et al. (2013) used linear models of CPTs for ASIC fab models and litho machine scheduling. Affine models extend linear models by adding a first wafer delay, where the production rate of the first wafer can be different than other wafers in a lot. They have been used to explore the optimality of scheduling for photolithography tools in Bitar et al. (2014). Flow line models have been used to model CPTs in (Wu and Chiou 2010; Chiou and Wu 2014; Park and Morrison 2015; Park and Hwang 2013); they consider process modules within a CPT but ignore wafer handling robots. Park et al. (2017a) explores the tradeoff between accuracy and computation time between affine and flow line models of CPTs, using a detailed petri-net model as the baseline. It was shown in Park et al. (2017a) and Morrison (2010) that flow line models are expressive and accurate enough to replace detailed models for modeling CPTs. Park et al. (2017b) proposes a new class of models called exit recursion models, which were shown to be close in accuracy to flow lines, while requiring much less computation time.

In this paper, we focus on the following equipment models: linear, affine, flow line and exit recursion models (ERMs). Similarly to Park et al. (2017b), we compare simpler models to more complex models and compare their performance in lot scheduling for CPTs.

1.3 Contribution and Organization

Park et al. (2017b) use simple FIFO heuristics to evaluate the prediction error. In this paper, we optimize lot sequence scheduling for CPT using various equipment models. We assume that flow line models of CPTs are exact and use them as our baseline. We construct mixed integer linear programs (MILPs) for linear, affine, exit recursion models and calculate the optimal sequence schedules for these reduced models. The optimal schedules are then compared to the optimal schedules obtained from the flow line model and used to evaluate the loss occurred when using a reduced model.

This paper is organized as follows. In Section Ⅱ, we describe the CPT system, and describe how to apply flow lines to CPTs. In Section Ⅲ, we develop how to derive optimal lot sequences for each of the equipment models and detail the exact mathematical equations of each MILP. Section Ⅳ shows the full simulation conditions used and compare the schedules obtained from other equipment models to those of flow lines. We discuss the experiment results here. In section Ⅴ, we make concluding remarks and consider directions for future research.

2 SYSTEM DESCRIPTION

2.1 Clustered Photolithography Tools

A clustered photolithography tool transfers patterns onto wafers using a mask called the reticle. Figure 1 illustrates the layout of an actual CPT used in the semiconductor industry in Yoon and Lee (2004). Each load ports can hold a lot of wafers which enter and exit the CPT. It consists of four clusters of modules, where each cluster contains a robot arm that can move a single wafer at a time between the modules. There are processes: hot plates (HP/HHP), low-pressure adhesions (LPAH), cold plate (CP), spin coaters (SC), post exposure bake hot plate (PEB), edge exposures (EE), and spin developers (SD). There may exist redundant chambers in a process module for batch processing of wafers. Groups of wafers called lots arrive at the CPT at the indexer, are processed according to their recipe, and exit the tool via the indexer. There may be several classes of lots according to their recipe. Modules are restricted to processing only one wafer class at a time, regardless if the module would allow parallel identical steps, to prevent overtaking and

contamination. While many different types of setups may occur within the tool. we only consider the reticle alignment setup in this paper. A reticle alignment setup is performed for the first wafer of every lot just before the stepper process. In this paper, we assume that reticle alignment setups are uniformly distributed. For a more detailed explanation of this CPT, see Park et al. (2017a) and Yoon and Lee (2004).

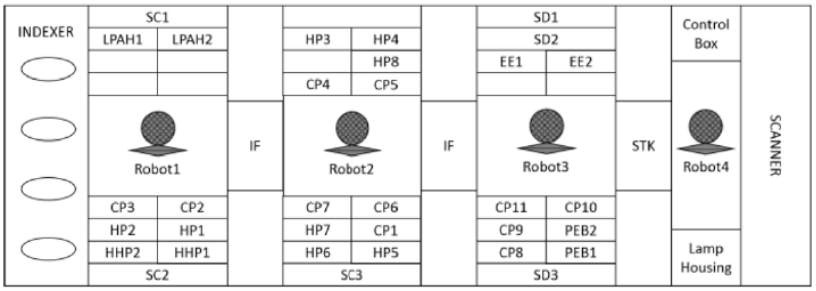

Figure 1: Example CPT layout.

2.2 Applications of Equipment Models to CPT

Table 1: Process time for parametric flow line.

