Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

QUEUE TIME PREDICTION METHODOLOGY IN SEMICONDUCTOR FAB

Donguk Kim
Byeongseon Lee
Sangchul Park

Department of Industrial Engineering
Ajou University
World Cup Street 206
Suwon, 16499, REPUBLIC OF KOREA

ABSTRACT

This paper presents a methodology for predicting queue times in semiconductor fabrication, where numerous complex and costly pieces of equipment are utilized. Queue time, occurring between continuous single or multi-processes, is a crucial factor affecting the quality of wafers, which can significantly impact costs. While most semiconductor fabrications use queue time limits as a key dispatching factor, some wafers may still be scrapped or reworked. By predicting queue times, we can reduce unnecessary waste by blocking or re-dispatching wafers. Two approximations are proposed and compared based on accuracy and prediction time: a machine learning model trained using experimental results and a multi-resolution simulation model with varying fidelity levels. The simulation model is validated using the SMAT2022 data set.

1 INTRODUCTION

A semiconductor FAB, a manufacturing system consisting of hundreds of complex and expensive equipment, has become increasingly larger in scale compared to the past. As a result, the automated material handling systems (AMHS) for transporting wafers inside the FAB, such as overhead hoist transport (OHT), track, buffer, and stocker, have become more complex. Therefore, modeling such semiconductor FAB can be performed by considering various technical constraints, hundreds of unit processes and reworks, and complex characteristics (Kopp et al. 2020). For example, some processes restrict the start time of the following process within a specified time to prevent oxidation or contamination on the surface of wafers (Scholl and Domaschke 1994). Accordingly, it is crucial to meet the queue time, the restricted time between such specific processes. Exceeding the queue time can result in wafers being reworked or scrapped, which can have negative impacts on production volume and cost efficiency (Tu et al. 2010). Queue time is generally defined as the period between the end time of the previous process and the start time of the following process. Moreover, there may exist single or multi processes between the two processes, which subject to queue time constraints. In most cases, constraints regarding queue time are incorporated into equipment dispatching rules. However, even between single processes, there may be pre-processing tasks such as measurement, batch size configuration, and preventive maintenance (PM) that must be performed, resulting in a violation of queue time constraints. If the queue time can be predicted in advance, it would be possible to prevent the violation by blocking wafers entering equipment or re-dispatching to other equipment.

However, the constraint of queue time makes production scheduling more difficult and complex (Klemmt et al. 2012). As a result, predicting queue time in semiconductor FABs with such complex production environments is very challenging. For this reason, most research on queue time in semiconductor FABs has focused on the constraint of queue time and scheduling methodologies. Robinson

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

and Giglio (1999) proposed a methodology for selecting inter-process time constraints and predicting the probability of rework in the process. Yu et al. (2013) proposed a lot scheduling methodology using time constraints between two processes. Sadeghi et al. (2015) proposed an approach using a disjunctive graph model and a list scheduling algorithm to estimate the probability of satisfying inter-process time constraints. However, there are only a few papers that address predicting queue time. Lee et al. (2020) collected predictive variables that are believed to influence waiting times based on information on work-in-process (WIP) at the start of the process and dispatching rule information. He used them to create a deep learning-based prediction model to forecast the waiting times of wafer lots with time constraints between processes. However, challenging to construct a robust predictive model for queue time in a complex FAB system using only WIP and dispatching information.

The objective of this study is to propose a new prediction methodology for queue time in a semiconductor FAB. Two approximations are proposed to predict the queue time and compared in terms of accuracy and prediction time. Section 2 explains two methodologies and compares their respective strengths and weaknesses. Section 3 evaluates how well each methodology can predict queue times through the experiments. Conclusions and future research directions are presented in Section 4.

