Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 100 (2021) 49–54

www.elsevier.com/locate/procedia

31st CIRP Design Conference 2021 (CIRP Design 2021)

Data analytics for time constraint adherence prediction in a semiconductor manufacturing use-case

Marvin Carl May[a,*], Sören Maucher[a], Andrea Holzer[b], Andreas Kuhnle[a], Gisela Lanza[a]

[a]*wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany*

[b]*Infineon Technologies AG, Wernerwerkstraße 2, 93049 Regensburg, Germany*

Abstract

Semiconductor manufacturing represents a challenging industrial environments, where products require more than several hundred operations, each representing the technical state-of-the-art. Products vary greatly in volume, design and required production processes and, additionally, product portfolios and technologies change rapidly. Thus, technologically restricted rapid product development, stringent quality related clean room requirements and high precision manufacturing equipment application enforce operational excellence, in particular time constraints adherence. Product specific time constraints between two or more successive process operations are an industry-specific challenge, as violations lead to additional scrapping or reworking costs. Time constraint adherence is linked to dispatching and currently manually assessed. To overcome this error-prone manual task, this article presents a data-based decision process to predict time constraint adherence in semiconductor manufacturing. Real-world historical data is analyzed and appropriate statistical models and scoring functions derived. Compared to other relevant literature regarding time constraint violations, the central contribution of this article is the design, generation and validation of a model for product quality-related time constraint adherence based on a real-world semiconductor plant.

Keywords: Smart Manufacturing Systems; Industry 4.0; Semiconductor Industry; Production Planning and Control

1. Introduction

Semiconductor manufacturing involves some of the most complex manufacturing processes and due to several hundred required operations operates on the verge of the physically and operationally possible [22]. Thus, operational excellence is focused amid global competition [19], inducing stringent quality requirements. Predominantly the adherence to time constraints that define a maximum time between two or more successive steps and high machine utilization remain pillars of competitiveness [1]. Any violation of time constraints can result in a batch being scrapped or requiring additional reworking [2].

Operational decisions, such as dispatching against the background of strict time constraints, in semiconductor manufacturing are currently often performed by humans, so that an operator takes the time-consuming and stressful evaluation whether or not a particular lot can be sent to the consecutive machine. If the consecutive machine, for instance, is fully utilized, the processing of the lot is expected to take longer, which implies that a corresponding time constraint might be violated [16]. Hence, a potential approach could be to automate this process with the benefit of minimizing time constraint violations and allowing more complex product designs.. Less time constraint violations result in fewer lots being scrapped or reworked and, thus, lead to lower costs. To design an automated system that mitigates time constraint violations, a promising approach is to integrate historical data. In general, this data can be extracted from the company's Manufacturing Execution System (MES) [21]. The approach is not limited exclusively to semiconductor manufacturing, however, the treatise, in this complexity, typically only arises in this complexity in high variety high volume production, such as semiconductor manufacturing.

Therefore, different time constraints are outlined in a literature review in Section 2. Section 3 introduces the require statistical foundations and enables the loss function derivation in Section 3.2. Following the prediction model introduction in Section 4, a case study is presented in Section 5. Lastly, a discussion and outlook conclude this paper in Sections 6 and 7.

* Corresponding author Tel:+49-1523-950-2624; Fax:+49-721-60845005
E-mail address: marvin.may@kit.edu (Marvin Carl May).

10.1016/j.procir.2021.05.008

2. Related Work

Technological restrictions, rapid product development and various product design interdependencies in the semiconductor industry lead to operational excellence as means a main enabler, in particular to tackling product design related time constraints. Operations management in the semiconductor industry is typically based on the Production Planning and Control (PPC) hierarchy [22], that distinguishes between planning, i.e. decisions that refer to a time horizon ranging from months to years and include for instance capacity, material requirements and demand planning, and control, i.e. short-term decisions about operation shifts, order release, scheduling released orders on machines and dispatching these under time constraints. Hence, a variety of time constraint definitions is explored and a comprehensive literature review on the lowest PPC level performed.