Proc #	TARC #1			TARC #2			BARC		
	Proc	***R***	***PT***	***Proc***	***R***	***PT***	***Proc***	***R***	***PT***
1	Dummy	1	0	Dummy	1	0	Op 1	1	5
2	Dummy	1	0	Dummy	1	0	Op 2	2	85
3	Dummy	1	0	Dummy	1	0	Op 3	2	95
4	Dummy	1	0	Dummy	1	0	Op 4	2	65
5	Op 1	1	5	Op 1	1	5	IF	1	5
6	Op 2	2	85	Op 2	2	85	Op 5	1	63
7	Op 3	2	95	Op 3	2	95	Op 6	2	95
8	Op 4	2	65	Op 4	2	65	Op 7	2	65
9	Op 5	2	70	Op 5	2	70	IF	1	5
10	IF	1	5	Op 6	2	95	Op 8	2	70
11	Op 6	1	63	IF	1	5	IF	1	5
12	Op 7	2	95	Op 7	2	65	Op 9	2	95
13	IF	1	5	Op 8	1	63	IF	1	5
14	Op 8	2	65	IF	1	5	Op 10	2	65
15	STK	15	5	STK	15	5	STK	15	5
16	Op 9	1	110	Op 9	1	110	Op 11	1	110
17	STK	1	5	STK	1	5	STK	1	5
18	Op 10	2	95	Op 10	2	95	Op 12	2	95
19	Op 11	2	65	Op 11	2	65	Op 13	2	65
20	Op 12	2	95	Op 12	2	95	Op 14	2	95
21	Op 13	3	135	Op 13	3	135	Op 15	3	135
22	IF	1	5	IF	1	5	IF	1	5
23	Op 14	2	95	Op 14	2	95	Op 16	2	95
24	Op 15	2	65	Op 15	2	65	Op 17	2	65
25	IF	1	5	IF	1	5	IF	1	5
26	Op 16	1	0	Op 16	1	0	Op 18	1	0

We use the procedure described in Park et al. (2017a) and Park et al. (2017b) in order to apply linear, affine, exit recursion, and flow line models to CPTs. Their detailed equations will not be listed here. As all of the models use historical data to calculate parameters, we use the parametric flow line model from Park et al. (2017a) to generate tool log data of the CPT and assume that it is exact. The flow line must be modified in order to be applied to a CPT; see Table 1. R is the number of redundant chambers in a process module, and PT is the process time in seconds.

The parameters for each of the linear, affine, and exit recursion models are then calculated from the tool log data generated by the flow line model.

2.3 Procedure to Evaluate Optimal Schedules

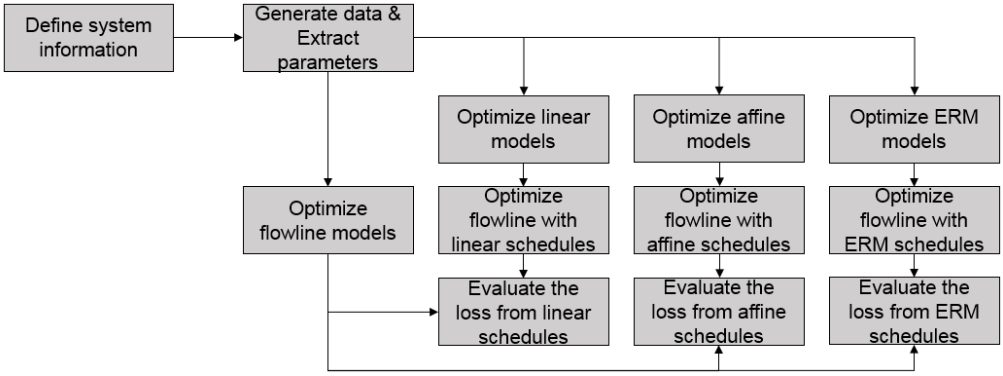

Figure 2: Schedule evaluation procedures.

The overall research procedure is shown in Figure 2. After defining system information such as process flow, process times, module redundancy and other configuration information, we generate historical data using the parametric flow line model and extract parameters from generated data for each equipment model. Note that this generated historical data is only used to calculate the model parameters, and is not used for the optimization models. After the model parameters are calculated, we construct mixed integer linear programs (MILPs) for each equipment model (linear, affine, ERM, and flow line) in order to determine the optimal lot sequence schedule and solve them. After the optimal schedules have been obtained from the reduced models (linear, affine, and ERM), the flow line MILP is solved once more where it is constrained to be equal to the each of the reduced model schedules. The optimal values obtained from these reduced models are then compared to the optimal value of the true system found previously. Here, the optimal schedule from the flow line MILP is assumed to be the true solution. This allows us to explore the consequences of using a simpler model in place of a more detailed model for lot sequence scheduling.

3 MATHEMATICAL MODELS FOR SEQUENCE SCHEDULING

3.1 Objective Functions and Notations

We consider three objective functions to minimize: average lot cycle time (CT), makespan, and tardiness. To calculate cycle time, we define C_l and a_l as completion and arrival time of lot l. L is the total number of lots to be scheduled. Average cycle time is calculated as:

$$O_1 : \sum_{l=1}^{L}(C_l - a_l)/L \tag{1}$$

To minimize makespan, we define C_{max}, the maximum completion time, as the total make span of a schedule, and add constraint (3). It ensures that C_{max} is the completion time of the last lot.

$$O_2 : C_{max} \tag{2}$$

$$C_{max} > C_l, \quad l = 1, \dots, \mathrm{L} \tag{3}$$

The last objective function is tardiness. To minimize the total tardiness, D_l is defined as the due date for lot l and T_l is a decision variable for the tardiness of lot l. T_l will be minimized when minimizing O_3. Constraints (5) and (6) are added to define the lower and upper bounds of T_l. The total tardiness is calculated as:

$$O_3 : \sum_{l=1}^{L} T_l \tag{4}$$
$$C_l - D_l \leq T_l, \quad l = 1, \dots, L \tag{5}$$
$$T_l \geq 0, \quad l = 1, \dots, L \tag{6}$$

As described in the previous section, model parameters are extracted using the generated data, which is then used as values in each MILP for linear, affine, exit recursion, and flow line models. Table 2 describes some notation. As reticle setups are uniformly distributed, we use the expected values. Note that 3 types of objective function do not mean multi-objective problems.