2 METHODOLOGY FOR QUEUE TIME PREDICTION

2.1 Artificial Intelligence Model

The artificial intelligence (AI) model used in this paper is a data-driven machine learning prediction model that trains and creates a model using explanatory and response variables and predicts the target using the generated model. The biggest characteristic of AI-based prediction models is that they are heavily influenced by data. Therefore, the accuracy of the model can vary significantly depending on the quantity and quality of the data used as explanatory variables. However, once the prediction model is created, the time required for making predictions is very short, and the performance of the model can be improved by using additional data.

2.2 Multi-Resolution Simulation Model

Multi-resolution modeling refers to representing a target system using models of different resolutions that are at different abstraction levels, depending on the objectives of the modeling. And each model at different resolutions represents a single target system in a complete form. Unlike AI prediction models, whose accuracy can vary significantly depending on the quantity and quality of the collected data, simulation-based prediction models can have significantly different prediction times depending on the level of resolution. Multi-resolution modeling allows simulations to be conducted at various resolutions as needed. In this approach, the objective can be achieved by performing additional modeling on the existing model (Hong 2014; Song et al. 2022).

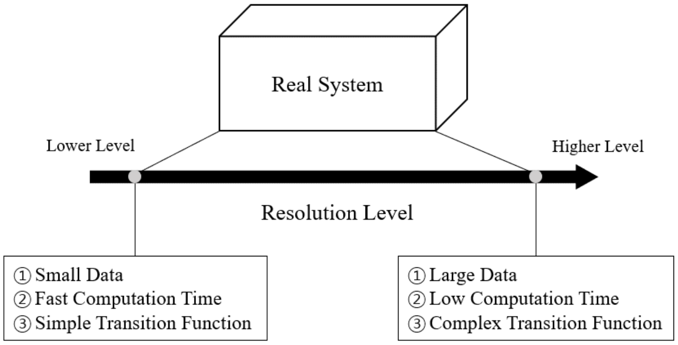

Figure 1: Multi-resolution modeling.

Figure 1 is a diagram that classifies the multi-resolution simulation model according to resolution levels. As the resolution level increases, simulation models require more data and can achieve higher accuracy, but there is a possibility of slower simulation performance. On the other hand, as the resolution level decreases, simulation models require fewer data, and although the model's accuracy is relatively lower, simulation performance may be faster due to the higher simulation acceleration.

3 EXPERIMENTS FOR PREDICTION METHODOLOGIES

3.1 SMT2020 Data Set

To validate the two queue time prediction methodologies proposed in this paper, a simulation model combining an automated material handling system (AMHS) and a FAB layout was used based on the SMT2020 dataset (Kopp et al. 2020). The SMT2020 dataset is classified into four types of datasets, as shown in Table 1, to provide a realistic experimental environment for discrete-event simulation researchers in the semiconductor manufacturing industry.

Table 1: Classification of SMT2020 data set.

Set	Feature	High volume Low Mix	Set	Low volume High Mix
1	Plan Type	Make to stock (Push)	2	Make to order (Pull)
	App. Target	Logic device		Logic or memory device
	Number of products	2 products		10 products
	Due date	No due date		Required
	Engineering Lot	Not contain		Not contain
3	Plan Type	Make to stock (Push)	4	Make to order (Pull)
	App. Target	Logic device		Logic or memory device
	Number of products	2 products		10 products
	Due date	No due date		Required
	Engineering Lot	Contain		Contain

For this experiment, the most complex and largest dataset 4 was used, which is suitable for mass production of small varieties of products, with a make-to-order (MTO) planning type that includes hot lots and engineering lots. There are a total of 10 product types in the dataset, with each product having 242-583 different process steps. Based on the characteristics of each process, they are broadly classified into 11 functional categories. In addition, the 11 process types are composed of 105 equipment groups with different types of equipment: table, batch, and cascading equipment types, depending on the characteristics of the equipment.

