Push and Time at Operation strategies for cycle time minimization in global fab scheduling for semiconductor manufacturing

Felicien Barhebwa-Mushamuka, S. Dauzere-Peres, C. Yugma

Push and Time at Operation strategies for cycle time minimization in global fab scheduling for semiconductor manufacturing

F. Barhebwa-Mushamuka [1], S. Dauzère-Pérès [2], C. Yugma [2]

Abstract— **This paper investigates two global scheduling strategies for cycle time minimization in semiconductor manufacturing. These global scheduling strategies represented as a linear programming models are compared to a First-in-First out dispatching rule. The first global scheduling strategy is a Push strategy, in which products are pushed to their final operations using high Work-In-Process holding costs on the first operations. The second global scheduling strategy is a Time at Operation strategy, where Work-In-Process quantities that have arrived at different times in an operation are penalized differently. The computational results performed on industrial data using the Anylogic simulation software coupled with IBM ILOG CPLEX show that the Time at Operation strategy minimizes the cycle time while maintaining a high throughput compared to the Push strategy and the simple First-In-First-Out dispatching rule. The paper also shows, when production targets are determined using the Push strategy, products with a large number of operations are prioritized.**

I. INTRODUCTION

We are surrounded by electronic devices, whether at home or at work, such as telephone, televisions, computers, advanced medical diagnostic equipment and other high-tech devices. The companies that produce chips (which are part of the composition of the devices) are semiconductor industries. These industries are constantly looking for efficient production strategies in order to be and remain competitive.

The process of semiconductor fabrication is probably the most complex manufacturing processes in existence [4]. In addition to common characteristics that can be found in classical manufacturing contexts, the semiconductor process includes characteristics that make such a production complex as re-entrant flows induced mainly by scarce and expensive resources, hundreds of operations for each product leading to very long cycle times, different types of scheduling problems, etc.Experiments in this paper are based on a High Mix Low volume (HMLV) production system.

The process of manufacturing Integrated Circuits can be summarized in two main parts. The first part, semiconductor wafer fabrication (wafer fab) or front-end, corresponds to the long and complex process of manufacturing silicon chips on silicon wafers. The second part, back-end, corresponds to the cutting and packaging of the chips and the final tests. In a wafer fab, different products require hundreds of operations, with re-entrant flows, performed on hundreds of machines of different types that are grouped in workcenters [4] and [5]. Each workcenter includes specific process characteristics, which increase the complexity of scheduling decisions such as batch process, a parallel process, auxiliary resources, etc. Hence, determining detailed scheduling decisions for the entire facility is very difficult in semiconductor manufacturing. Cycle time is one of the main Key Performance Indicators in semiconductor manufacturing, as it is a lever that decision makers can use to be competitive.

Cycle time covers the life of a product in a factory, combining value-added and non-value-added processes. Many parameters influence cycle times in semiconductor manufacturing some key factors are provided in [1] such as equipment availability, utilization, product mix, variability, hot lots, re-entrant flows, etc. The minimization of cycle times has an impact on several other metrics and key performance indicators such as throughput, yield, on-time delivery, etc. Short cycle times also help to reduce wafer risk contamination, yield loss and the inventory that should be maintained [2]. This paper studies two global scheduling strategies for cycle time minimization. A Push strategy, in which products are pushed to their final operations using high Work-In-Process holding costs on the first operations and a Time at Operation strategy, where Work-In-Process quantities that have arrived at different times in an operation are penalized differently. The goal is to prioritize the processing of Work-In-Process quantities that have spent more time in the operation. Since products share the same resources at multiples stages of their processes in semiconductor manufacturing, regulating the competition between products on the different shared resources is critical. In this work, this is well managed by the Time at Operation strategy, which ensures that the waiting time of products in each operation is not preventing the minimization of cycle times.

In scheduling problems, most of the criteria are derived from the completion times of products, which constitute the main information to compute the cycle times of products, see e.g., [9]. Cycle time reduction refers to the strategy of decreasing the time a product spends in the factory from its release to its last operation. Shorter cycle times drive a better on time delivery, help to decrease Work-In-Process and ensure good production quality (higher yield). Several strategies have been studied, essentially based on the management of factors that influence the cycle time. Variability is considered as one of the cycle time killers,

This work was supported by the European Unions Horizon 2020 program under grant agreement No 737459 (project Productive 4.0).