2.1. Time Constraint definitions

According to the basic definition of Lima et al. [16], each time constraint (t_{s_1,e_1}) has a start (s_1), an end process step time (e_1) and a time limit (t_{limit}) that should not be exceeded for the lot to retain its expected properties: $e_1 - s_1 \leq t_{limit}$. In the context of semiconductor manufacturing, minimal waiting time constraints are not considered, as those time constraints can trivially be fulfilled by holding the corresponding lot back. Wang et al. [24] hence use the appropriate term of limited waiting time constraints, indicating that the waiting time of a lot should be limited. Maleck and Eckert [18] uses the term timelink area for two or more consecutive steps, where simple timelinks span two consecutive production steps. More complex forms are known as time constraint tunnels (TCT), which are consecutive time constraints that overlap [17]. Formally described, this means that there are at least two time constraints t_{s_1,e_1}, t_{s_2,e_2} such that $s_1 \leq e_2$ and $s_2 \leq e_1$. These differ from timelinks insofar as the individual time constraints can be competing, sometimes requiring holding back lots before the second time constraint starts. Secondly, the process that is time constrained can be classified as described by Arima et al. [2], one example is the queue time constraint, which in their paper is defined as limiting the time between two consecutive process steps. Another type is the transportation time constraint, limiting the time from the intermediate buffer storage to the target machine, as described by Kim et al. [11]. A third category is the wafer residency time constraint mentioned by Pan et al. [23] and Yang et al. [29], which limit the sojourn time in different process modules, such as chemical vapor deposition. All in all, this paper addresses simple queue timelink areas.

2.2. Literature Review

The literature identified throug a grounded theory literature review can be clustered according to their PPC classification, where scheduling refers to assigning released orders to machines prescriptively and dispatching controls operations on the fly. Thus, approaches can be clustered according to their objective, modeling and solution technique as shown in Table 1.

Table 1. Classification of relevant literature

PPC	Objective	Modeling	Solution	Ref.
Scheduling	Optimal schedule deriv.	Enum. Tree	Branch & Bound	[28]
	Auxiliary time constraints	Experiments	Genetic Alg.	[20]
	Min. violations, cycle time	MILP	heuristic	[10]
	Reduce violations	MILP	Branch & Bound	[12]
	Max. production rate	MILP	Numerical anal.	[14]
	Predict WIP & thresholds	Forecast model	Neural Network	[5]
	Improve cycle time	MILP & Exp.	Optimization	[13]
Dispatching	Reduce failure rate	MES & Exp.	Heuristic	[15]
	Queue length opt. policy	MDP	Optimization	[27]
	Flow line control policy	MDP & Exp.	Optimization	[26]
	Sojourn time control	MILP	Branch & Bound	[18]
	Completion time bounds	Experiments	Heuristic	[25]
	Predict violation prob.	Graph model	Heuristic	[17]
	Increase utilization	MDP	RL	[1]

In respect to scheduling, the literature proposes the solution through branch-and-bound schemes to derive optimal cyclic schedules for instance for simplified two-machine [28, 14] or three-machine flow shops [12]. In scheduling modeling is often performed with Mixed Integer Linear Programming (MILP), which either minimizes the number of time constraint violations in the objective function [10] or models time constraint as linear auxiliary constraints [13]. Another approach controls the Work-in Progress (WIP) Level as higher WIP thresholds increase the time constraint violation likelihood. When maximizing the production rate with relaxed time constraints Lee and Li [14] achieve 0.2% violation rate. The dire need to provide quick, yet not perfect but acceptable, solutions is addressed by heuristics [10], for instance full and greedy batching [20]. Finally, data-based approaches start to emerge by predicting WIP levels and imposing rigid thresholds through a hybrid decision tree and neural network approach [5].

