Multi-fidelity Simulation-Based Optimisation for Large-Scale Production Release Planning in Wafer Fabs

Zhengmin Zhang, Zailin Guan, Yeming Gong, Qian Shen

Multi-Fidelity Simulation-Based Optimisation for Large-Scale Production Release Planning in Wafer Fabs

Zhengmin Zhang[1[0000-0002-5229-9142]], Zailin Guan[1[0000-0003-2295-6698]], Yeming Gong[2[0000-0001-9270-5507]], and Qian Shen[1]

[1] Huazhong University of Science and Technology, Wuhan, China
hust_zzm@hust.edu.cn
[2] EMLYON Business School, 23 avenue Guy de Collongue, 69134 Ecully cedex, France
gong@em-lyon.com

Abstract. This paper focuses on large-scale production release planning problems in wafer fabs. Due to the complexity and dynamic characteristic of production lines, it is useful to develop simulation models to evaluate different production plans and select the optimal one. However, detailed simulation requires large computational costs, especially for large-scale problems. Therefore, we provide a multi-fidelity optimisation with ordinal transformation and optimal sampling (MO2TOS) based on the multiple population evolutionary algorithm to accelerate computational efficiency and address large-scale release planning problems for wafer fabs. In the low-fidelity approximation process of the proposed approach, we employ open queuing theory to estimate cycle times. The experimental results confirm that the low-fidelity model shows a high consistency with the high-fidelity model and the proposed multi-fidelity optimisation method is effective at solving large-scale release planning problems.

Keywords: Queuing Theory, Multi-fidelity Simulation-based Optimisation, Wafer Fabs, Production Release.

1 Introduction

The manufacturing of wafer fabs is a highly complex and time-consuming process because of massive process steps and complex process requirements [1]. Thus, it is difficult to provide an effective production release plan for the production line. In wafer fabs, the production release planning problem seeks an optimal or near-optimal solution to optimise the output and the work-in-process of a production system by determining the release rate of each product type.

Some basic approaches, such as mathematical programming [2-3], simulation-based optimisation [4-5], and analytical modelling [1,6], have been developed to address production release planning problems. Simulation-based optimisation [7-8] refers to the approaches that search for the optimal solution by using simulation models to directly evaluate the objective values of candidate solutions. Compared with mathematical modelling, this technology can provide accurate estimates of given production plans.

Besides, simulation-based optimisation plays an important role in digital twins and has huge advantages in the application of the digital twin platform [9].

However, simulation-based optimisation consumes high computational costs because of its high-precision modelling capability [10]. Real-world production planning problems require considerable computational time to run the entire solution space by the simulation model [11]. Therefore, simulation models are often used to verify the solution quality of mathematical models [12-14]. Some approaches [15-17] address the order release planning problems by iterative simulation and linear programming. Nevertheless, these methods may take a long time to converge if the initial release plan is ineffective or the convergence process is not stable. Therefore, the research question of this paper is: *How to solve large-scale production release planning problems in wafer fabs effectively by simulation-based optimisation?*

This research therefore designs a multi-fidelity optimisation with ordinal transformation and optimal sampling (MO2TOS) based on the multiple population evolutionary algorithm for production release planning problems. The proposed method is developed based on the MO2TOS proposed by Xu et al. [18]. The high-fidelity model is developed using the discrete-event simulation and the low-fidelity model is established as a mathematical expression based on the queuing theory. In the proposed MO2TOS, we provide a multiple population evolutionary algorithm to accelerate the solution space searching and obtain the approximate solution space. Compared with other methods, this proposed method obtains the same optimal solution as MO2TOS when solving large-scale problems, and saves about 90% of computational time.