Each equipment model is a function of input parameters and model parameters, and outputs the start and completion times of lot l. Thus the decision variables are:

S_l : Start time of lot l,

C_l : Completion time of lot l,

$$y_{k,l} : \begin{cases} 1, & if\ lot\ k\ is\ direct\ predecessor\ of\ lot\ l \\ 0, & otherwise. \end{cases}$$

S_l and C_l are continuous variables, and $y_{k,l}$ is binary. Note that $y_{k,l}$ is the lot sequence.

Table 2: Notations.

Parameters	Description
L	Total number of lots
W_l	Number of wafers in lot l
a_l	Arrival time of lot l
G_l	Class of lot l
D_l	Due date of lot l
A_{G_l}	Bottleneck process time value of lot class G_l
B_{G_k,G_l}	First wafer delay when lot class G_k is direct predecessor G_l
$B^0_{G_l}$	First wafer delay of lot class G_l for the first lot
$A^1_{G_l}$	Bottleneck process time with non-bottleneck contention of lot class G_l
$A^2_{G_l}$	Bottleneck process time with bottleneck contention of lot class G_l
$D^1_{G_l}$	Vacating time related parameters without class change
$D^2_{G_l}$	Vacating time related parameters with class change
FWD_{G_l}	First wafer delay of lot class G_l
MS	Total number of module stages
B	Index of scanner module in flow line
D_{G_l}	Number of dummy modules for lot class G_l
$R_{G_l,m}$	Number of redundant chambers at module m for lot class G_l
$\tau^{pt}_{G_l,m}$	Deterministic process time per wafer of lot class G_l at module m
$\tau^{\mathrm{sr}}_{G_k,G_l}$	Reticle alignment setup time for first wafer of lot l at reticle stage when lot class G_k is predecessor of lot class G_l
$\tau^{sr_0}_{G_l}$	Reticle alignment setup time for first wafer of lot l at reticle stage when the first lot class is G_l and there is no predecessor
$y^{sch}_{k,l}$	Solution schedule from reduced models

3.2 Linear and Affine Models

We describe constraints used for linear and affine models.

$$S_l \geq a_l, \quad l = 1, \dots, L \tag{7}$$
$$S_l \geq C_k - M\left(1 - y_{k,l}\right), \quad k, l = 1, \dots, L \tag{8}$$

Constraints (7) and (8) are start time constraints for each lot, for both linear and affine models. Constraint (7) enforces the start time to be greater or equal to the arrive time of lot *l*. (8) enforces that the start time of lot *l* be greater or equal to the completion time of lot *k* if lot *k* is the predecessor of lot *l*. Here M is Big M, an arbitrarily large number. If lot *k* is not the predecessor of lot *l*, then constraint (8) will automatically be satisfied.

$$C_l = S_l + A_{G_l} W_l, \quad l = 1, \dots, L \tag{9}$$
$$C_l = S_l + A_{G_l}(W_l - 1) + \sum_{k=1}^{L} B_{G_k G_l} y_{k,l} + B^0_{G_l}\left(1 - \sum_{k=1}^{L} y_{k,l}\right), \quad l = 1, \dots, L \tag{10}$$

The completion time constraint for linear models is shown in (9). The completion time for lot *l* is calculated as $A_{G_l} W_l$, where A_{G_l} is the model parameter and W_l is the number of wafers in lot *l*. Equation (10) is the completion time constraint for affine models. In this case, the time from start to completion is calculated in the form of $A_{G_l}(W_l - 1) + B$, where the first wafer delay is taken into account. This first wafer delay can depend on the current and previous lot class. Using the binary variable $y_{k,l}$, the value $B_{G_k G_l}$ is only used when the predecessor of lot *l* is lot *k*. If there is no predecessor of lot *l* (i.e. *l* = 1, the first lot), the value of $\sum_{k=1}^{L} B_{G_k G_l} y_{k,l}$ is 0, and the $B^0_{G_l}$ value is used.

$$\sum_{k=1}^{L} \sum_{l=1}^{L} y_{k,l} = L - 1 \tag{11}$$
$$\sum_{k=1}^{L} y_{k,l} \leq 1, \quad l = 1, \dots, L \tag{12}$$
$$\sum_{l=1}^{L} y_{k,l} \leq 1, \quad k = 1, \dots, L \tag{13}$$
$$y_{k,l} = 0, \quad k, l = 1, \dots, L, k = l \tag{14}$$
$$S_l, C_l \geq 0, y_{k,l} \; is \; binary, \quad k, l = 1, \dots, L \tag{15}$$

Equations (11)-(14) are predecessor constraints for ensuring a correct sequence of lots. (15) are non-negativity and binary constraints. Equations (11)-(15) are common constraints that are also used for ERMs and flow line models.