Regarding dispatching, deriving a control policy [27] or predicting violation probabilities [17] serve as objectives, whereas modeling is predominantly performed through the application of queuing theory [26], MILP [18] or the simulation of Markov Decision Processes (MDP) [1]. Beyond optimization approaches [27, 18], heuristic solutions focus on comparing algorithms within an MES integration [15], triggering dispatching based on thresholds [17] as well as determining suitable upper and lower acceptance bounds [25]. Traditional methods in practice, however, are typically based on operators that trigger dispatching.

2.3. Research Question derivation

A commonality of the identified literature, regardless of their PPC level, are their strong assumptions, i.e. few machines or simple machine setups. Additionally, most variables, such as processing and arrival times, are assumed to follow predetermined distributions. If applicable, results are evaluated in simulations, yet real-world data is neglected. Thus, the goal of this research is to design a production control method for real-world time constraint adherence improvements , that is based on observable real-time data, with no restrictions to the number of machines, states or detailed distributional knowledge.

3. Statistical Foundation

The designed time constraint dispatching support system, in contrast to the reviewed literature, is based on individual historic transition time realizations. Hence, it is necessary to review prediction intervals, as well as accurate loss functions to determine the most suitable prediction interval.

3.1. Prediction Intervals

Prediction intervals are intervals from which the next realized data point will be sampled with a given probability. Consider stochastic variables $y_1, ..., y_n$ that are independently and identically distributed with a normal distribution of $N(\mu, \sigma^2)$. The corresponding prediction interval is based on a $1-\alpha$ coverage level, yet the real μ and σ are not known, and thus the student's t distribution and estimators $\bar{y}$ and s are applied. As the goal is to predict upper bounds for future realizations, one-sided prediction intervals are regarded (Equation 1).

$$\left(-\infty,\ \bar{y} + t_{n-1;1-\alpha}\, s \sqrt{1+\frac{1}{n}}\right] \tag{1}$$

3.2. Loss Function derivation for Prediction Intervals

In order to compare the performance of multiple approaches for prediction intervals, a suitable performance measure is necessary. One approach is to apply a classification loss to the number of intervals that contain the realized value. However, a meaningless but arbitrarily broad coverage $(1-\alpha)$ can be chosen, necessitating a suitable trade-off between interval width and coverage. Thus, an inverse trade-off between length and coverage is a desirable trait of a loss function L for prediction intervals [3]. On the other hand, arbitrarily short miscalibrated intervals shall be avoided [4] and no specific knowledge about the underlying data generation process necessary [3].

The most commonly used interval-forecast loss function is the Winkler Loss $L_{winkler}(y, d, \lambda)$ (see Equation 2) [3], where smaller values indicate better prediction intervals. The interval length is represented by d, while d^l and d^u are the lower and upper bounds of the predicted interval. The first term in Equation 2 thus penalizes large intervals, and the second and third terms penalize values outside of the predicted interval. λ_l and λ_u are parameters balancing the trade-off between the interval width and coverage and should be set to $1/\lambda_l + 1/\lambda_u = \alpha$ [8].

$$L_{winkler}(y, d, \lambda) = |d| + \lambda_l (d^l - y)\mathbb{1}\{y < d^l\} + \lambda_u (y - d^u)\mathbb{1}\{y > d^u\} \tag{2}$$

As the data analytics approach aims at finding the best interval for a desired coverage, quantile-based one-sided prediction intervals, that require adaptations, are considered. As explained in the deviation of the Winkler Loss in the work of Gneiting and Raftery [8], the two-sided interval score is derived from the prediction of multiple quantiles at $r_1, ..., r_k$ (Equation 3).

$$L(r_1, ..., r_k, y, \alpha) = \sum_{i=1}^{k} [\alpha_i s_i(r_i) + (s_i(y) - s_i(r_i))\mathbb{1}\{y \le r_i\}] + h(y) \tag{3}$$

