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

WORK-IN-PROCESS BALANCING CONTROL IN GLOBAL FAB SCHEDULING FOR SEMICONDUCTOR MANUFACTURING

Félicien Barhebwa-Mushamuka
Stéphane Dauzère-Pérès
Claude Yugma

Mines Saint-Etienne
Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
CMP, Department of Manufacturing Sciences and Logistics
Gardanne, F-13541, FRANCE

ABSTRACT

This paper addresses the problem of controlling the Work-In-Process (WIP) in semiconductor manufacturing by using a global scheduling approach. Global fab scheduling steers scheduling decisions at work-center level by providing objective in terms of production targets, i.e. product quantities to complete for each operation and at each period on a scheduling horizon. A WIP balancing strategy is proposed to minimize the product mix variability in terms of throughput and cycle time. This strategy is enforced using a global scheduling optimization model which is formulated as a linear programming model. The global scheduling model is coupled with a generic multi-method simulation model for evaluation purpose. Computational results on industrial data show that the WIP balancing strategy provides a better control of the WIP in the system and helps to minimize product mix variability while maintaining high throughput.

1 INTRODUCTION

Semiconductor manufacturing creates integrated circuits through two main stages. In the first stage, the Front-End, wafers are used as starting material on which chips are produced. In the Back-End stage, testing and packaging operations are performed before delivering the final products to customers. Features of the Front-End stage, such as re-entrant flows, hundreds of operations for each product and different processing schemes for workshops, make semiconductor manufacturing probably the most complex manufacturing system (Mönch et al. 2011).

In this paper, the problem under study involves the scheduling decisions in Front-End manufacturing facilities (fabs) in which products are manufactured on different machines grouped into work-centers. Each work-center corresponds to machines with the same capabilities.

The Work-In-Process (WIP) corresponds to the products already in the fab but not yet completed. Balancing the Work-In-Process is of great importance because it allows a good capacity utilization by ensuring that products are properly distributed in the fab. It also ensures that all products steadily move forward to the completion of their operations.

It is difficult to improve Key Performance Indicators (KPIs) such as cycle time, throughput, and on time delivery without a thorough management of the WIP. WIP balancing control (ensuring that Work-In-Process is properly distributed throughout the whole manufacturing system) is considered as an efficient method to improve KPIs (Lee and Lee 2003). Strategies to avoid unbalanced WIP have been studied in different ways:

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

- WIP balancing and WIP controlling strategies based on the operation view point (see (Dabbas and Fowler 2003; Li et al. 1996; Leachman et al. 2002; Fordyce et al. 1992; Bureau et al. 2007)). Priorities and scheduling policies are used to balance Work-In-Process on the different operations of products. In (Bureau et al. 2007), different layers of operations are created. WIP targets are defined for each layer and the balancing is achieved by minimizing the deviation between the current WIP and the defined WIP target in each layer. In (Leachman et al. 2002), different methodologies and algorithms are proposed for Short Cycle Time and Low Inventory Management (SLIM). Continuous time target output schedule or continuous-time target cycle times are translated into target profile of WIP through the sequence of operations for each product. Instead of individual lots, operations are considered as the main scheduling object in SLIM. (Fordyce et al. 1992) propose a daily output planning by using WIP target for each operation of a product. The goal is to provide quantities of lots that should be processed in each operation at a given period in order to meet immediate demand or to anticipate future demand.
- WIP balancing strategies based on the work-center view point (see (Zhou and Rose 2010; Chung and Jang 2009; Lee and Lee 2003)). WIP targets are defined for each work-center, generally bottleneck work-centers. The balancing is achieved by minimizing the deviation between the current WIP and the defined WIP target.

Using both the operation and work-center view points, WIP control is either performed by using targets and/or priorities. (Chung and Jang 2009) study a WIP balancing procedure using production targets for throughput maximization in semiconductor manufacturing. The balancing is achieved by sending detailed target production quantities to bottleneck work-centers. These targets are transformed from production quantities sent from production planning. The same control is implemented in (Lee and Lee 2003). Besides production quantities, targets are also based on the WIP or on the cycle time. The idea is to divide every route (sequence of operations for one product) in layers (a layer is a set of consecutive operations) which correspond to a logical separation that allows intermediate controls on products during manufacturing. WIP targets or cycle time targets are estimated for each layer (see (Lee et al. 2008; Bureau et al. 2007)). The objective is then to ensure that the difference between the current WIP (resp. current cycle time) and the WIP target (resp. cycle time target) is minimized for each layer.