3.2 SMAT2022 Data Set

In this experiment, we added an AMHS-based logistics model to the SMT2020 process model for a more specific semiconductor FAB environment. Lee et al. (2022) added logistics information such as bay layout, buffer, and vehicles to the SMT2020 model to create a new dataset called SMAT2022. In SMAT2022, the OHTs move along the bay layout and transport lots. The logistics layout was configured using the spine type commonly used in FABs.

Table 2: Classification of SMT2020 data set.

Category	Group	Feature
Production	All features of SMT2020 data set	
Material Handling	Layout	Spine configuration
		3 Inter-bays
		40 Intra-bays
	AMHS	500 OHTs
		734 ZCUs
	Buffers	18000 STB/UTBs
		40 Stockers

Table 2 shows the process and logistics models used in SMAT2022. Based on Tables 1 and 2, the simulation model based on SMAT2022 and the FAB layout used in the experiment were constructed, resulting in Figure 2. The SMT2020 dataset 4 used in the experiment contains queue time defined between various process steps, and the experiment was conducted by selecting one representative step of the main process with queue time constraints in the table, batch, and cascading type equipment.

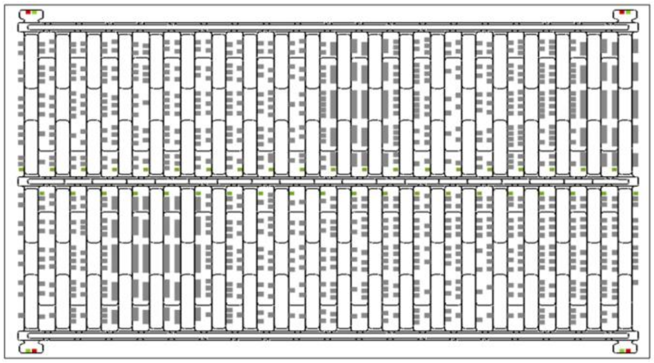

Figure 2: Layout of FAB model.

3.3 Two Consideration for AI base Queue Time Prediction Model

In the first experiment, we performed queue time prediction based on an AI model. To predict queue time using the AI model, we considered two main factors. The first is determining the prediction timing, which can be broadly classified into two categories in this experiment, as shown in Figure 3.

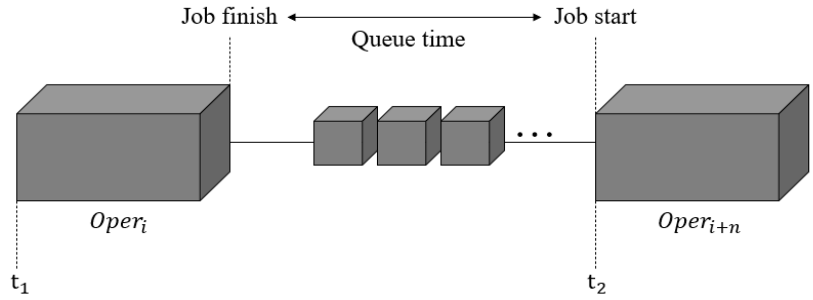

Figure 3: Two points of prediction time t_1, t_2.

In the case of t_1, it represents the dispatching timing of the starting process with queue time constraints. If the prediction timing is set to t_1, it would be possible to prevent the waste of wafers by preventing the input of lots that exceed the predicted queue time at the dispatching timing or by re-dispatching them. On the other hand, t_2 represents the dispatching timing of the ending process where the queue time constraint ends. In this case, since the processes with queue time constraints have already been performed, it is necessary to start the ending process within the remaining time. Therefore, it is not possible to prevent lot input itself, and if there are no idle equipment available, there may be cases where the queue time is exceeded. In this experiment, t_1 was set as the prediction timing so that the prediction results could be used for equipment re-dispatching.

The second consideration was which variables to use as explanatory variables for training the prediction model. Since t_1 was set as the prediction timing, only variables that could be collected at t_1 had to be used for training the model. We defined the six variables with the highest explanatory power, as shown in Table 3. Table 4 represents the variables which are considered before the variables of Table 3.