The MILPs to obtain optimal lot sequence schedules for each of the objective functions when a linear model is used is:

$$\text{Minimize } (1) \text{ or } (2) \text{ or } (4)$$
$$\text{Subject to } (7) - (9), \; (11) - (15) \; or \; add \; (3) \text{ or add } (5) - (6).$$

The affine model MILPs are similar.

$$\text{Minimize } (1) \text{ or } (2) \text{ or } (4)$$
$$\text{Subject to } (7) - (8), \; (10) - (15) \; or \; add \; (3) \text{ or add } (5) - (6).$$

3.3 Exit Recursion Models

For ERMs, two more decision variables are required.

V_l^1 : Vacating time of lot *l* without class change

V_l^2 : Vacating time of lot *l* with class change

V_l^1 and V_l^2 are both decision variables for the vacating time, which is defined as the time that lot *l* has vacated the modules required for lot *k* to start processing with $y_{k,l} = 1$. As a process module may only process one class at a time, two variables are required. V_l^1 is the time when the last wafer of lot *l* has vacated at least one chamber from the first module so that the next lot (which must be of the same class as lot *l*) can start processing. V_l^2 is the time when the last wafer of lot *l* has vacated all chambers of the first module so that the next lot (which must be of a different class as lot *l*) can start processing. These variables are mutually exclusive; one of either V_l^1 or V_l^2 is used in the constraints for each lot, depending on whether a class change occurs. Due to the vacating time parameters, there are additional constraints for the start and completion times; they are listed below.

$$S_l \geq V_k^1 - M(1 - y_{k,l}), \quad k, l = 1, \dots, L \tag{16}$$
$$S_l \geq V_k^2 - M(1 - y_{k,l}), \quad k, l = 1, \dots, L \tag{17}$$
$$V_l^1 = C_l - D_{G_l}^1, \quad l = 1, \dots, L \tag{18}$$
$$V_l^2 = C_l - D_{G_l}^2, \quad l = 1, \dots, L \tag{19}$$

Constraints (16) and (17) are for the start times of each lot. If a class change does not occur between lots *k* and *l*, constraints (16) is used, and (17) otherwise. Constraints (18) and (19) define the vacating times for each case.

$$C_l \geq S_l + A_{G_l}^1(W_l - 1) + FWD_{G_l}, \quad l = 1, \dots, L \tag{20}$$
$$C_l \geq C_k + A_{G_l}^2(W_l - 1) + \sum_{k=1}^{L} B_{G_k G_l} y_{k,l} + B_{G_l}^0\left(1 - \sum_{k=1}^{L} y_{k,l}\right) - M(1 - y_{k,l}), \quad k, l = 1, \dots, L \tag{21}$$

The completion time constraints take a similar form to those of the affine model. However, as the ERM equations use a max function to determine completion time, a total of two constraints are required: one for each argument of the max function. As the completion time must be greater than both of these constraints, the max function is satisfied.

$$V_l^1, V_l^2 \geq 0, \quad l = 1, \dots, L \tag{22}$$

Lastly, we add non-negativity constraints (22) for the new decision variables. The lot sequence constraints are the same as those of linear and affine models.

The MILPs to solve lot sequence scheduling using ERMs are expressed as:

$$\text{Minimize (1) or (2) or (4)}$$
$$\text{Subject to (11)} - \text{(15)}, \ \text{(16)} - \text{(22)} \ \textit{or add} \ \text{(3) or add (5)} - \text{(6)}.$$

3.4 Flow Line Models

Flow line models consider the individual process modules within a CPT. As these modules use the module entry times of each wafer in a lot to calculate start and completion times of the lot, more decision variables are required.

$X_{l,w,m}$: Entry time of lot *l* of wafer *w* into stage *m*

$$X_{l,1,1} \geq a_l, \quad l = 1, \dots, L \tag{23}$$
$$X_{l,w,1} \geq X_{l,w-R_{G_l,1},2}, \quad l = 1, \dots, L, \ R_{G_l,1} < w \leq W_l \tag{24}$$
$$X_{l,w,1} \geq X_{l,w-1,1}, \quad l = 1, \dots, L, w = 2, \dots, W_l \tag{25}$$

Constraints (23)-(25) are for the entry times of the first process module in a CPT. (23) ensures that the entry time of the first module for the first wafer of lot l is greater or equal to the lot arrival time. (24) considers redundancy of the first module and allows the appropriate wafer to enter the first module once a single chamber is emptied. Constraint (25) prevents the overtaking of wafers.