As it is difficult to estimate the exact WIP level for each operation or each layer (Lee et al. 2002), another way to balance the WIP in the factory is achieved by using a combination of several dispatching rules. (Zhou and Rose 2011) propose a new composite dispatching rule which for instance combines the operation due date rule, the shortest processing time rule and the least work at next queue rule to consider several objectives simultaneously.

Besides these methods which use a route subdivision in layers, another approach based on a so-called WIP control table is discussed in (Zhou and Rose 2010). In this approach, each upstream work-center maintains a WIP control table, which contains the current WIP information of the downstream work-centers such as the WIP target, the current WIP and the difference between the WIP target and the current WIP. This WIP control table is regularly updated and allows the upstream work-centers to optimally supply lots to the downstream work-centers. Those targets are estimated either based on historical data or by simulation.

WIP balancing strategies can also use priorities to speed up or slow down lots in layers for the purpose of smoothing the workload in different layers (Bureau et al. 2007). According to the due date and the workload information, a priority matrix table can also be used to assign lot priorities to manage the WIP to balance the overall workload of the manufacturing system. (Zhou and Rose 2012) provide a priority matrix to control the flow of lots in the system and a WIP calibration method whose purpose is to recover the WIP balance due to an event such as an unpredictable machine failure.

Due to the complexity of semiconductor manufacturing, scheduling decisions are usually taken rather independently for individual work-centers or grouped work-centers (Mönch et al. 2011). This is for instance the case when scheduling lots on non-identical parallel machines in the photolithography work-center, see e.g. (Bitar et al. 2016) or (Ham and Cho 2015), or when scheduling on batching machines in the cleaning

and diffusion work-center, see e.g. (Yugma et al. 2012; Jung et al. 2014) or (Knopp et al. 2017). Dispatching rules, that select a lot among lots waiting in front of a machine when it becomes available, are still often used in practical settings. A review of dispatching rules can be found in (Varadarajan and Sarin 2006; Mönch et al. 2011).

Independent scheduling is shortsighted because work-centers do not see what is globally happening in the fab. Every work-center takes decisions based on its local information and objectives. Thus, unbalanced flows of WIP can be observed, which implies the deterioration of some global KPIs. A literature survey on scheduling in semiconductor manufacturing can be found in (Mönch et al. 2011).

The global scheduling approach in this paper adopts two views of the operational level: The global level (fab level) and the local level (work-center level). The global level uses global information (Work-In-Process in the whole fab, lot releases and aggregate resource capacity, etc.), while the local level uses local information (waiting times of lots, lots currently in queues, etc.). The global level aims at determining production targets to be followed at the work-center level, that are regularly revised in a rolling horizon. Different strategies can be implemented in the global scheduling approach depending on the KPIs that the fab manager wants to optimize. In this paper, the strategy is enforced using a global scheduling model written as a linear programming model. Its objective function embeds some Work-In-Process management goals such as the minimization of the remaining WIP in each operation and a Work-In-Process balancing control. The global scheduling model is called in a rolling horizon scheme, and a generic multi-method fab simulation model is used to represent the local view. The fab simulation model helps to evaluate the impact of our global scheduling approach. In this model, the First-In-First-Out (FIFO) dispatching rule is used in work-centers, combined with a rule to ensure that global scheduling production targets are followed.

Our approach differs from approaches in the literature which use production targets, in the sense that it does not only focus on bottleneck work-centers. In addition, our approach takes into account the interaction between work-centers, thus preventing the myopic view of independent scheduling decisions. Instead of imposing WIP level or WIP targets to the local level, it provides production targets for each product to complete at each operation and in each period on a scheduling horizon. Production targets are determined based on the global information (fab level) and the WIP balancing strategy implemented in the global sheduling model.

The paper is structured as follows. Section 2 describes the linear optimization model as well as the Work-In-Process management strategies. Section 3 presents and analyzes computational results on industrial instances. Finally, conclusions and perspectives are provided in Section 4.