The two-sided interval score is a special case where d^l and d^u represent the lower and upper bounds corresponding to the $1 - \frac{\alpha}{2}$ quantile. Yet, in the reagarded one-sided case, only the upper limit d^u is considered. Furthermore, α_1 is set to $1-\alpha$ since the upper bound at the $1-\alpha$ coverage level is considered. The functions s and h must be at most polynomial in x and s shall be non-decreasing and, hence, set according to $s_1(x) = \frac{x}{\alpha}$ and $h(x) = -\frac{x}{\alpha}$, yielding to the one-sided loss function (Equation 4).

$$L_{one-sided}(d^u, y, \alpha) = d^u + \frac{1}{\alpha}(y - d^u)\mathbb{1}\{y > d^u\} \tag{4}$$

The loss function described in the literature evaluates based on only one data point [8]. In this setting, however, a large test data set is used. The loss is, thus, defined as the average over multiple data points. This approach further enables the comparison of more complex models with individual prediction intervals for different transitions. Thus, the final custom loss function in this paper is defined as shown in Equation 5:

$$L_{one-sided}(\bullet) = \frac{1}{n}\left(\sum_{i=1}^{n} d_i^u + \frac{1}{\alpha}\sum_{i=1}^{n}[(y_i - d_i^u)\mathbb{1}\{y_i > d_i^u\}]\right) \tag{5}$$

4. Prediction Model

All in all, the final prediction model combines a point estimator and a prediction interval in order to obtain a reasonable upper bound for the next transition time. In case the predicted upper bound does not exceed the time constraint, the corresponding dispatching action is permissible. Figure 1 shows an exemplary manifestation, where different prediction intervals and point estimators can be combined. Based on the newly introduced, adapted one-sided Winkler Loss, the best prediction interval can be selected, while point estimators can be ranked according to traditional metrics. In order to fulfill simple queue timelink areas controlling the transition time of lots is sufficient, i.e. as long as the transition time is lower than the time constraint, a dispatching action is permissible.

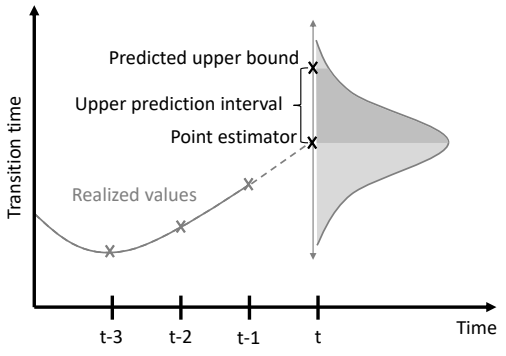

Fig. 1. Final model as a combination of point estimator and prediction interval

4.1. ARMA point estimator

Regarding real world data, dispatching in a semiconductor fab is characterized by high volume high variance [1], leading

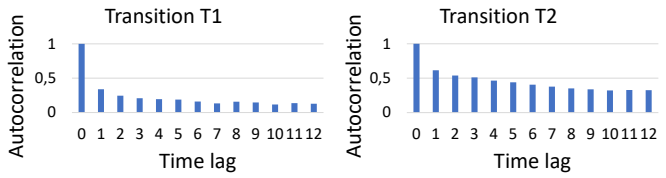

Fig. 2. Exemplary degree of auto-correlation regarding previous time-steps

to non-stationary transition times, when regarding transitions between two machines. Nevertheless, as visually observable in the autocorrelogram in Figure 2, the transition times $X_t, X_{t-1}, ...$ of many transitions exhibit high degrees of auto-correlation $corr(X_t, X_{t-j})$ towards their predecessors. The presence of auto-correlation can formally be tested with the Durbin-Watson Test, which is fulfilled for T2. Since the majority of transitions is auto-correlation wise visually comparable to T2, a random pick, and even in the lowest degree of auto-correlation transition T1 auto-correlation can be observed, it is assumed in the following and exploited for building a point estimator.