$$X_{l,w,m} \geq X_{l,w,m-1} + \tau^{pt}_{G_l,m-1},\ \ l = 1,\dots,L, w = 1,\dots,W_l \tag{26}$$
$$X_{l,w,m} \geq X_{l,w-R_{G_l,m},m+1},\ \ l = 1,\dots,L,\ \ R_{G_l,m} < w \leq W_l, m = 2,\dots,MS-1 \tag{27}$$
$$X_{l,w,m} \geq X_{l,w-1,m},\ \ l = 1,\dots,L, w = 2,\dots,W_l, m = 2,\dots,MS-1 \tag{28}$$
$$X_{l,w,MS} \geq X_{l,w,MS-1} + \tau^{pt}_{G_l,MS-1}, l = 1,\dots,L, w = 1,\dots,W_l \tag{29}$$
$$X_{l,w,MS} \geq X_{l,w,MS} + \tau^{pt}_{G_l,MS},\ \ l = 1,\dots,L, R_{G_l,MS} < w \leq W_l \tag{30}$$
$$X_{l,w,MS} \geq X_{l,w-1,MS},\ \ l = 1,\dots,L, w = 2,\dots,W_l \tag{31}$$

(26)-(31) are module constraints for wafers within a lot, as wafer w is processed from module 2 to MS. (26) enforces the entry time to module m to be greater or equal to the sum of the entry time to module $m+1$ and the module process time. Constraints (27) and (28) serve a similar purpose to (24) and (25). (27) defines the entry time wafer w of module m to be at least the entry time of wafer $w - R_{G_l,m}$ to module $m+1$, considering the redundancy of module m. (28) prevents overtaking as before. Constraints (29)-(31) are similar and are for the last module stage.

$$X_{l,w,1} \geq X_{k,W_k-R_{G_l,1}+w,2} - M(1-y_{k,l}), k,l = 1,\dots,L, 1 \leq w < R_{G_l,1}, G_k = G_l \tag{32}$$
$$X_{l,w,1} \geq X_{k,W_k,2} - M\big(1-y_{k,l}\big), k,l = 1,\dots,L, 1 \leq w < R_{G_l,1}, G_k \neq G_l \tag{33}$$
$$X_{l,w,m} \geq X_{k,W_k-R_{G_l,m}+w,m+1} - M\big(1-y_{k,l}\big),\ \ k,l = 1,\dots,L, m = 2,\dots,MS-1, 1 \leq w < R_{G_l,m}, G_k = G_l \tag{34}$$
$$X_{l,w,m} \geq X_{k,W_k,m+1} - M\big(1-y_{k,l}\big),\ \ k,l = 1,\dots,L, m = 2,\dots,MS-1, 1 \leq w < R_{G_l,m}, G_k \neq G_l \tag{35}$$
$$X_{l,w,MS} \geq X_{k,W_k-R_{G_l,MS}+w,MS} + \tau^{m}_{G_k,MS} - M\big(1-y_{k,l}\big), k,l = 1,\dots,L, 1 \leq w < R_{G_l,M}, G_k = G_l \tag{36}$$
$$X_{l,w,m} \geq X_{k,W_k,MS} + \tau^{m}_{G_k,MS} - M\big(1-y_{k,l}\big), k,l = 1,\dots,L, 1 \leq w < R_{G_l,MS}, G_k \neq G_l \tag{37}$$
$$X_{l,1,B+1} \geq X_{l,1,B} + \tau^{m}_{G_l,B} + \textstyle\sum_{k=1}^{L} \tau^{sr}_{G_k,G_l} y_{k,l}\ , + \tau^{sr_0}_{G_l}\big(1 - \textstyle\sum_{k=1}^{L} y_{k,l}\big) - M\big(1-y_{k,l}\big),\ \ l = 1,\dots,L \tag{38}$$

Constraints (32)-(38) are module constraints for wafers between different lots and consider setups. They are similar to (26)-(31) but consider whether lot k is the predecessor of lot l and use Big-M notation to ignore non-successive lots. In particular, (38) includes the reticle alignment setup between successive lots.

$$S_l = X_{l,D_{G_l}+1,1},\ \ l = 1,\dots,L \tag{39}$$
$$C_l = X_{l,W_l,MS} + \tau^{MS}_{G_l},\ \ l = 1,\dots,L \tag{40}$$
$$X_{l,w,m} \geq 0,\ \ l = 1,\dots,L, w = 1,\dots,L,\ \ m = 1,\dots,MS \tag{41}$$

(39) defines the start time of lot l as the entry time of the first wafer into the first stage, considering dummy modules (process flows of different classes may have different lengths and so dummy modules are used to coerce the same number of module stages for each class). (40) defines the completion time of lot l as the sum of the entry time of the last wafer at the last module stage and its process time.

The optimal lot sequence schedule using flow lines is solved by:

$$\text{Minimize (1) or (2) or (4)}$$
$$\text{Subject to (11)} - \text{(15)},\ \ \text{(23)} - \text{(41)}\ \textit{or add}\ \text{(3) or add (5)} - \text{(6)}.$$

3.5 Loss and Error Evaluation

Using the MILPs described above for the linear, affine and exit recursion models, we can obtain the optimal sequence schedule for the three objective functions. In effect, we are modeling the true system with reduced models to solve a sequence scheduling problem. Using the solutions from these reduced models, we add an extra constraint (42) to the flow line MILPs to enforce that the same schedule is used, and solve for objective values. $y_{k,l}^{sch}$ is the optimal sequence obtained from the reduced models.

$$y_{k,l}^{sch} = y_{k,l}, \quad k,l = 1,\dots,L \tag{42}$$

By reusing the optimal sequence from reduced models and solving the flow line model again, we can evaluate the loss that occurs. Loss is defined as the ratio between objective values of reduced models and flow line from each solutions. Error is defined as the ratio between objective values of flow line from reduced schedules and objective value of flow line. Loss means how much inaccuracy occurs when we use abstract models and error means how far prediction goes wrong.