[1] IMT Atlantique, Department of Automation, Production and Computer Sciences, Nantes, France

[2] Mines Saint-Etienne, Univ. Clermont Auvergne CNRS, UMR 6158 LIMOS CMP, Departement of Manufacturing Sciences and Logistics F-13541 Gardanne, France

felicien.barhebwa@emse.fr
dauzere-peres@emse.fr
Yugma@emse.fr

[1]. [10] provides a three-step procedure for cycle time reduction: (1) Identification of controllable factors that influence the product cycle time, (2) Investigation of the relationship between the controllable factors and product cycle time and (3) Finally, based on this relationship, actions should be planned to shorten the product cycle time. In [12], cycle time reduction is done by using a hierarchical approach based on two schedulers. A mid-term scheduler that maximizes the weighted production flow and ensures on time delivery as well and a short-term schedule, which slices mid-term scheduling results into more detailed schedules. To reduce cycle time, [13] addresses a planning problem, which determine how many lots have to be released during the next planning period and which target cycle times have to be assigned to each lot (including both new releases and the initial WIP) such that both cycle time and the deviation of fab output from the master production schedule are minimized.

Other factors that influence the cycle time have been used as a lever for cycle time reduction such as Work-In-process management [11], batch size [14], lot size [15] and [16], queue time management and priority management. Equipment management, essentially the study of preventive maintenance segregation, is proposed in [17] with the goal to determine the optimum preventive maintenance policy that results in reduced fabrication cycle times. [18] provide a set of methodologies and scheduling applications for managing the cycle time in semiconductor manufacturing called SLIM (Short cycle time and Low Inventory Manufacturing).

The minimization of the mean, variance and standard deviation of the cycle time is also widely studied. Scheduling policies are one of the levers used for mean and variance cycle time reduction in semiconductor manufacturing [2] and [19]. For more information about the minimization of mean and variance of cycle times, see [3], [20] and [21].

To our knowledge, the Push and Time at Operation strategies are not yet studied especially using a global scheduling approach. In these paper, these strategies are implemented through mathematical programming models (global scheduling models), see for instance [6] and [7] for other strategies than the one in this paper. These global scheduling strategies determine production targets, i.e., product quantities to complete in each operation and each period on a scheduling horizon. The Push strategy characteristics are outlined while the performance of the Time at Operation strategy in terms of cycle time and throughput is shown in computational results. Both strategies are evaluated in a rolling horizon scheme using the generic multi-method simulation model proposed in [8].

The paper is structured as follows. Section II describes the Push and Time at Operation strategies. Section III presents and analyzes computational results on industrial data. Finally, conclusions are provided in Section IV.

II. Push Strategy versus Time at Operation Strategy

To minimize cycle times, the push strategy and the time at operation strategy are oriented towards the management of the Work-In-Process at the operations in the routes of each product. These strategies are described below.

- The Push strategy, consists of setting costs in decreasing order from the first operation to the last operation allowing products to advance as quickly as possible towards their last operations. Assume that UB is the maximum number of operations in the product mix. UB is decreased forward on the set of operations of each product, to ensure that products are pushed forward toward their last operations. Figure 1 illustrates the Push strategy.

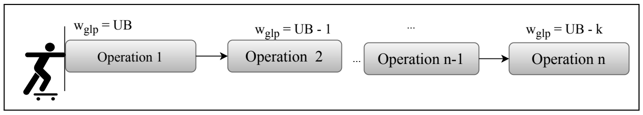

Fig. 1: Push strategy

Table I summarizes the parameters and decision variables used in the global scheduling models for the push strategy and the time at operation strategy. In our experiment P = H = 99 (scheduling horizon) as shown in section III-A.