Thus, an Autoregressive Moving Average (ARMA) model, which predicts the current value X_t based on its predecessors according to Equation 6, where $\phi_p, \theta_q \neq 0$, $and\ \sigma_w^2 > 0, \epsilon_t \sim wn(0, \sigma_w^2)$ denote a white noise sequence, can be set up. A p-th order autoregressive process (AR) and a q-th order moving average (MA) process are mixed, so that c is an internal constant value, while $(\phi_1, \phi_2, ..., \phi_p)$ and $(\theta_1, \theta_2, ..., \theta_q)$ denote the AR and MR process parameters. Selecting p, q is crucial to enable parameter estimation through least square minimization.

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-i} \tag{6}$$

4.2. Prediction interval modeling

In order to find the most suitable prediction interval, three different approaches are compared applying the 75% coverage level and the previously derived Winkler Loss. The approaches are based on three different statistical distribution assumptions. The normal model assumes a normal distribution, while the logarithmic model first applies a logarithmic transformation on the data before deriving the quantiles. As a third model, the Chebyshev model is chosen, which derives a prediction interval without any distribution assumptions as described by Jørgensen and Sjoeberg [9] and thus represents the base case. Table 2 shows the calculated Winkler Loss scores which indicate that the logarithmic model performs best. This result can intuitively be explained by the fact that the distributions of realized transition times are right-skewed in most cases. Thus, a logarithmic transformation is necessary before assuming normal distribution and deriving the quantile for the prediction intervals , which is statistically analyzed in Section 4.3.

4.3. Logarithmic prediction interval

The application of the one-sided prediction interval to determine a lots scheduling feasibility under time constraints

Table 2. Calculation of the custom Winkler Loss for different statistical models

Model	Winkler Loss Score
Chebyshev model	6,855
Normal model	5,867
Logarithmic model	5,790

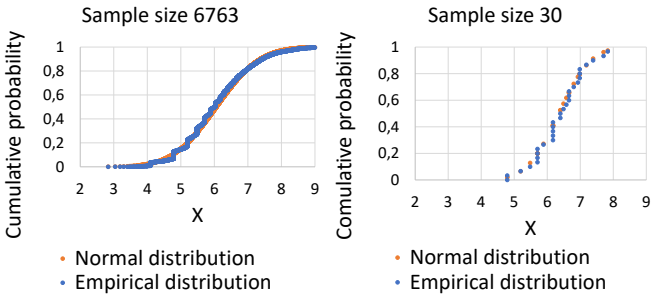

Fig. 3. Comparison of accumulated empirical distribution and normal distribution for transition T1

formally requires an underlying normal distribution, which is checked for the logarithmic model. The Kolmogorov-Smirnov test (K-S test) is a nonparametric goodness-of-fit test whether an observed distribution fits a predefined underlying distribution [6] , where the null hypothesis H_0 assumes the observed distribution to fit the predefined.

Transition *T1* is analyzed with a K-S test as it has the most available data points and visually appears to closely follow a normal distribution. A rejection of H_0 for *T1*, thus, most likely induces a rejection for any of the other transitions. Plotting the empirical cumulative distribution results in Figure 3, which suffices in visual comparison. Statistically, the cutoff for critical value for rejection with 5% significance lies approximated at 0.0473 and as the critical value $K_{n;1-\alpha/2} = 0.0165$ is smaller, the H_0 hypothesis assuming equal distributions is rejected at the 0.05 significance level, indicating that the observed distribution cannot be described as normally distributed.

Filion [7], however, mention, that for large sample sizes the critical value vanishes and as distributions are typically not perfectly symmetrical, H_0 is rejected very often. Hence, only regarding small subsets, for instance the first 30 recorded values as in Figure 3, results in not rejecting H_0 for the same significance level (i.e. $0.2417 > 0.1095$). Thus, despite the missing statistical significance, in the following it is assumed that in the relevant range the empirical distributions accurately enough resemble a normal distribution for the prediction interval application.