Table 3: Input & output variables of the prediction model.

Input variable	x_1	Average queue time of the last 3 measurements
	x_2	WIP count of tool group $oper_{i+n}$
	x_3	Average of the time stored in the buffer for the last 3
	x_4	The ratio of available equipment count to the total equipment
	x_5	Lot priority
	x_6	Critical ratio
Output variable	λ	Queue time

Table 4: Previous variables of the prediction model.

Input variable	x'_1	The average length of the route that the last three lots passed through.
	x'_2	The average delivery time of the last three lots
	x'_3	Whether the lot is a hot lot or not
	x'_4	Whether the lot is an engineering lot or not
	x'_5	Total number of equipment
	x'_6	The ratio of WIP count to equipment count
	x'_7	The number of equipment with preventive maintenance (PM)
	x'_8	The number of OHTs passing through the target route
Output variable	λ'	Queue time

x_1 in Table 3 represents the average queue time of the last three measurements passing through the same $oper_i$ and $oper_{i+n}$, x_2 represents the WIP count of the tool group $oper_{i+n}$, and x_3 represents the average time stored in the last three buffers. Additionally, x_4 represents the ratio of available equipment in the tool group $oper_{i+n}$ to the total equipment, x_5 represents the priority of the lot, and x_6 indicates the ratio of the remaining delivery date and remaining process time for the lot.

3.4 Simulation Environment for Experiments

In the first experiment, the variables in Table 3 are collected from SMAT2022 simulation model mentioned in Section3.2. Simulation time was set to 180 days of warm-up period and 180 days of simulation period. The target step for the experiment was limited from 034 wet etch to 035 lithography, with a tool group of table type and a queue time limit of 2 hours. The process time is set at 1.14min and 2.78 min respectively.

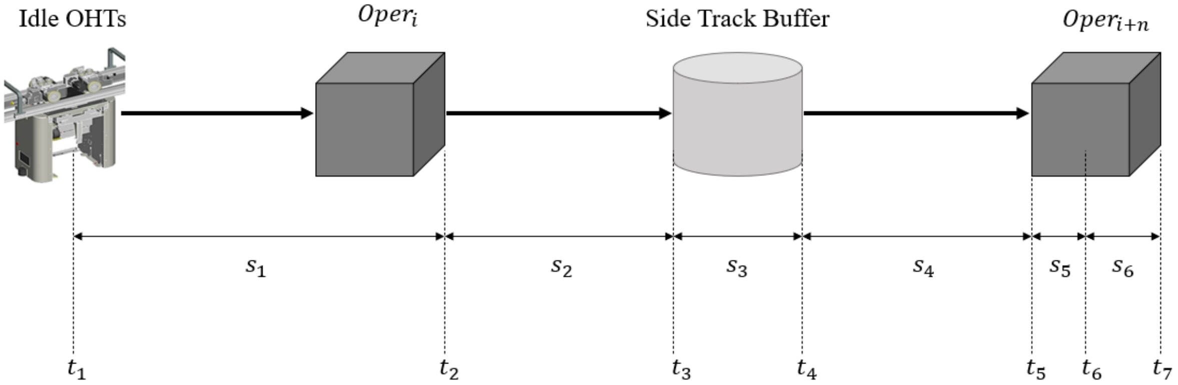

Figure 4: Collecting time points of input variables in the simulation model.

Each variable was collected at points t_1 through t_7, as shown in Figure 4. In this case, the queue time λ can be obtained by summing the differences between t_2 and t_6, which corresponds to the sum of s_2 through s_5. Table 5 represents annotations from t_1 to t_6.