TABLE I: Notations

Parameters:	
$\mathcal{G}$	Set of products,
$\mathcal{K}$	Set of work-centers,
$\mathcal{L}_g$	Set of operations of product g,
$\mathcal{LK}(k)$	Set of operations and products that must be processed in work-center k, i.e $(g, l) \in \mathcal{L}K(k)$ means that operation l of product g must be processed in work-center k,
P	Number of periods in planning horizon,
IW_{gl}	Initial WIP at operation l of product g,
R_{gp}	Release quantity of product g in period p,
α_{gl}	Unit process time at operation l of product g,
C_{kp}	Capacity of work-center k in period p,
w_{glp}	Unit WIP holding cost at operation l of product g in period p.
Decision variables:	
X_{glp}	Quantity of product g arriving in operation l in period p,
Y_{glp}	Quantity of product g completing operation l in period p,
Z_{glp}	WIP of product g at operation l at the end of period p,
Z_{glpt}	WIP of product g at operation l at the end of period p that arrived in period t ($t \leq p$ and $\sum_{t=1}^{p} Z_{glpt}$ = Z_{glp}).

Below, the Linear Program that models the global scheduling Push strategy is written.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{P} w_{glp} Z_{glp} \quad (1)$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \quad (2)$$

$$Z_{g11} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathcal{G} \quad (3)$$

$$Z_{gl1} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \quad (4)$$

$$Z_{g1p} = Z_{g1(p-1)} + R_{gp} - Y_{g1p} \quad \forall g \in \mathcal{G},\ p = 2,\ \ldots,\ P \quad (5)$$

$$Z_{glp} = Z_{gl(p-1)}+X_{glp}-Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \geq 2,\ p = 2,...,\ P \tag{6}$$

$$\sum_{(g,l)\in \mathcal{LK}(k)} \alpha_{gl} Y_{glp} \leq C_{kp} \quad \forall k \in \mathcal{K},\ p = 1,\ ...,\ P \tag{7}$$

$$Z_{glp},\ Y_{glp},\ X_{glp} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ ...,\ P \tag{8}$$

The objective function (1) ensures that the Work-In-Process is moving forward to the last operations of products. The costs w_{glp} are chosen in such a way that $w_{glp} \leq w_{gl-1p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ ...,\ P$. Constraints (2) tie consecutive operations. Constraints (3)-(6) are flow constraints linking the Work-In-Process of each product at each operation in each period with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (7) are resource capacity constraints.

In the Push strategy, when production targets are determined, products with a large number of operations are prioritized. This is because the larger the number of operations, the higher the cost for holding Work-In-Process. This can be seen as a downside because often decision makers do not seek to prioritize products with a large number of operations. This is confirmed in the computational results of section III. Let us illustrate this fact with an example with two products, product P_1 with 5 operations and product P_2 with 3 operations. Assume that product P_1 in operations 1 and 2 shares the same resource R with product P_2 in all its operations. Resource R has a limited capacity. The unit holding cost for the Work-In-Process is given in Figure 2.

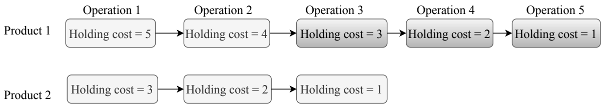

Fig. 2: Push strategy drawback illustration.

If there is not enough products P_1 to fill the capacity of resource R, then both products P_1 and P_2 will be processed. As product P_1 has a larger priority based on its unit holding cost, if there is enough product P_1 to fill the capacity of resource R, then product P_2 will not be produced. The time at operation strategy is designed to overcome this drawback.

The Time at Operation strategy ensures that the Work-In-Process of product g arriving in period p at operation l and which remains at the end of period p, does not have the same holding cost β as the Work-In-Process that arrived at period $t < p$ at operation l. Figure 3 illustrates the time at operation strategy.