4.4. Performance measure

Application in the high volume semiconductor industry, however, requires as few time constraint violations as possible. Thus, a conversion from statistical intervals and probabilities to traditional performance measures that correctly predict individual constraint adherence is necessary. In order to determine the probability of the time constraint adherence and given the prediction interval and duration of the time constraint, it is pos-

sible to derive the adherence probability, as shown in Equation 7. The formula is initially used to derive the upper bound of the prediction interval d^u by using the point estimator $\hat{y}$, the t-value $t_{n-1;1-\alpha}$, the standard deviation s, and the sample size n. As shown in the second part of Equation 7, the formula can be transformed to determine $t_{n-1;1-\alpha}$. Instead of specifying the α value, the upper bound d^u is set equal to the duration of the time constraint, and the corresponding t-value is calculated. Given $t_{n-1;1-\alpha}$, the corresponding α can be derived. $1-\alpha$ then refers to the probability of the realized transition time to be smaller or equal to the time limit specified by the time constraint.

$$d^u = \hat{y} + t_{n-1;1-\alpha} s \sqrt{1+\frac{1}{n}} \Rightarrow t_{n-1;1-\alpha} = \frac{d^u - \hat{y}}{s\sqrt{1+\frac{1}{n}}} \tag{7}$$

Using the mean $\bar{x}$ and standard deviation s, the upper bound d^u for any given α can be calculated according to Equation 8. The higher the expected coverage, the higher the corresponding upper bound. One crucial aspect to understand when implementing the model is the trade-off between high coverage and a small upper bound. In general, an arbitrarily large upper bound can be chosen to achieve almost 100% coverage. This upper bound, however, does not provide any meaningful information, since it is arbitrarily large. Therefore, the goal is to find the lowest upper bound that still has an acceptable coverage level. This acceptable coverage level depends on the domain and must be specified by domain experts.

$$d^u = \bar{x} + t_{n-1;1-\alpha} s \sqrt{1+\frac{1}{n}} \tag{8}$$

5. Case Study

The proposed data analytics model is tested by implementing a three stage process where 64% of the data is used for training, 16% for validation, i.e. selecting the best performing hyperparameters, and the remaining 20% for testing, i.e. comparing models and analyzing their ability to generalize from test and validation data to the unknown testing. For the final model, point estimations are created for all transitions limited by time constraints. Based on multiple interviews with domain experts, a minimum of 75% coverage was identified and applied in the context of the case study.

The proposed model is applied to a complex semiconductor job shop that incorporated more than 20,000 unique paths between machines within the regarded timeframe. A four-digit number of transitions was found to be limited by time constraints. The results show an *accuracy* of 99.08%, a *recall* of 66.66%, and a *precision* of 15.38%. While the *accuracy* of the model is high, the *recall* and *precision* are relatively low. The overall explanation for this result is that the data is strongly biased. There are only few cases with a positive condition in comparison to thousands of cases with a negative condition, i.e. not violating the time-constraint. Even a model exclusively predicting negative conditions for all data points would score a high accuracy of close to 100%. A positive aspect of the final model, however, is that two third critical transitions could be recognized. Furthermore, a majority of the wrongly classified transitions were only slightly above the defined interval coverage. Changing the model in a benignly small way, such as by raising the required coverage to 76%, would, in this case, increase the recall significantly.

6. Discussion

This research provides a real-world validated approach to estimating time constraint adherence probabilities in a complex, matrix-shaped job shop. The approach is based on splitting the model into point and interval estimators that each can be uniquely modeled. The disadvantage of calibrating and combining two different models and model types is offset by the possibility to integrate a designated coverage that helps in reaching targeted statistical levels. The fact, that the current Production Control already controls operations very well tremendously complicates the application of traditional data analytics. Despite these adverse circumstances, the proposed approach outperforms the current predictions by two thirds.