2 GLOBAL SCHEDULING APPROACH

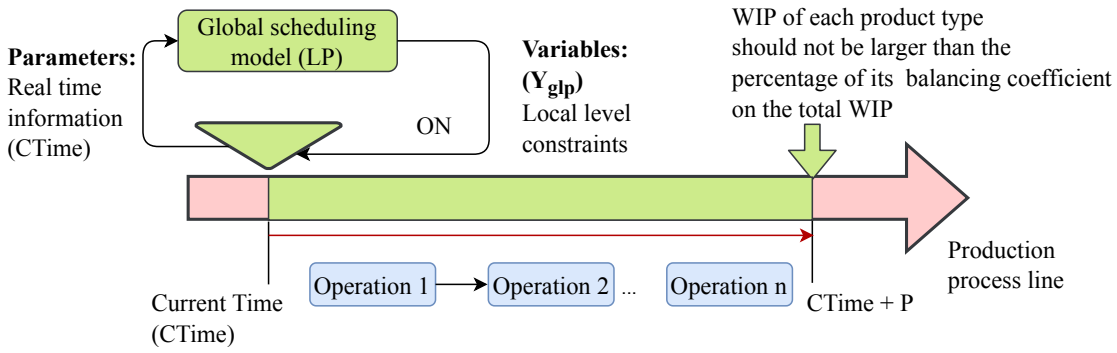

Figure 1: Global scheduling approach.

Thee global scheduling approach illustrated in Figure 1 aims at providing production targets, i.e. product quantities to complete for each operation and at each period on a scheduling horizon. Production targets

should be followed at work-center level and revised regularly in a rolling horizon. The global scheduling approach can use different strategies depending on the KPIs that the factory manager aims at optimizing. In this paper, relying on a linear programming model (global scheduling model), the strategy consists in minimizing the product mix variability and should allow products to be accelerated.

The global scheduling model is called in a rolling horizon within the simulation model by taking the current status of the simulation (real time information), and provides production targets Y_{glp} for each route g, operation l and period p which are tracked in the simulation model. As shown on Figure 1, the optimization model controls the WIP in such a way that, at the end each period, no route should have a WIP that is larger than a given percentage δ_g of the current total WIP in the factory.

2.1 Notation

Parameters:

- $\mathscr{G}$: Set of all routes (a route is a sequence of operations for a product),
- $\mathscr{K}$: Set of all work-centers,
- $\mathscr{L}_g$: Set of operations in route g,
- $\mathscr{LK}(k)$: Set of operations and routes that must be processed in work-center k, i.e $(g, l) \in \mathscr{LK}(k)$ means that operation l of route g must be processed in work-center k,
- P: Number of periods in the scheduling horizon,
- IW_{gl}: Initial WIP in operation l of product in route g,
- R_{gp}: Release quantity of product in route g in period p,
- α_{gl}: Processing time for operation l of product in route g,
- C_{kp}: Capacity of work-center k in period p,
- μ: Balancing penalty,
- δ_g: Balancing coefficient of product in route g.

Decision variables:

- Y_{glp}: Quantity of product in route g completing operation l in period p,
- W_{glp}: WIP of product in route g at operation l at the end of period p,
- X_{glp}: Quantity of product in route g arriving in operation l in period p,
- Z_{gp}: WIP maximum balancing deviation of route g in period p.

2.2 Global Scheduling Model Without WIP Balancing Control

$$Min \sum_{g\in\mathscr{G}} \sum_{l\in\mathscr{L}_g} \sum_{p=1}^{P} W_{glp} \tag{1}$$

Subject to :

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

$$W_{g11} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathscr{G} \tag{3}$$

$$W_{gl1} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathscr{G},\ \forall l \in \mathscr{L}_g,\ l \geq 2 \tag{4}$$

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

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

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

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

The objective function (1) ensures that the WIP at each operation at the end of each period is minimized. 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 aggregate resource capacity constraints.