Table 5: Annotations from t_1 to t_6.

t_1	Dispatching time of tool group $oper_i$
t_2	Process completion time in tool group $oper_i$
t_3	Arrival time at STB
t_4	Departure time at STB
t_5	Arrival time at tool group $oper_{i+n}$
t_6	Process starting time in tool group $oper_{i+n}$

3.5 LGBM based Queue Time Prediction Model

In the first experiment, a total of 1,638 queue time logs were collected, and we utilized this data as explanatory variables to build a light gradient boosting machine (LGBM) regression model. LGBM is a framework based on the gradient boosting algorithm, which utilizes tree-based learning. It has the advantage of high accuracy, efficient memory usage and fast speed (Ke et al. 2017). Hyperparameters that affect the model's performance and learning rate were set to build a LGBM prediction model. Table 6 below shows the hyperparameters we use.

Table 6: Annotations from t_1 to t_6.

num_leaves	31
$lerning\_rate$	0.1
$num\_estimator$	100
$\max\ \_depth$	-1
$\min\ \_child\_samples$	20
reg_lambda	0

Figure 5 shows the queue time data collected from the simulation results, where the x-axis represents the name of the lot, and the y-axis represents the queue time of the lot. A total of 102 data points, accounting for approximately 6% of the entire dataset, exceeded the queue time constraint of 2 hours.

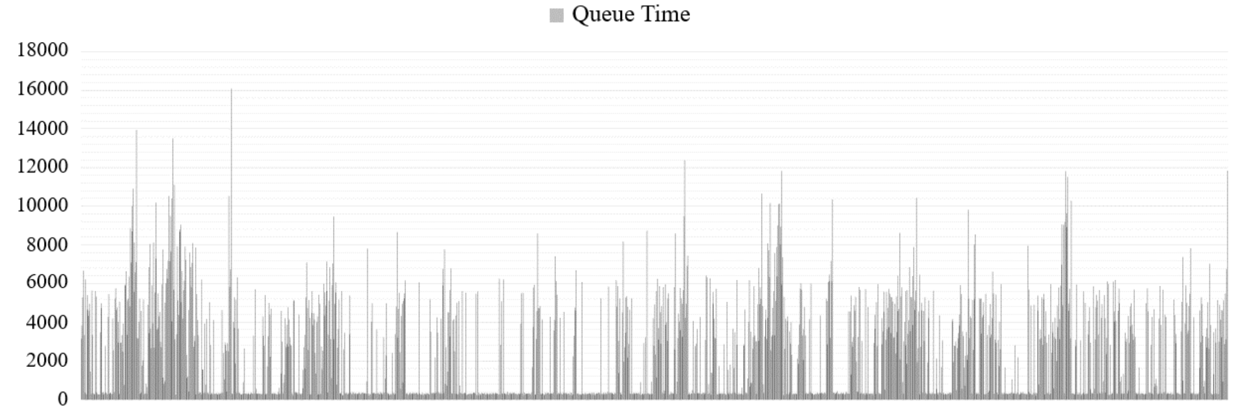

Figure 5: Collected queue time data.

The training results are shown in Table 7, indicating that the percentage of predicted queue times within ±300 seconds of the actual queue time was 31%, within ±600 seconds was 50%, and within ±900 seconds was 65%. Considering the processing time of the target process, which is 167.04 seconds, these results imply that the performance of the prediction model is not very high. The reason for the low model performance can be found in the explanatory variables used in the model training. This is because the characteristics of the variables collected at t_1 cannot fully reflect the situation at t_6. That is, due to changes in the WIP state and equipment status resulting from the dispatching results during the time between the two points, it is difficult to accurately predict the queue time using only the explanatory variables collected at t_1.

Table 7: Prediction results of AI model.

ε bound (sec)	Average (sec)	Processing time (sec)	In-bound rate (%)
± 300	2,414	167.04	31
± 600			50
± 900			65

3.6 Experiments for Multi-resolution Simulation Model based Queue Time Prediction Model

In the second experiment, queue time prediction was performed based on both a high-resolution simulation model and a low-resolution simulation model, and their prediction accuracy and time were compared. The high-resolution simulation model was also constructed based on the SMAT2022 dataset, incorporating all process models and OHT levels. In contrast, the low-resolution simulation model was developed by collecting OHT delivery times from the simulation logs of the high-resolution model. Consistent with the previous experiment, we conducted simulations over a 180-day warm-up period followed by a 180-day simulation period.