4 EXPERIMENTS

4.1 Experimental Design

Table 3 describes the design of experiments. The number of lots is varied from 1 to 100. As the number of lots increases, the optimization models cannot be solved optimally in adequate time. We thus set the time limit as 1 hour. As described before, we conduct experiments about 3 objectives. We consider two cases of reticle alignment setups, where setups are either independent of or dependent on lot class changes. In the first case, the reticle setup is uniformly distributed from [210, 260] regardless of class changes between lots. In the second case, we use different distributions for class changes, where the reticle setup is uniformly distributed in [210,260] if there is no class change. For cases with class change, a factor from the non-symmetric of Table 3 is applied to the same distribution. We use 3 types of arrival. First one is a classical assumption of scheduling problems in Bitar et al. (2014). Other two arrival case is more similar with reality. Lot length means the possible number of wafer in a lot and Average lot size means the average number of wafers in a lot. We also include FIFO, a simple rule-based heuristic, as a guideline to compare the mathematical models developed in this paper. 10 replications are conducted for each of test cases.

Table 3: Design of experiments.

Category		Description	Category		Description
No. of lots	1	1, 5, 10, 20, 50, 100 lots in optimization	Average lot size	1	12, { 11, 12, 13 }
Objectives	1	Mean cycle time, Total makespan, Total tardiness		2	14. { 13, 14, 15 }
Reticle setup distribution	1	Uniformly distributed, [210,260] for every lot change		3	16, { 15, 16, 17 }
	2	Uniformly distributed, [210,260] for no-class change Uniformly distributed, (2or3) * [210,260] for class change		4	18, { 17, 18, 19 }
Arrival times	1	Interarrival times of all lots are 0		5	20, { 19, 20, 21 }
	2	20 % of lots have 0, the other follow exponential distribution		6	22, { 21, 22, 23 }
	3	All the Interarrival follow exponential distribution		7	24, { 23, 24, 25 }
Lot length	1	3, { 23, 24, 25 }	Time limits		60 min
	2	6, { 20, 21, 22, 23, 24, 25 }	Optimality gap		1e-4
	3	9, { 17, 18, 19, 20, 21, 22, 23, 24, 25 }			
	4	12, { 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 }			
	5	15, { 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 }			
	6	Bimodal distribution, { 11, 12, 13 } & { 23, 24, 25 }			

4.2 Result and Discussion

FIFO is used as the initial schedule for all optimization models. We consider loss to compare each model, where it is calculated as the final objective value of the reduced model divided by the final objective value of the flow line model. Note that as there are time limits and the flow line optimization model contains many more decision variables than the other reduced models. Reduced models may have negative loss, meaning that they can outperform our true system in our experiments.

In the most variation case, there is chance to decrease 20% or 30% inefficiencies, if we use MILP schedules instead of FIFO or LPT. Even all our mathematical model show similar results about loss, using ERM makes to be better predicted more than 5% linear and affine models. Next step is to conduct sensitivity analysis from table 3. We will show interesting results among them.

Figure 3 shows the cycle time loss of each types of setup distribution. With the same distribution, there is not much differences between the models. Otherwise, the presence of non-symmetric setup makes affine and ERM outperform other models. Linear and FIFO cannot reflect the sequence dependent characteristics.

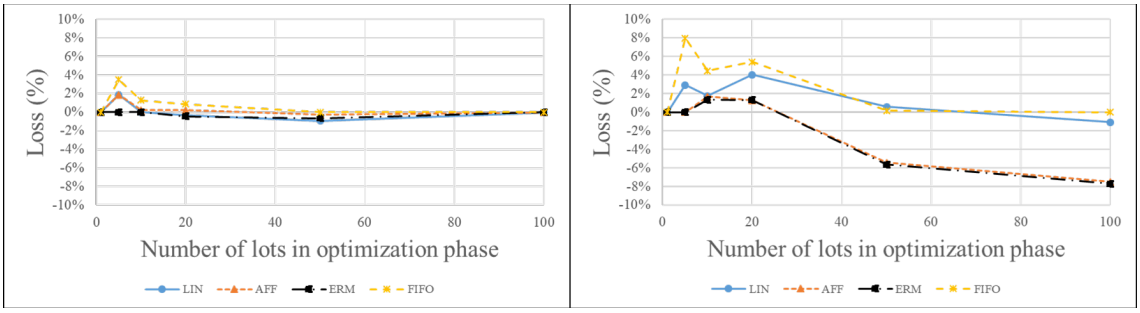

Figure 3: Mean cycle time loss from setup distribution (left: same / right: non-symmetric).