2.3 Global Scheduling Model With WIP Balancing Control

We are now including WIP balancing control in the global scheduling model. In the new objective function, the total deviation of the WIP of each product is also minimized as shown in (9). The penalty μ controls the WIP balancing strategy. It penalizes the maximal deviation on the WIP for each product. The maximum balancing deviation of each product in each period Z_{gp} ensures the feasibility of the solution if some products can be completed while others are still in the system.

$$Min \sum_{g \in \mathscr{G}} \sum_{l \in \mathscr{L}_g} \sum_{p=1}^{P} W_{glp} + \mu \sum_{g \in \mathscr{G}} \sum_{p=1}^{P} Z_{gp} \tag{9}$$

Constraints (10) and (11) are also added to Constraints (2)-(8) in the global scheduling model. These new constraints balance the WIP between products depending on the balancing coefficients δ_g. They ensure that the WIP in route g at the end of each period cannot be larger than δ_g expressed as a percentage of the current total WIP in the factory. δ_g can be set to its trivial lower bound 100% divided by the number of products if all products have the same demand or depending on the percentage of each product in the product mix. In our experiments, δ_g depends on the release scheme.

$$\sum_{l \in \mathscr{L}_g} W_{glp} \ \leq \ \delta_g \sum_{g' \in \mathscr{G}} \sum_{l \in \mathscr{L}_{g'}} W_{g'lp} \ + \ Z_{gp} \quad \forall g \in \mathscr{G},\ \forall p \tag{10}$$

$$Z_{gp} \ \leq \ (1 \ - \ \delta_g) \sum_{g' \in \mathscr{G}} \sum_{l \in \mathscr{L}_{g'}} W_{g'lp} \quad \forall g \in \mathscr{G},\ \forall p \tag{11}$$

$$Z_{gp} \geq 0 \quad \forall g \in \mathscr{G},\ p = 1,\ \ldots,\ P \tag{12}$$

2.4 Description of the Multi-method Simulation Model

In Front-End manufacturing, each product has a route with a sequence of operations. Products of the same type are grouped in lots (a lot contains at most 25 wafers). A fab has hundreds of machines which are grouped into work-centers and each work-center is dedicated to a set of specific operations.

The simulation model, coded with the AnyLogic simulation software (version 8.4), is a multi-method simulation model which combines Discrete Event (DE) and Agent Based (AB) simulation methods. The notion of queues in Discrete-Event Simulation is used and the flexibility, behavior, and communication of agents in Agent Based simulation are used (Borshchev 2013). The main types of agents are the lots and the work-centers. Secondary agents are non-physical components such as operations and routes. For more information on the data driven generic multi-method simulation model, see (Sadeghi et al. 2016).

3 COMPUTATIONAL RESULTS

Computational results are provided in this section to analyze the performance of the global scheduling optimization model and the impact of the WIP balancing strategy on cycle times and throughput.

Sections 3.1 and 3.2 present the experiment design and configuration. Section 3.3 studies the impact of the WIP balancing control in the global scheduling optimization model. The computational experiments

show how the optimization model controls the WIP so that the flow of each product depends on its balancing coefficient. The product mix variability (on cycle time and throughput) is also analyzed by comparing the results of the simulation model coupled with the global scheduling model with and without WIP balancing control. In addition, the results of the simulation model coupled with the global scheduling model with WIP balancing control are compared to the results of only the simulation model where a FIFO dispatching rule is used in work-centers. Finally, Section 3.4 illustrates how balancing coefficients can be used to speed up products.

3.1 Experimental Design

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 WIP levels in work-centers, and parameters, such as future releases and aggregate resource capacities, the global scheduling model is solved to determine production targets. In the mean time, the simulation model is paused. When the global scheduling optimization is completed, its solution Y_{glp} 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 starts again and begins to track those production quantities using the rule detailed below.

A controller variable is set-up which indicates whether the target for a particular product g is reached at a given operation in each period. In some situations, depending on the status of the queue of the resource and of the production targets, a product can have high throughput with long cycle times or short cycle times with lower throughput. This is due to the fact that, if a product reaches its target, then its production is temporally stopped in order to track targets of other products. A lot of a stopped product will continue to be produced when all products reach their respective targets or when it is the only one in the queue of a resource.

In this paper, each period corresponds to one shift (8 hours) and the global scheduling model is called in the simulation model every 3 periods (24 hours) more than 300 times. The scheduling horizon in the global scheduling model is fixed to 33 days ($P = 99$, i.e. 792 hours).