Major drawbacks are the restriction to simple time constraints, i.e. time link areas that do not span more than two consecutive transactions, and the exclusive restriction to observed transition times, as further possible data sources, such as information about queues and failure behavior is not explicitly implemented. Thus, the main contribution lies in regarding a complex real-world system with few constraints and the simple interpretability of the proposed approach. Additionally, the simple implementation enables easily achievable improvements for complex job shops with time constraints.

7. Outlook

The proposed model addresses dispatching decision for *simple time constraints* in a real production plant. A more comprehensive review based on simulated data can help in building better models that are consequently transferred to validation in a real use-case. Thus, beyond regarding simulations, in principle two different, yet not mutually exclusive further approaches seem promising. First, the model van be improved by considering much more data, i.e. queue information, failure behavior and domain knowledge about implemented priority rules. Secondly, the scope can be extended towards integrating more complex time constraints and decisions.

The first can leverage larger data sets, that include longer observation periods to build more stable statistical model, the integration and generation of pre-processed and more complex features, e.g. queue length, failures, and the research on more comprehensive point estimators such as recurrent neural networks or deep learning models in general. Additionally, more complex statistical distributions can be regarded, that better fit to the observed data. The latter approach can focus on regarding multiple step time constraints, neighboring or even overlapping time constraints and hence, solve that practically more relevant complex time constraints in a comprehensive framework.

Furthermore, different industries can be considered and the regarded use case can be shifted towards a scheduling perspective, insofar as the optimal next machine is selected, decreasing the probability of violating time constraints. An intuitive approach to this problem lies in calculating the adherence probabilities based on the presented model and consequently selecting the most suitable machine. However, future research shall incorporate and regard the interrelation between active selection of actions and possible difficulties for the prediction models if applicable.

Acknowledgements

This research work was undertaken in the context of DIGIMAN4.0 project ("DIGItal MANufacturing Technologies for Zero-defect Industry 4.0 Production", http://www.digiman4-0.mek.dtu.dk/). DIGIMAN4.0 is a European Training Network supported by Horizon 2020, the EU Framework Programme for Research and Innovation (Project ID: 814225).

References

[1] Altenmüller, T., Stüker, T., Waschneck, B., Kuhnle, A., Lanza, G., 2020. Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints. Production Engineering 14, 319–328.

[2] Arima, S., Kobayashi, A., Wang, Y.F., Sakurai, K., Monma, Y., 2015. Optimization of re-entrant hybrid flows with multiple queue time constraints in batch processes of semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 28, 528–544.

[3] Askanazi, R., Diebold, F.X., Schorfheide, F., Shin, M., 2018. On the comparison of interval forecasts. Journal of Time Series Analysis 39, 953–965.

[4] Casella, G., Hwang, J.G., Robert, C., 1993. A paradox in decision-theoretic interval estimation. Statistica Sinica 3, 141–155.

[5] Chien, C.F., Kuo, C.J., Yu, C.M., 2020. Tool allocation to smooth work-in-process for cycle time reduction and an empirical study. Annals of Operations Research 290, 1009–1033.

[6] Dodge, Y., 2008. The concise encyclopedia of statistics. Springer Science & Business Media.

[7] Filion, G.J., 2015. The signed kolmogorov-smirnov test: why it should not be used. Gigascience 4, 13742–015.

[8] Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association 102, 359–378.

[9] Jørgensen, M., Sjoeberg, D.I.K., 2003. An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. Information and software Technology 45, 123–136.

[10] Jung, C., Pabst, D., Ham, M., Stehli, M., Rothe, M., 2014. An effective problem decomposition method for scheduling of diffusion processes based on mixed integer linear programming. IEEE Transactions on Semiconductor Manufacturing 27, 357–363.

[11] Kim, H., Lim, D.E., Lee, S., 2020. Deep learning-based dynamic scheduling for semiconductor manufacturing with high uncertainty of automated material handling system capability. IEEE Transactions on Semiconductor Manufacturing 33, 13–22.