Figure 4 shows the error with average lot size. The left graph describe the type 1 arrival result and the right one illustrate the type 3 arrival result. With type 1 arrival, the error is decreasing with increasing average lot size. With type 3 arrival, the error is independent with the average lot size. ERM show small error about any average lot size and any type of arrival. This is due to linear and affine model cannot reflect the parallel characteristics into their model otherwise ERM can do.

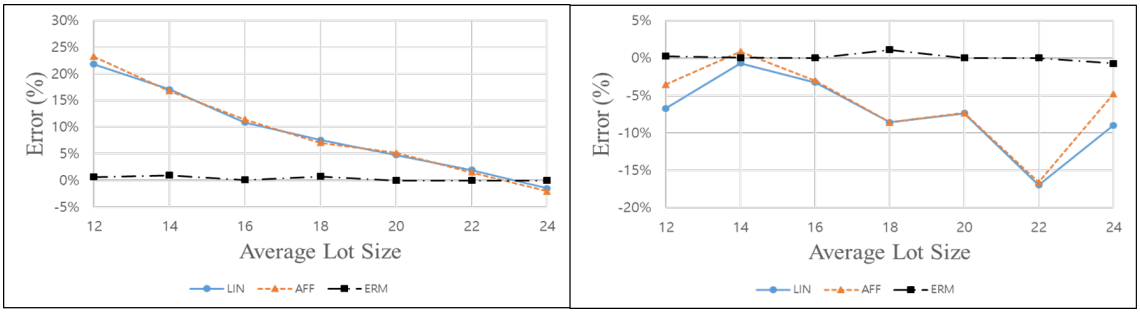

Figure 4: Mean cycle time error by average lot size (left: type 1 arrival / right: type 3 arrival).

Figure 5 shows the loss by the increasing lot length. The left graph is the loss for mean cycle time and the right one is the loss for total makespan. By the longer lot length, the loss tends to increase because longer lot length means that processing time variation is much larger.

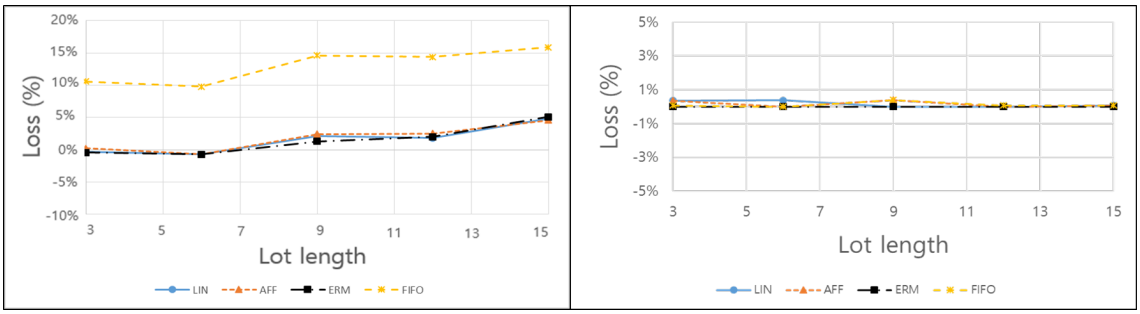

Figure 5: Loss by lot length (left: mean cycle time / right: total tardiness).

The total tardiness case shows similar results to the mean cycle time case and will not be shown. This may be due to the fact that the total tardiness optimization problems have a similar structure to the mean cycle time problems, using due dates information D_l instead of arrival time a_l.

Table 4: Run time.

		Computation Time (s)											
		Mean cycle time				Total makespan				Total tardiness			
		LIN	AFF	ERM	FL	LIN	AFF	ERM	FL	LIN	AFF	ERM	FL
Number of Lots	1	0.016	0.016	0.016	0.031	0.005	0.007	0.007	0.021	0.005	0.007	0.0100	0.020
	5	1.186	1.435	1.841	10.452	0.421	0.564	0.487	3.786	0.046	0.0400	0.0260	6.246
	10	358.660	480.018	1097.429	3600.950	78.390	100.043	128.170	3601.600	3600.261	3600.521	3600.261	3601.550
	>20	3600.464	3600.282	3600.233	3600.960	3600.418	3601.822	3600.901	3601.370	3600.717	3600.558	3600.980	3600.890

Table 4 describe the run time results. All the above 20 lots case, it spend whole our time limits to solve problems. Without this case, we tried to calculate the relative computation time. ERM needs 1.5 times and 1.2 times computation than linear and affine model.

5 CONCLUDING REMARKS

In this paper, we developed MILP sequence scheduling models different from the error prediction using FIFO schedules in park et al (2017b) and evaluate the loss of using reduced models in place of the true model for three objective functions. The reduced models considered have low loss in most cases and even outperform the flow line model for a large number of lots under time limits. ERMs exhibit the best performance among them in terms of error and loss. Affine models perform similarly to ERMs in terms of loss in all cases, but are shown to have high errors. The linear model cannot properly express class dependent setups and exhibits high loss and high errors. We used different average lot sizes and different setup distributions in order to give more variance to the lot schedules.