2 Problem Formulation and Modelling

Consider the following scenario: k ($k = 1 \ldots K$) types of products are produced by m ($m = 1 \ldots M$) workstations of a wafer production line that demand exceeds supply. Product k have r ($r = 1 \ldots R_k$) alternative processing routes, each route has $L_{k,r}$ operations. $M_{k,r,l}$ represents the workstation for the lth operation of product k with processing route r. Products of different types have similar processing flows but not exactly the same. Any processing routes may have re-entrant flows. The demand d_k of product k is known, the optimal release rate and the routing allocation scheme for each product type need to be calculated to increase productivity. We need to find the optimal release plan for different products and obtain the optimal routing allocation plan for each product type with different processing routes. The objective of the problem is to complete the customer's requirements as soon as possible and minimise the maximum work-in-process of workstations in the production line. To maintain balanced production, we put forward the following two assumptions:

1. The release rate of each product in a planning period is stable and changeless.
2. Under any release plan, the product demand capacity cannot exceed the production capacity of the machine. Otherwise, the system cannot reach a steady state and may lead to infinite accumulation of work-in-process.

We provide a mathematical model to describe the proposed problem on the basis of the above assumptions. The objective function are as follows:

$$Min(w1 \cdot makespan + w2 \cdot work\_in\_process) \quad (1)$$

$$makespan = max\, F\left(k, r, p, L_{k,r}\right), \forall k, r, p, m \quad (2)$$

$$work\_in\_process = max\left(\frac{\sum_{p=1}^{\lceil F(k,r,p,l)\rceil} \sum_{k=1}^{K} \sum_{r=1}^{R_k} WIP(k,r,p,m)}{\lceil F(k,r,p,l)\rceil}\right), \forall m \quad (3)$$

Subject to:

$$\sum_{p=1}^{\lceil F(k,r,p,l)\rceil} \sum_{r=1}^{R_k} y_{k,r}^{p} \geq d_k, \forall k, m \quad (4)$$

$$x_{k,r}^{p}, y_{k,r}^{p}, F(k,r,p,l), WIP(k,r,p,m) \geq 0, \forall k, r, p, m \quad (5)$$

Equations (1)-(3) represents the objectives of this problem, $w1$ and $w2$ means the weights of two objectives. $F(k,r,p,l)$ means the complete time for the lth operation of product k with processing route r released in period p. $WIP(k,r,p,m)$ is the cumulative work-in-process for product k with processing route r at workstation m in period p, where $WIP(k,p,m) = \sum_{r=1}^{R_k} WIP(k,r,p,m)$. Constraint (4) provides the minimum production quantity for each product. $y_{k,r}^{p}$ represents the external release material quantity for product k with processing route r in period p. $y_{k,r}^{p}$, $F(k,r,p,l)$ and $WIP(k,p,m)$ are decision variables linked with $x_{k,r}^{p}$ (external release material quantity for product k with processing route r in period p). In the proposed mathematical model, we need to derive the relationship between $x_{k,r}^{p}$ and other decision variables.

3 MO2TOS Based on the Multiple Population Genetic Algorithm

3.1 Low-Fidelity Approximation Based on Queuing Theory

We propose a queuing theory-based low-fidelity approximate method to derive the relationship between $x_{k,r}^{p}$ and other variables. According to $x_{k,r}^{p}$, we can approximate the product arrival distribution and calculate the waiting time and the queue length according to queuing theory. Then, $F(k,r,p,l)$ and $WIP(k,p,m)$ can be calculated accordingly, and $y_{k,r}^{p}$ is obtained according to $x_{k,r}^{p}$ and $F\left(k,r,p,L_{k,r}\right)$. Here is a list of the notations we use:

τ_m	The processing time for m in processing any operations
ρ_m^p	The utilisation of workstation m
$P_{k,r,l}$	The processing time of the lth operation of product k with route r
$W(k,r,p,l)$	The waiting time for the lth operation of product k with route r released in period p
λ_m^p	The expected arrival rate for all products at workstation m in period p
$\lambda_m^p(k,r)$	The expected arrival rate for product k at workstation m in period p
$\lambda_{0,m}^p(k,r)$	The external arrival rate for product k at workstation m in period p

$\lambda_{n,m}^{p}(k,r)$ The arrival rate for product *k* from workstation *n* to workstation *m* in period *p*