3.2 Experiment Configuration

Numerous tests have been conducted on industrial data. The instances include 570 machines in 329 work-centers, which are shared between operations of various types of products. Products have between 104 and 315 operations in their routes. Five products are considered and lots are continuously released in the system in a uniform scheme, i.e. one lot for each product in the product mix every 280 minutes, 360 minutes, 480 minutes, 480 minutes and 480 minutes for products I, II, III, IV and V, respectively.

The linear programming model and the data driven generic multi-method simulation model developed in (Sadeghi et al. 2016) were implemented using the Anylogic software (version 8.4) which interacts with the standard solver IBM ILOG CPLEX (version 12.6). 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.

As we assume that we are not working with a new factory, six months of warm-up time (time to load the factory) were used. These six months are excluded when collecting statistical data to ensure that the system is analyzed correctly. The simulation is ran for 18 months including the six months of warm-up time. The balancing penalty μ is fixed to 60,000 which is large enough to penalize the WIP deviation.

The throughput percentage that is achieved is based on the estimated throughput. The estimated throughput is computed as the difference between the release quantities from the end of the warm-up time to the end of the simulation horizon and the release quantities that correspond to the theoretical cycle time (3 months).

3.3 Comparing the Global Scheduling Model With and Without WIP Balancing Control

This section presents an analysis of the WIP management of products based on the WIP balancing coefficients. The First-In-First-Out (FIFO) rule is used as the primary dispatching rule in the generic simulation model. Table 1 shows the remaining WIP in the global scheduling model (when the optimization is completed) for each product at the 75^{th}, 85^{th} and 95^{th} calls of the global scheduling model without WIP balancing control. While Table 2 shows the remaining WIP in the global scheduling model for each product at the 75^{th}, 85^{th} and 95^{th} calls of the global scheduling model when using WIP balancing control. In Table 1, the flow scheme is not imposed to the products, while the WIP is much better controlled in Table 2 where each product flow depends on its associated balancing coefficient.

Table 1: Global scheduling model without WIP balancing control.

	Remaining WIP in global scheduling model					
Products	75^{th} call		85^{th} call		95^{th}call	
Product I	103	14.8%	125	16.0%	151	16.5%
Product II	118	17.0%	136	17.4%	162	17.7%
Product III	249	35.8%	273	34.9%	361	39.4%
Product IV	218	31.4%	195	24.9%	178	19.4%
Product V	7	1.0%	53	6.8%	64	7.0%
Total	695	100%	782	100%	916	100%

Table 2: Global scheduling model with WIP balancing control.

		Remaining WIP in global scheduling model					
Products	δ_g	75^{th} call		85^{th} call		95^{th}call	
Product I	45%	268	30.0%	295	31.0%	340	34.5%
Product II	16%	181	20.3%	196	20.6%	183	18.5%
Product III	13%	147	16.5%	160	16.8%	165	16.7%
Product IV	13%	150	16.8%	150	15.8%	149	15.1%
Product V	13%	147	16.5%	150	15.8%	149	15.1%
Total	100%	893	100%	951	100%	986	100%

Note that the smaller the balancing coefficient of a product, the higher the speed of the product's flow. But, in some particular cases, it can be difficult to slow down a product with a very small number of operations or to speed up a product with a large number of operations. Table 2 shows the flows of Products III, IV and V at almost the same rate based on their balancing coefficients. The comparison of Tables 5 and 4 shows the benefit on product mix variability (on cycle time and throughput) of imposing a flow scheme for each product.

The results of the simulation model coupled with the global scheduling model with (Table 5) and without (Table 4) WIP balancing control are provided as well as the results of the simulation model only, i.e. without the global scheduling approach (Table 3). The simulation model was run using the First-In-First-Out dispatching rule, combined with the rule for production targets when the global scheduling approach is used.

The product mix variability is analyzed by using the InterQuartile Range (IQR), which is a robust variability measure. It indicates the central 50% dispersion of values in the data set and is computed based on the median. It is shown in Tables 3, 4 and 5 and on Figure 2 that the global scheduling approach using the WIP balancing strategy leads to the best finished product mix variability.

Table 3: Simulation without global scheduling approach.