Numerical results in Section III show that the time at operation strategy reduces the average cycle time compared to the Push strategy by 15% and increases the overall throughput of the factory by 8%. The Linear Program that models the Time at Operation global scheduling strategy is written bellow:

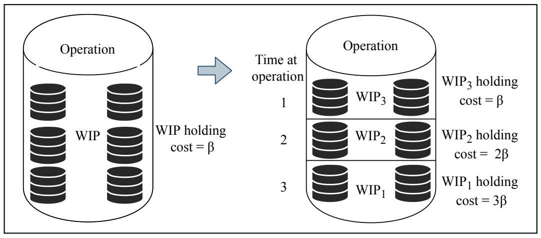

Fig. 3: Time at operation strategy

$$\mathcal{M}in \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \sum_{t=1}^{p} [(p+1)-t] Z_{glpt} \tag{9}$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \tag{10}$$

$$Z_{g111} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathcal{G} \tag{11}$$

$$Z_{gl11} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \tag{12}$$

$$\sum_{t=1}^{p} Z_{g1pt} \geq \sum_{t=1}^{p-1} Z_{g1(p-1)t} + R_{gp} - Y_{g1p} \quad \forall g \in \mathcal{G},\ p = 2,\ ...,\ H \tag{13}$$

$$\sum_{t=1}^{p} Z_{glpt} \geq \sum_{t=1}^{p-1} Z_{gl(p-1)t} + X_{glp} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2,\ ...,\ H \tag{14}$$

$$\sum_{t=1}^{m} Z_{glpt} \geq \sum_{t=1}^{m} Z_{gl(p-1)t} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2,\ ...,\ H,\ m \leq p-1 \tag{15}$$

$$\sum_{(g,l)\in\mathcal{LK}(k)} \alpha_{gl} Y_{glp} \leq C_{kp} \quad \forall k \in \mathcal{K},\ p = 1,\ ...,\ H \tag{16}$$

$$Z_{glpt},\ Y_{glp},\ X_{glp} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ ...,\ H,\ t \leq p \tag{17}$$

The objective function (9) ensures that the Work-In-Process is pushed forward to the last operations by taking into account how long a product has been waiting in an operation. The holding cost of the Work-In-Process is increasing with the number of periods it remains in an operation. The objective function ensures that the Work-In-Process arriving in period p in an operation and still at the operation at the end of p does not have the same holding cost as the Work-In-Process which arrived in period $t < p$ in the operation. The Constraints (10) tie consecutive operations. Constraints (11) model the first operation in the first period upon which the initial Work-In-Process and release quantities must be considered. Constraints (12) model the Work-In-Process for the remaining operations in the first period based on the completed quantity (Y variables) in the first period and the initial Work-In-Process. Constraints (13) are for the first operation in which the release must be considered and, based on the completed quantity (Y variables) in the first operation, model the flow of the Work-In-Process from period t to period p ($t \leq p$). Constraints (14)-(15) compute the remaining Work-In-Process in each of the remaining operations and its flow from period t to period p ($t \leq p$) with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (16) are resource capacity constraints.

III. COMPUTATIONAL RESULTS

A. Experimental design

Using Anylogic, a multi-method simulation software (version 8.4) which interacts with the standard solver IBM ILOG CPLEX (version 12.6), the simulation model starts by creating the required agents such as the routes of products, product operations, work-centers, etc. These agents are then fed with data from Excel files (data related to the fab such as work-centers, number of machines in each work-center, processing times, etc). Finally, the parameters of the simulation model and of the global scheduling model are initialized, and lots of products are generated following the product release scheme. The run of the simulation model starts using the FIFO dispatching rule for scheduling decisions. Next, the global scheduling optimization model is called in a rolling horizon by a simulation trigger event. After collecting dynamic parameters from the current status of the simulation model, such as current Work-In-Process levels in work-centers, and static parameters, such as future releases and aggregate resource capacities, the global scheduling model is solved to determine production targets. In the meantime, the simulation model is paused. When the optimization is completed the production targets Y_{glp} for product g in operation l and period p determined by the global scheduling model is then imposed as constraints at the work-center level in terms of production quantities of each product to complete at each operation in each period. Then, the simulation model resumes and tracks these production quantities using the FIFO dispatching rule combined with the so-called Production Target Dispatching Rule (PTDR). The Production Target Dispatching Rule (PTDR) ensures that the target for a particular product g is reached in a given operation in each period using a controller variable. As we assume that we are not working with a new factory, a warm-up time (time to load the factory) of six months was considered. The warm-up time is excluded when collecting statistical data in order to make sure that the system is analyzed with relevant data. We consider a uniform and continuous release of lots. The scheduling horizon (optimization horizon) is set to 33 days, the period length in the scheduling horizon is set to a shift (8 hours) and the global scheduling model is called in the simulation once a day (24 hours) more that 300 times. The simulation is stopped after 18 months. The calculation time for each execution of the global scheduling model is less than one minute.