$\lambda_{m}^{p}(k,r)$ comprises the external arrival flows and the re-entrant flows. Thus, $\lambda_{m}^{p}(k,r)$ can be calculated as:

$$\lambda_{0,m}^{p}(k,r) = x_{k,r}^{p} \cdot [M_{k,r,1} = m] \tag{6}$$

$$\lambda_{n,m}^{p}(k,r) = \sum_{n=1}^{M}(\sum_{l=1}^{L_{k,r}-1}(x_{k,r}^{p} \cdot [M_{k,r,l} = n, M_{k,r,l+1} = m])), n \neq m \tag{7}$$

Therefore, $\lambda_{m}^{p}(k,r)$ can be described as:

$$\lambda_{m}^{p}(k,r) = \lambda_{0,m}^{p}(k,r) + \lambda_{n,m}^{p}(k,r), \forall m,k,r,p \tag{8}$$

λ_{m}^{p} is then given by:

$$\lambda_{m}^{p} = \sum_{k=1}^{K}\sum_{r=1}^{R_k}\lambda_{m}^{p}(k,r), \forall m,p \tag{9}$$

Therefore, the utilisation of workstation *m* can be provided as:

$$\rho_{m}^{p} = \tau_{m} \cdot \lambda_{m}^{p}, \forall m,p \tag{10}$$

We refer the approximation method of a GI/G/1 system in Whitt [19] to estimate the expected waiting time of each product lot in any workstations.

$$EW = \frac{\tau\rho(C_a^2+C_s^2)g}{2(1-\rho)} \tag{11}$$

Where g is defined as:

$$g(\rho, C_a^2, C_s^2) = \begin{cases} \exp\left[-\frac{2(1-\rho)}{3\rho}\frac{(1-C_a^2)^2}{C_a^2+C_s^2}\right], C_a^2 < 1 \\ 1 \qquad\qquad\qquad\qquad , C_a^2 \geq 1 \end{cases} \tag{12}$$

τ denotes the service time for a product, ρ means the utilisation of this server, C_a^2 represents the squared coefficient of variation of the arrival interval distribution at this server, while C_s^2 is the squared coefficient of variation for the service-time distribution. C_a^2 is the variance of the renewal interval divided by the square of its mean, and C_s^2 is the variance of the service time divided by the square of its mean.

In our low-fidelity approximation model, $W(k,r,p,l)$ can be represented by the expected waiting time EW. Accordingly, $F(k,r,p,l)$ can be recorded as:

$$F(k,r,p,l) = P_{k,r,l} + W(k,r,p,l) \tag{13}$$

According to the Little's law:

$$WIP(k,p,j) = W(k,r,p,l) \cdot \lambda_{m}^{p} \tag{14}$$

$F\left(k,r,p,L_{k,r}\right)$ means the makespan of product *k* released in period *p*. Because the release rate in each period is the same, we can consider $F\left(k,r,1,L_{k,r}\right)$ as the cycle time

for product k with processing route r. Thus, $y_{k,r}^{p}$ can be estimated according to the following formula:

$$y_{k,r}^{p} = x_{k,r}^{p} + F\left(k, r, 1, L_{k,r}\right) \tag{15}$$

3.2 Framework of MO2TOS Based on the Multiple Population Genetic Algorithm

The steps of MO2TOS based on the multiple population genetic algorithm (MPGA) are exhibited in **Fig. 1**. Traditional MO2TOS uses the low-fidelity model to estimate all the results of the solution space. Then, it selects high-quality solutions by the ordinal transformation and optimal sampling strategy, and uses the high-fidelity simulation model to calculate these solutions and choose the optimal solution. However, the solution space of large-scale production planning problems may be more than several million. Even if we use the low-fidelity model to run the entire solution space, the computational costs and runtime are not negligible. To reduce computational burden, we design a multiple population evolutionary algorithm to accelerate the solution space search efficiency based on the original MO2TOS. We choose a multiple population evolutionary algorithm because solution spaces of production problems are generally multi-peak, and multiple population evolutionary algorithms are conducive to jumping out of local optimal and accelerating the convergence process.