	Product I	Product II	Product III	Product IV	Product V
Average Cycle Time (days)	63.0	63.2	64.5	50.9	50.6
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,398	1,097	816	868	870
Percentage Throughput Achieved	98.4%	99.3%	98.4%	104.7%	104.9%
InterQuartile Range (Cycle Time)	25.2	22.9	23.9	19.6	19.8

Table 4: Global scheduling approach without WIP balancing control.

	Product I	Product II	Product III	Product IV	Product V
Average Cycle Time (days)	52.5	53.9	79.8	30.5	65.9
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,474	1,157	749	1050	803
Percentage Throughput Achieved	103.8%	104.7%	90.3%	126.6%	96.9%
InterQuartile Range (Cycle Time)	20.0	15.4	33.7	4.9	25.1

Table 5: Global scheduling approach with WIP balancing control.

	Product I	Product II	Product III	Product IV	Product V
Balancing coefficients δ_g	45%	16%	13%	13%	13%
Average Cycle Time (days)	65.1	54.4	53.4	49.9	51.8
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,380	1,142	858	884	855
Percentage Throughput Achieved	97.2%	103.3%	103.4%	106.6%	103.1%
InterQuartile Range (Cycle Time)	23.3	21.9	19.4	18.5	20.2

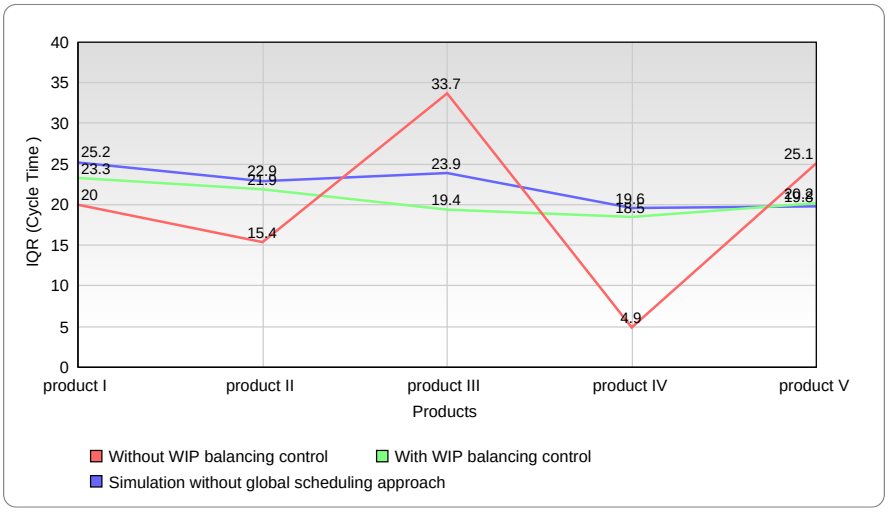

Figure 2: Product InterQuartile Ranges (Cycle Times).

3.4 Using Balancing Coefficients to Speed up Products

To speed up a product, its balancing coefficient should be set to a small value. A small value of the balancing coefficient indicates that the product should have a small remaining WIP in the fab at the end of the scheduling horizon in the global scheduling model. Tables 6 and 7 show the remaining WIP for each product at the 75^{th} , 85^{th} and 95^{th} calls of the global scheduling model for different values of the balancing coefficients. In Table 6, Products I and II are accelerated by changing their balancing coefficients respectively from 45% to 10% for Product I and from 16% to 15% for Product II.

Table 6: Speeding up Products I and II.

		Remaining WIP in global scheduling model					
Products	δ_g	75^{th} call		85^{th} call		95^{th}call	
Product I	**10%**	149	12.5%	153	12.8%	158	12.8%
Product II	**15%**	185	15.5%	188	15.6%	192	15.5%
Product III	25%	309	25.9%	314	26%	320	25.9%
Product IV	25%	272	22.9%	275	22.8%	280	22.7%
Product V	25%	279	23.4%	278	23.0%	284	23.0%
Total	100%	1194	100%	1208	100%	1234	100%

In Table 7, Products II and III are accelerated by changing their balancing coefficients respectively from 15% to 14% for Product II and from 25% to 14% for Product III.

Table 7: Speeding up Products II and III.