In the future, we will consider greater variation in the processing times and allow different arrival or ready times for each lot. Additionally, as the flow line MILP model was computationally prohibitive and time limits had to be used, learning techniques or heuristics could be utilized to produce more accurate results. Comparing the loss of using ERMs and of using the heuristics/other approaches could be an interesting direction for further research.

REFERENCES

Bitar, A., S. Dauzère-Pérès, and C. Yugma. 2014. "On the Importance of Optimizing in Scheduling: The Photolithography Workstation". In *Proceedings of the 2014 Winter Simulation Conference*, A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 2561-2570. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cakici, E. and S. J. Mason. 2007. "Parallel Machine Scheduling Subject to Auxiliary Resource Constraints". *Production Planning and Control* 18(3):217-225.

Chiou, C. W. and M. C. Wu. 2014. "Scheduling of Multiple In-line Steppers for Semiconductor Wafer Fabs". *International Journal of Systems Science* 45(3):384-398.

Kabak, K. E., C. Heavey, V. Corbett and P. J. Byrne. 2013. "Impact of Recipe Restrictions on Photolithography Toolsets in an ASIC Fabrication Environment". *IEEE Transactions on Semiconductor Manufacturing* 26(1):53-68.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011. "A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations". *Journal of scheduling* 14(6):583-599.

Morrison, J. R. 2010. "Deterministic Flow Lines with Applications". *IEEE Transactions on Automation Science and Engineering 7*(2): 228-239.

Morrison, J. R., and D. P. Martin. 2007. "Performance Evaluation of Photolithography Cluster Tools". *OR spectrum* 29(3):375-389.

Park, Y. J., and H. R. Hwang. 2013. "Minimization of Total Processing Time in Semiconductor Photolithography Process". In *Applied Mechanics and Materials* 325: 88-93.

Park, K., and J. R. Morrison. 2015. "Controlled Wafer Release in Clustered Photolithography Tools: Flexible Flow Line Job Release Scheduling and an LMOLP Heuristic". *IEEE Transactions on Automation Science and Engineering* 12(2):642-655.

Park, J. Y., K. Park, and J. R. Morrison. 2017(a). "Models of Clustered Photolithography Tools for Fab-level Simulation: From Affine to Flow Line". *IEEE Transactions on Semiconductor Manufacturing* 30(4): 547-558.

Park, J. Y., K. Park, and J. R. Morrison. 2017(b). "Exit Recursion Models of Clustered Photolithography Tools for Fab Level Simulation". *IEEE Transactions on Semiconductor Manufacturing 30*(1): 39-51.

Ramírez-Hernández, J. A., and E. Fernandez. 2009. "A Simulation-based Approximate Dynamic Programming Approach for the Control of the Intel Mini-fab Benchmark Model". In *Proceedings of the 2009 Winter Simulation Conference*, M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1634-1645. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Riedmiller, S. and M. Riedmiller. 1999. "A Neural Reinforcement Learning Approach to Learn Local Dispatching Policies in Production Scheduling". In *International Joint Conference on Artificial Intelligence,* July 31th -August 6th, Stockholm, Sweden, 2, 764-771.

Wein, L. M. 1988. "Scheduling Semiconductor Wafer Fabrication". *IEEE Transactions on semiconductor manufacturing 1*(3):115-130.

Wu, M. C. and C. W. Chiou. 2010. "Scheduling Semiconductor In-line Steppers in New Product/Process Introduction Scenarios". *International Journal of Production Research* 48(6):1835-1852.

Yan, B., H. Y. Chen, P. B. Luh, S. Wang, and J. Chang. 2013. "Litho Machine Scheduling with Convex Hull Analyses". *IEEE Transactions on Automation Science and Engineering* 10(4):928-937.

Yoon, H. J. and D. Y. Lee. 2004. "Deadlock-free Scheduling of Photolithography Equipment in Semiconductor Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 17(1):42-54.

AUTHOR BIOGRAPHIES

HYEONG-OOK KIM is M.S. student in the Department of Industrial and Systems Engineerings, KAIST, Daejeon, South Korea. He holds B.S. degrees in Industrial Engineering and minor in Economics from Ajou University, South Korea. His email address is hyung1405@kaist.ac.kr.

SE-HYEON PARK is M.S. student in the Department of Industrial and Systems Engineerings, KAIST, Daejeon, South Korea. He holds B.S. degrees in Mathematical Science and minor in Management Science from UNIST, South Korea.. His email address is sehyeon.park@kaist.ac.kr.

JUNG YEON PARK received the M.S. and B.S. degrees in the Department of Industrial and Systems Engineering from KAIST, Daejeon, South Korea. He worked as a software engineer at Samsung Electronics, DS division, Suwon, South Korea. His email address is jpark0@kaist.ac.kr

JAMES R. MORRISON is an Associate Professor in the Department of Industrial and System Engineering, KAIST, Daejoen, South Korea. He holds B.S. degrees from the University of Maryland at College Park, USA and M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, USA. He is a co-Chair of the IEEE RAS Technical Committee on Semiconductor Manufacturing Automation. His email address is james.morrison@kaist.edu