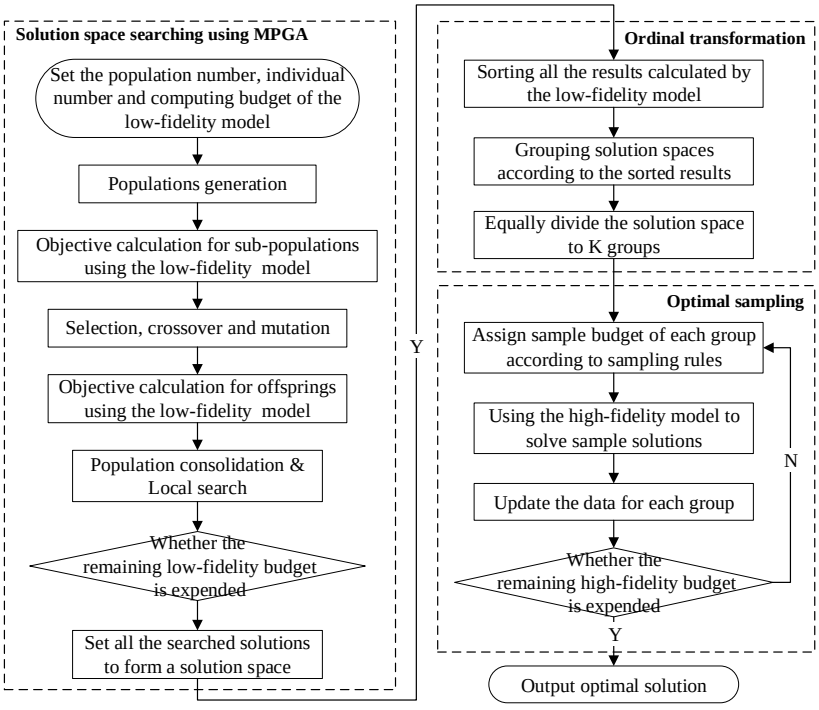

Fig. 1. MO2TOS based on the multiple population genetic algorithm

4 Computational Experiments

We simplify the actual wafer production system and develop some typical wafer production lines to test and estimate the effectiveness of the proposed method. The simplified cases retain the characteristics of batch processing, machine failures and mixed-flow production in actual wafer fabs. First, we develop a small-scale production release planning case to verify the accuracy of the low-fidelity model. We calculate the solution space via the high-fidelity simulation model and the low-fidelity model. **Fig. 2** plots the output of the solution space after ordinal transformation. According to **Fig. 2**, we find the consistency of the output of the solution space using two models is obvious, which proves the proposed low-fidelity approximation method is feasible. Therefore, the low-fidelity model we proposed in this paper is feasible.

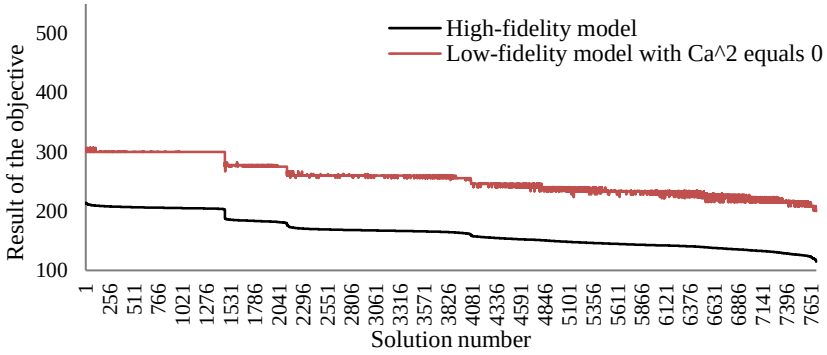

Fig. 2. Output of the solution space after ordinal transformation

We further develop a large-scale production release planning case and employ the proposed method to solve the case. We first examine the optimisation performance of the proposed multi-population genetic algorithm. **Fig. 3** shows the convergence curves of the optimal results for the same case when the same low-fidelity budget (10000) is allocated. From **Fig. 3**, we conclude that the proposed algorithm has stable optimisation capability in solution space searching process.