Numerical tests have been conducted on industrial instance with 449 machines in 203 work-centers, which are shared between operations of various types of products. Products have between 352 and 622 operations. Five product families are considered and the release scheme is one lot for each product every 205 minutes. Experiments were performed on a computer with windows 10 as operating system, processor Intel(R)Xeon(R) CPUE3-1240v5, 2*3.50 GHz and 32 Go of RAM.

Av. CT represents the average cycle time and *R.Quantities* the release quantities; *%Achieved T.* is the percentage of achieved throughput computed based on the estimated throughput. The estimated throughput is the quantity of products released into the system after the warm-up time (time to load the factory) up to the end of the simulation horizon minus the maximum cycle time of all products. This implies that the estimated throughput should be completed before the end of the simulation horizon. In our computational experiments, we considered a maximum cycle time for all products equal to three months. *WT.Av.CT* is the weighted total average cycle time where the average cycle time of each product is multiplied by its throughput before calculation, and finally, *T. Throughput* is the total throughput of the system.

B. Analysis

Tables II, III and IV present the numerical results obtained with respectively the simulation model without any global scheduling strategy, the simulation model coupled with the push strategy (model of push strategy) and the simulation model coupled with the Time at Operation strategy.

TABLE II: Results of simulation model without any global scheduling strategy

			Products		
Indicators	1	2	3	4	5
Av. CT	**44.2**	**53.3**	*41.1*	85.4	*59.7*
R.Quantities	2,529	2,529	2,529	2,529	2529
Throughput	**2,099**	**2,009**	*2,125*	1,715	*1,954*
%Achieved T.	108.2%	103.6%	109.6%	88.4%	100.8%
WT.Av.CT	**55.6**				
T. Throughput	**9,902**				

Compared to Table II, the results in Table III show that the throughput of product 3 is significantly reduced by 9.7% and the cycle time increases by 39.9%. The same can be said for product 5 with respectively a decrease of 20.3% and an increase of 36.8% of the throughput and the cycle time. This is because products 3 and 5 have a relatively small number of operations (352 and 415 respectively) and because they compete for the same resources than products 1 and 2. Products 1 and 2 have a larger number of operations (501 and 440 respectively), and are thus prioritized by the push strategy. The throughput of products 1 and 2 increase both by 4.8%, and their cycle times decrease respectively by 24.0% and 2.6%. This shows the drawback of the push strategy discussed in Section II. Product 4 has the largest number of operations (622 operations) and, based on the results in Tables II, III and IV, is not competing for the same resources than the other products. This is why the cycle times of product 4 do not change much from Table II to Table III.

TABLE III: Results of simulation model with push global scheduling strategy

	Products				
Indicators	1	2	3	4	5
Av. CT	**33.6**	**51.9**	*57.5*	86.5	*81.7*
R.Quantities	2,529	2,529	2,529	2,529	2529
Throughput	**2,200**	**2,106**	*1,919*	1694	*1557*
%Achieved T.	113.5%	108.6%	99.0%	87.4%	80.3%
WT.Av.CT	**59.9**				
T. Throughput	**9,476**				

We can observe in Table IV, Figures 4 and 5 for the Time at operation strategy, that the cycle time (throughput) of all products is lower (larger) than in Table II for the simulation model without any global scheduling strategy. In addition, the total throughput is larger with a lower total average cycle time compared to Tables II and III. This is explained by the fact that the Time at Operation strategy handles the Work-In-Process so that quantities of product arriving at different times in the queue of an operation are penalized differently. The Work-In-Process that has been waiting the most is prioritized and not the entire Work-In-Process of a single product. Thus, the Work-In-Process of all products might remain at the end of each period in the global scheduling model.