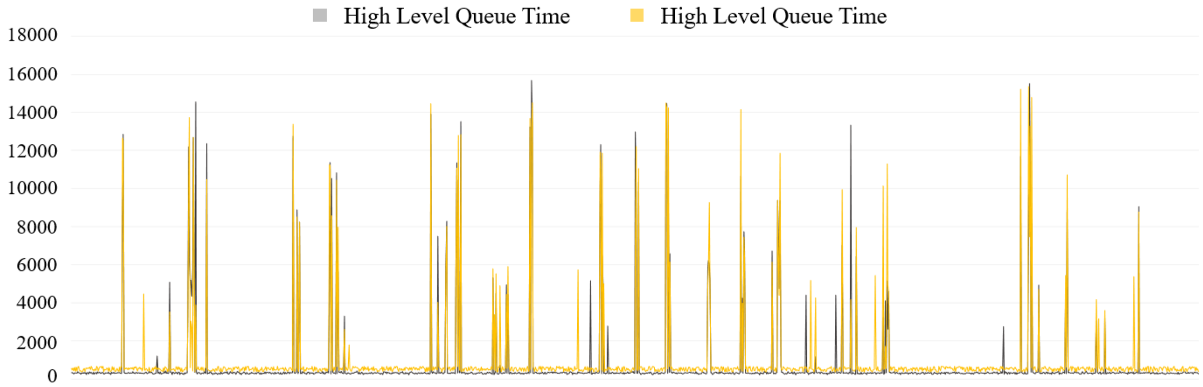

Figure 6: Comparison of multi-resolution simulation models.

The black data in Figure 6 represents the results from the high-resolution simulation, while the yellow data represents the results from the low-resolution simulation. When comparing the results of each simulation, Table 8 shows that the error range ε of the low-resolution simulation results compared to the high-resolution simulation model were within ±300 seconds for 76% of the cases, ±600 seconds for 95% of the cases, and ±900 seconds for 96% of the cases, indicating a high level of fidelity. Furthermore, the average computation time of each model was found to be 7,066ms for the high-resolution simulation model and 221 ms for the low-resolution simulation model.

Table 8: Prediction results of AI model.

ε bound (sec)	Computation time of high-resolution model (ms)	Computation time of low-resolution model (ms)	In-bound rate (%)
± 300	7,066	221	76
± 600			95
± 900			96

4 CONCLUSION

In this study, two approaches for predicting queue time in semiconductor fabs were presented: an AI-based queue time prediction methodology and a multi-resolution simulation-based queue time prediction methodology. The strengths and weaknesses of each approach were discussed, and experiments were conducted. To predict queue time based on AI models, we set the prediction time and explanatory variables and trained using the LGBM learning algorithm, which is a tree-based algorithm. The percentage of predicted queue times within ±300 seconds was 31%, within ±600 seconds was 50%, and within ±900 seconds was 65%, indicating that the model's performance was not satisfactory considering the processing time of the target process.

To improve the prediction accuracy, we performed queue time prediction based on multi-resolution simulation and lowered the model resolution to secure fast computation time. Comparing the low-resolution model with the high-resolution model, the low-resolution model showed a 76% match with the high-resolution model in terms of the error range of ±300 seconds and a 95% match for the error range of ±600 seconds and a 96% match for the error range of ±900 seconds. The average computation time required for queue time prediction was $221ms$ for the low-resolution model, which is more than 30 times faster than the high-resolution model. This indicates that when it is not possible to extract the correct explanatory variables for AI model construction in queue time prediction that requires both accurate prediction results and fast prediction times, a multi-resolution simulation model can be used.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation (NRF-2020R1A2C1004544) grant by the Korean government (MSIT); the Institute for Information and Communications Technology Promotion (IITP-2021000292) grant by the Korean government (MSIT); and the Ministry of Trade, Industry & Energy (MOTIE, Korea) of the Republic of Korea (RS-2022-00155650).