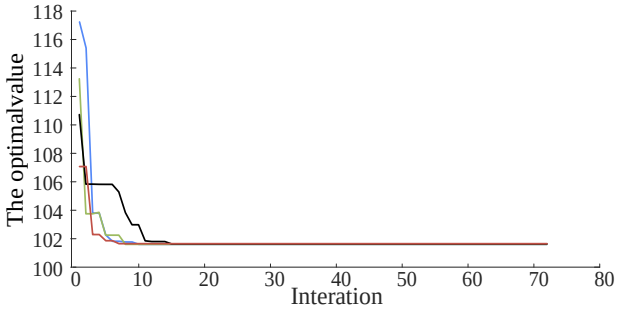

Fig. 3. The convergence curve of the optimal results

We compare the proposed method with the high-fidelity simulation optimisation (HFSO) method and MO2TOS used by Zhang et al. [4]. HFSO selects specific amount of solutions randomly according to the high-fidelity simulation budget, uses only the high-fidelity model to run all the selected solutions, and searches for the optimal solution according to the results. The MO2TOS method solves the entire solution space by using a low-fidelity model, selects high-fidelity samples using ordinal transformation and optimal sampling, and runs the high-fidelity samples to find the optimal solution. The MO2TOS based on the multi-population genetic algorithm differs from MO2TOS in the first phase. It uses the multi-population genetic algorithm combined with a low-fidelity model to solve the solution space.

Table 1. The optimal results for the large-scale release planning case.

High-fidelity budget	The optimal result			Runtime (s)		
	HFSO	MO^{2}TOS	MO2TOS & MPGA	HFSO	MO^{2}TOS	MO^{2}TOS & MPGA
200	128.048	113.475	***113.292***	425	2221	***358***
400	120.651	113.447	**113.292**	1351	2874	**824**
600	121.166	***113.292***	**113.292**	3028	***3944***	**1672**
800	124.113	113.292	**113.292**	4908	5503	**2785**
1000	120.061	113.292	**113.292**	7948	7569	**4906**
1200	117.943	113.292	**113.292**	10244	10348	**7149**

Table 1 provides the optimal results with different methods for the large-scale release planning case. We can conclude: 1. The more fidelity budgets are allocated, the longer the run time is required. 2. Regardless of the amount of high-fidelity budget allocated, HFSO is worse than the compared methods. **3. The method proposed in this paper can find the optimal solution when the high-fidelity budget is only 200, and save about 90% of the computing time than MO2TOS when solving large-scale problems.**

The studied actual wafer production system uses a static lead time release planning model in Manufacturing Requirements Planning (MRP) to develop release plans at present. This model is oversimplified because it considers lead times as workload-independent constants. The static lead time is usually estimated from historical data and experience. This method releases the materials corresponding to the demand of each cycle directly according to the static lead time, which often causes a large amount of work-in-process accumulation. Consider the proposed case as an example, the optimal result of this method is 134.515, which is 18.733% worse than the method proposed in this manuscript.

This method is suitable for real-world wafer fabs that demand exceeds supply. First, managers need to know the demands of multiple planning periods in the future and roughly estimate the upper and lower limits of release rates based on demands. Then, managers can use the proposed low-fidelity estimation method to estimate the revenues of different release plans and obtain the appropriate solution set. Next, they can use the high-fidelity discrete simulation model to run these plans and select the optimal one.

This method can greatly reduce computation time and is suitable for large-scale problems. Note that due to some uncertain events or parameters in actual production scenarios, this method may need to update release plans periodically to keep the simulation model consistent with the actual production system.

5 Conclusions

In this research, we present a MO2TOS based on the multi-population evolutionary algorithm to address large-scale production release planning problems in wafer fabs. The low-fidelity estimation model is an open queue approximation model and the high-fidelity simulation model is established based on discrete-event simulation. We test a large-scale case to evaluate the proposed method. The proposed method is compared with the MO2TOS framework proposed by Zhang et al. [4] and the high-fidelity simulation optimisation method. Experiment results show that the proposed method can obtain the same optimal solution as Zhang et al. [4] and save computing time.