TABLE IV: Results of simulation model with time at operation global scheduling strategy

	Products				
Indicators	1	2	3	4	5
Av. CT	43.8	44.6	34.4	79.1	58.7
R.Quantities	2,529	2,529	2,529	2,529	2529
Throughput	2,125	2,132	2,232	1,792	1976
%Achieved T.	109.6%	110.0%	115.1%	92.4%	101.9%
WT.Av.CT	**50.9**				
T. Throughput	**10,257**				

IV. CONCLUSIONS

This paper presents and studies two different global scheduling strategies to minimize cycle times: The push strategy and the Time at Operation strategy. The Time at Operation strategy provides very good results in terms of cycle times and throughput compared to the push strategy and compared to not using any global scheduling strategy (only the simulation where a First-In-First-Out dispatching rule is implemented). This is because the time at operation strategy better manages the Work-In-Process over time. Different quantities of Work-In-Process that arrived at different times in an operation are penalized differently, to prioritize the processing of the Work-In-Process that has been waiting the most in the operation. As future research, we will investigate multi-objective approaches combining objectives such as the maximization of the productivity combined, cycle time minimization and cycle time variability minimization. We will also, compare our approach with others sophisticated dispatching rules.

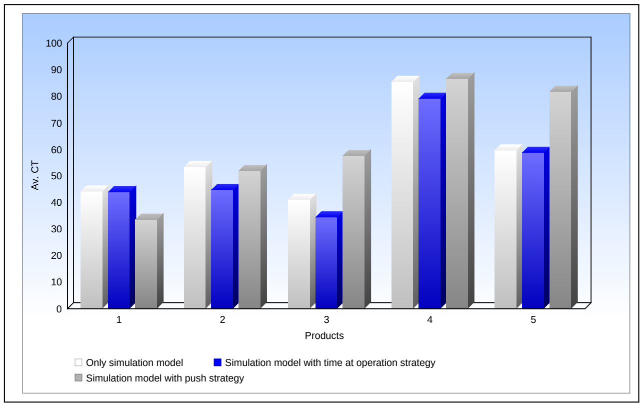

Fig. 4: Comparison of Av. CT between simulation and the two global scheduling strategies

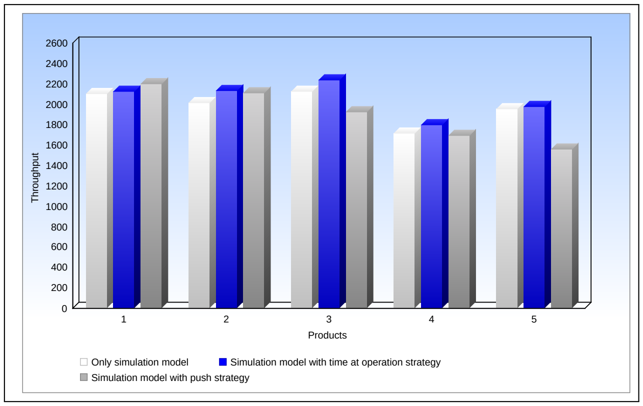

Fig. 5: Comparison of Throughput between simulation and the two global scheduling strategies

V. ACKNOWLEDGMENT

This project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737459 (project Productive 4.0). This Joint Undertaking receives support from the European Unions Horizon 2020 research and innovation program and Germany, Austria, France, Czech Republic, Netherlands, Belgium, Spain, Greece, Sweden, Italy, Ireland, Poland, Hungary, Portugal, Denmark, Finland, Luxembourg, Norway, Turkey.

REFERENCES

[1] J. K. Robinson, Understanding and improving wafer fab cycle times. Semiconductor FabTech 17(April), 2002.

[2] S. C. Lu, D. Ramaswamy, and P. Kumar, Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants. IEEE Transactions on Semiconductor manufacturing 7 (3): 374-388, 1994.

[3] S. C. Lu, D. Ramaswamy, and P. Kumar, Scheduling semiconductor manufacturing plants to reduce mean and variance of cycle-time. Proceedings. IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, IEEE, pp. 83?85, 1993.

[4] L. Mönch, J. W. Fowler, and S. J. Mason,Production planning and control of semiconductor wafer fabrication facilities: Modeling, analysis, and systems. NewYork: Springer, 2012.