REFERENCES

Hong, S. Y., 2014. "Integration Architecture for Multi-Resolution Modeling and Simulation", In *Proceeding of The Korea Society For Simulation Conference*, November 8th, Ajou University, Korea, 73-75.

Klemmt, A., and L. Mönch. 2012. "Scheduling Jobs with Time Constraints between Consecutive Process Steps in Semiconductor Manufacturing". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, 7412-7423. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2000. "Integrating Critical Queue Time Constraints into SMT2020 Simulation Models". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1813-1824. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2000. "SMT2020—A Semiconductor Manufacturing Testbed". *IEEE Transactions on Semiconductor Manufacturing*, 33(4):522–531.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", In *Advances in Neural Information Processing Systems 30(NIPS 2017)*, December 5th-4th, Long Beach Convention Center, California, USA, 30.

Lee, G. H., S. U. Cheon, and S. C. Park, 2020, "Queue-Time Prediction for Wafer Lot with Time-Constraints, Production Control in Semiconductor Manufacturing with Time Constraints". *Korean Journal of Computational Design and Engineering*, 25(4):343-349.

Lee, K. W., S. Y. Song, D. S. Chang, and S. C. Park. 2022. "A new AMHS Testbed for Semiconductor Manufacturing". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, 3318-3325. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Robinson, J. K., and R. Giglio. 1999. "Capacity Planning for Semiconductor Wafer Fabrication with Time Constraints between Operations". In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 880-887. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sadeghi, R., S. Dauzère-Pérès, Yugma, C., and G. Lepelletier. 2015. "Production Control in Semiconductor Manufacturing with Time Constraints", In *26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). IEEE*, May 3rd-6th, Saratoga Springs, New York, USA, 29-33.

Scholl, W., and J. Domaschke. 1994. "Implementation of Modeling and Simulation in Semiconductor Wafer Fabrication with Time Constraints between Wet Etch and Furnace Operations". *IEEE Transactions on Semiconductor Manufacturing*, 13(3):273-277.

Song, S. Y., K. W. Lee, and S. C. Park. 2022. "Semiconductor FAB AMHS Simulation using Multi-Resolution Model", In *Proceeding of Computational Design and Engineering Conference*, Incheon National University, Korea, 2251-2252.

Tu, Y.-M., H.-N. Chen, and T.-F. Liu. 2010. “Shop-Floor Control for Batch Operations with Time Constraints in Wafer Fabrication”. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 17(2):142-155.

Yu, T. S., H. J. Kim, C. Jung, and T. E. Lee. 2013. “Two-stage lot scheduling with waiting time constraints and due dates”. In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3630-3641. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

DONGUK KIM received a bachelor degree (2018) in industrial engineering and a master degree (2021) in industrial engineering, Ajou University, Korea. He is now a Ph. D candidate in industrial engineering, Ajou University, Korea. He is interested in simulation-based scheduling and planning, digital manufacturing, and PHM for plant. His email address is dek1603@ajou.ac.kr

BYEONGSEON LEE received a bachelor degree (2022) in industrial engineering, Ajou University, Korea. He is now a Master’s student in industrial engineering, Ajou University, Korea. He is interested in simulation-based digital manufacturing systems, production scheduling and automated material handling system. His email address is tjs5007@ajou.ac.kr

SANGCHUL PARK was granted his B.S.(1994), M.S.(1996) and Ph.D.(2000) degrees in industrial engineering, Korea Advanced Institute of Science and Technology(KAIST). He is a professor in Department of industrial engineering, Ajou University, Republic of Korea, since 2004. He is interested in modeling and simulation, combat simulation for defense, and digital manufacturing system. His email address is scpark@ajou.ac.kr