In the future, we will consider how to improve the estimation accuracy and speed of the low-fidelity model simultaneously. Moreover, we will study how to let high-fidelity model estimation results feedback to the low-fidelity model and reduce estimation gaps between the low- and the high-fidelity models.

References

1. Chung, S. H., Lai, C.: Job releasing and throughput planning for wafer fabrication under demand fluctuating make-to-stock environment. The International Journal of Advanced Manufacturing Technology 31(3), 316-327 (2006).
2. Asmundsson, J., Rardin, R. L., Uzsoy, R.: Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Transactions on Semiconductor Manufacturing 19 (1), 95–111 (2006).
3. Leachman, R. C.: Semiconductor production planning. In handbook of Applied Optimisation, Eds. New York, NY, USA: Oxford Univ. 746–762 (2001).
4. Zhang, F., Song, J., Dai, Y., Xu, J.: Semiconductor wafer fabrication production planning using multi-fidelity simulation-based optimisation. International Journal of Production Research 58(21), 6585-6600 (2020).
5. Thuerer, M., Stevenson, M., Land, M., Fredendall, L.: On the combined effect of due date setting, order release, and output control: an assessment by simulation. International Journal of Production Research 57(6), 1741-1755 (2019).
6. Schneckenreither, M., Haeussler, S., Gerhold, C.: Order Release Planning with Predictive Lead Times: A Machine Learning Approach. International Journal of Production Research (2020).
7. Chen, C., Lee, L.: Stochastic Simulation Optimisation (An Optimal Computing Budget Allocation). Back Matter 175-227 (2010).
8. Xu, J., Nelson, B., Hong, L.: Industrial strength COMPASS: A comprehensive algorithm and software for optimisation via simulation. Acm Transactions on Modeling & Computer Simulation 20(1), 1-29 (2010).

9. Zhang, Z., Guan, Z., Gong, Y., Luo, D., Yue, L.: Improved multi-fidelity simulation-based optimisation: application in a digital twin shop floor. International Journal of Production Research (2020).
10. Asmundsson, J., Rardin, R. L., Turkseven, C. H., Uzsoy, R.: Production planning with resources subject to congestion. Naval Research Logistics 56 (2), 142–157 (2009).
11. Fowler, J., Mönch, L.: Modeling and Analysis of Semiconductor Manufacturing. Advances in Modeling and Simulation. Springer, Cham (2017).
12. Bang, J., Kim, Y.: Hierarchical Production Planning for Semiconductor Wafer Fabrication Based on Linear Programming and Discrete-Event Simulation. IEEE Transactions on Automation Science and Engineering 7(2), 326-336 (2010).
13. Kopp, D., Monch, L., Pabst, D., Stehli, M.: Qualification Management in Wafer Fabs: Optimisation Approach and Simulation-Based Performance Assessment. IEEE Transactions on Automation Science and Engineering 17(1), 475-489 (2019).
14. Ziarnetzky, T., Kacar, N., Monch, L., Uzsoy, R.: Simulation-based performance assessment of production planning formulations for semiconductor wafer fabrication. Winter Simulation Conference. IEEE, Huntington Beach (2015).
15. Missbauer, H.: Order release planning by iterative simulation and linear programming: theoretical foundation and analysis of its shortcomings. European Journal of Operational Research 280(2), 495-507 (2020).
16. Kim, S. H., Lee, Y. H.: Synchronized production planning and scheduling in semiconductor fabrication. Computers & Industrial Engineering 96(6), 72-85 (2016).
17. Kim, B., Kim, S.: Extended model for a hybrid production planning approach. International Journal of Production Economics 73(1), 165-173 (2001).
18. Xu, J., Zhang, S., Huang, E., Chen, C., Lee, L., Celik, N.: MO2TOS: Multi-Fidelity Optimisation with Ordinal Transformation and Optimal Sampling. Asia-Pacific Journal of Operational Research 33 (3), 1650017 (2016).
19. Whitt, W.: The Queueing Network Analyser. Bell Labs Technical Journal 62(9), 2779 - 2815 (1983).