[5] G. S. May, and C. J. Spanos, Fundamentals of semiconductor manufacturing and process control. John Wiley & Sons, 2006.

[6] F. Barhebwa-Mushamuka, S. Dauzère-Pérès, C. Yugma, Multi-objective optimization for work-in-process balancing and throughput maximization in global fab scheduling, IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE, pp. 697?702, 2019.

[7] F. Barhebwa-Mushamuka, S. Dauzère-Pérès, C. Yugma, Work-in-process balancing control in global fab scheduling for semiconductor manufacturing, In proceedings of the Winter simulation Conference, pp. 2257-2268, 2019.

[8] R. Sadeghi, S. Dauzère-Pérès, C. Yugma, A Multi-Method Simulation Modelling for Semiconductor Manufacturing. IFAC-PapersOnLine 49 (12): 727-732, 2016.

[9] Y. Mati, S. Dauzère-Pérès, and C. Lahlou, A general approach for optimizing regular criteria in the job-shop scheduling problem, European Journal of Operational Research 212(1): 33 ? 42, 2011.

[10] T. Chen, A systematic cycle time reduction procedure for enhancing the competitiveness and sustainability of a semiconductor manufacturer. Sustainability 5(11): 4637?4652, 2013.

[11] C.-F. Chien, and C.-H. Hu, Segmented wip control for cycle time reduction, IEEE International Symposium on Semiconductor Manufacturing, IEEE, pp. 265?268, 2006.

[12] T.-K. Hwang, and S.-C. Chang, Design of a lagrangian relaxation-based hierarchical production scheduling environment for semiconductor wafer fabrication. Transactions on Robotics and Automation 19(4): 566?578. 2003.

[13] P. O. Kriett, S. Eirich, and M. Grunow, ,Cycle time-oriented mid-term production planning for semiconductor wafer fabrication. International Journal of Production Research 55, 4662?4679, 2017.

[14] D. Babbs, and R. Gaskins, Effectiveness of small batch size on cycle time reduction in a conventional 300mm factory. IEEE/SEMI Advanced Semiconductor Manufacturing Conference, IEEE, pp. 105?110, 2007.

[15] E. Zarifoglu, J. J.Hasenbein, and E. Kutanoglu, Lot size management in the semiconductor industry: Queueing analysis for cycle time optimization, IEEE Transactions on Semiconductor Manufacturing 26(1): 92?99, 2012.

[16] D. Eberts, S. Keil, F. Peipp, and R. Lasch, Shortening of cycle time in semiconductor manufacturing via meaningful lot sizes, 2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), IEEE, pp. 34?41, 2015.

[17] K. Rozen, and N. M. Byrne, Using simulation to improve semiconductor factory cycle time by segregation of preventive maintenance activities, Proceedings of the 2016 Winter Simulation Conference, IEEE Press, pp. 2676?2684, 2016.

[18] R. C. Leachman, J. Kang, and V. Lin, Slim: Short cycle time and low inventory in manufacturing at samsung electronics, Interfaces 32(1): 61?77, 2002.

[19] M. Mittler, and A. K. Schoemig, Comparison of dispatching rules for semiconductor manufacturing using large facility models, In proceedings of the Winter simulation Conference, pp. 709?713, 1999.

[20] M. Mittler, A. Schoemig, and N. Gerlich, Reducing the variance of cycle times in semiconductor manufacturing systems. In International Conference on Improving Manufacturing Performance in a Distributed Enterprise: Advanced Systems and Tools, 1995.

[21] H. J. Yoon, and D. Y. Lee, A control method to reduce the standard deviation of flow time in wafer fabrication, IEEE transactions on Semiconductor Manufacturing 13(3): 389?392, 2000.

[22] A. N. Swe, A. K. Gupta, A. I. Sivakumar, and P. Lendermann, Cycle time reduction at cluster tool in semiconductor wafer fabrication, 2006 8th Electronics Packaging Technology Conference, IEEE, pp. 671?677, 2006.

[23] E. Akcalt, K. Nemoto, and R. Uzsoy, Cycle-time improvements for photolithography process in semiconductor manufacturing, IEEE Transactions on Semiconductor Manufacturing 14(1): 48?56, 2001.