[12] Kim, H.J., Lee, J.H., 2017. A branch and bound algorithm for three-machine flow shop with overlapping waiting time constraints. IFAC-PapersOnLine 50, 1101–1105.

[13] Kim, H.O., Park, S.H., Park, J.Y., Morrison, J.R., 2019. On the consequences of un-modeled dynamics to the optimality of schedules in clustered photolithography tools. 2019 Winter Simulation Conference (WSC) , 2224–2235.

[14] Lee, J.H., Li, J., 2017. Performance evaluation of bernoulli serial lines with waiting time constraints. IFAC-PapersOnLine 50, 1087–1092.

[15] Lee, Y.Y., Chen, C., Wu, C., 2005. Reaction chain of process queue time quality control. ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005. , 47–50.

[16] Lima, A., Borodin, V., Dauzère-Pérès, S., Vialletelle, P., 2017. Analyzing different dispatching policies for probability estimation in time constraint tunnels in semiconductor manufacturing. 2017 Winter Simulation Conference (WSC) , 3543–3554.

[17] Lima, A., Borodin, V., Dauzère-Pérès, S., Vialletelle, P., 2020. A sampling-based approach for managing lot release in time constraint tunnels in semiconductor manufacturing. International Journal of Production Research 59, 860–884.

[18] Maleck, C., Eckert, T., 2017. A comparison of control methods for production areas with time constraints and tool interruptions in semiconductor manufacturing. 2017 40th International Spring Seminar on Electronics Technology (ISSE) , 1–6.

[19] Maleck, C., Nieke, G., Bock, K., Pabst, D., Schulze, M., Stehli, M., 2019. A robust multi-stage scheduling approach for semiconductor manufacturing production areas with time contraints. 2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC) 1, 1–6.

[20] Mason, S.J., Kurz, M.E., Pfund, M.E., Fowler, J.W., Pohl, L.M., 2007. Multi-objective semiconductor manufacturing scheduling: A random keys implementation of nsga-ii. 2007 IEEE Symposium on Computational Intelligence in Scheduling , 159–164.

[21] McClellan, M., 1997. Applying manufacturing execution systems. CRC Press.

[22] Mönch, L., Fowler, J.W., Mason, S.J., 2012. Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems. volume 52. Springer Science & Business Media.

[23] Pan, C., Zhou, M., Qiao, Y., Wu, N., 2017. Scheduling cluster tools in semiconductor manufacturing: Recent advances and challenges. IEEE transactions on automation science and engineering 15, 586–601.

[24] Wang, H.K., Chien, C.F., Gen, M., 2015. An algorithm of multi-subpopulation parameters with hybrid estimation of distribution for semiconductor scheduling with constrained waiting time. IEEE Transactions on Semiconductor Manufacturing 28, 353–366.

[25] Wang, M., Srivathsan, S., Huang, E., Wu, K., 2018. Job dispatch control for production lines with overlapped time window constraints. IEEE Transactions on Semiconductor Manufacturing 31, 206–214.

[26] Wu, C.H., Cheng, Y.C., Tang, P.J., Yu, J.Y., 2012. Optimal batch process admission control in tandem queueing systems with queue time constraint considerations. Proceedings of 2012 Winter Simulation Conference , 1–6.

[27] Wu, C.H., Lin, J.T., Chien, W.C., 2010. Dynamic production control in a serial line with process queue time constraint. International Journal of Production Research 48, 3823–3843.

[28] Yang, D.L., Chern, M.S., 1995. A two-machine flowshop sequencing problem with limited waiting time constraints. Computers & Industrial Engineering 28, 63–70.

[29] Yang, F., Wu, N., Qiao, Y., Zhou, M., 2016. Efficient and optimal scheduling of time-constrained hybrid multi-cluster tools in semiconductor industry. 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC) , 1–6.