		Remaining WIP in global scheduling model					
Products	δ_g	75^{th} call		85^{th} call		90^{th}call	
Product I	32%	373	32.0%	382	32.0%	388	32.3%
Product II	**14%**	163	14.0%	167	14.0%	170	14.1%
Product III	**14%**	163	14.0%	167	14.0%	170	14.1%
Product IV	20%	233	20.0%	239	20.0%	242	20.1%
Product V	20%	233	20.0%	239	20.0%	232	19.3%
Total	100%	737	100%	872	100%	890	100%

The impact of the WIP balancing strategy on the final results is shown in Table 8 as well as in Table 9 after imposing production targets from the global scheduling model as constraints in the simulation model.

Table 8: Impact of balancing coefficients on cycle time and throughput of Products I and II.

	Product I	Product II	Product III	Product IV	Product V
δ_g	10%	15%	25%	25%	25%
Average Cycle Time (days)	**52.4**	**54.4**	76.9	51.0	50.5
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,475	1,153	760	867	870
Percentage Throughput Achieved	**103.8%**	**104.3%**	91.7%	104.5%	104.9%
Weighted Total Average Cycle Time	55.9				
Total Throughput	5,125				

A product can be accelerated using its WIP balancing coefficient as shown in Table 8 and the contribution of each product on the overall throughput follows the balancing coefficients. In comparison with the results presented in Table 5, Table 8 shows the acceleration of Products I and II due to their small WIP balancing coefficients while other products are slowed down to an acceptable level.

Table 9 shows how Products II and III are accelerated while other products are slowed down due to their larger balancing coefficients. Comparing to the results in Table 8, Product II reaches 105.3% of achieved throughput and Product III reaches 107.2% of achieved throughput, while Product I is slowed down with a change of its balancing coefficient from 10% to 32% but reaches 102.9% of achieved throughput.

Table 9: Impact of balancing coefficients on cycle time and throughput of Products II and III.

	Product I	Product II	Product III	Product IV	Product V
δ_g	32%	14%	14%	20%	20%
Average Cycle Time (days)	55.9	**53.8**	**47.3**	54.2	53.9
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,462	1,164	889	856	854
Percentage Throughput Achieved	102.9%	**105.3%**	**107.2%**	103.2%	103.0%
Weighted Total Average Cycle Time	53.3				
Total Throughput	5,229				

Note also that there is a limit on the acceleration of a product, which is due to the fact that the product is very slow (long process times) or has more operations. Slowing down a product is also limited because the product is very fast(short process times) or has fewer operations. This is illustrated on Product I where, even with a balancing coefficient $\delta_g = 32\%$, a throughput of 102.9% is still achieved with a cycle time of 55.9 days.

4 CONCLUSION AND PERSPECTIVES

This paper proposed a WIP balancing strategy in a global scheduling approach for semiconductor manufacturing. The approach uses the global information at fab level and determines production targets, that should be followed by work-centers at the local scheduling level. These targets are production quantities to complete for each product at each operation and in each period on a scheduling horizon. A global scheduling model, more precisely a linear programming model, is proposed to optimize production targets. The global scheduling approach allows the interaction between work-centers to be taken into account, thus preventing the myopic view of independent scheduling decisions in each work-center. The global scheduling model is called in a rolling horizon within a simulation model to validate the approach.

The WIP balancing strategy in this paper aims at controlling the flow of products to minimize the product mix variability and to speed up products. To enforce this strategy, smoothing constraints are added in the global scheduling model and a WIP balancing penalty in the objective function. The effectiveness of the global scheduling approach and the impact of the WIP balancing strategy are demonstrated by computational results on industrial data that are discussed.

Our future agenda includes the integration of additional constraints and objective functions within a multi-objective optimization approach, and the development of a novel solution approach to solve the resulting and more complex model with a large number of products.

ACKNOWLEDGMENTS

This project has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737459 (project Productive4.0). 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

Bitar, A., S. Dauzère-Pérès, C. Yugma, and R. Roussel. 2016. "A memetic algorithm to solve an unrelated parallel machine scheduling problem with auxiliary resources in semiconductor manufacturing". *Journal of Scheduling* 19(4):367–376.

Borshchev, A. 2013. *The big book of simulation modeling: multimethod modeling with AnyLogic 6*. AnyLogic North America.

Bureau, M., S. Dauzère-Pérès, C. Yugma, L. Vermariën, and J.-B. Maria. 2007. "Simulation results and formalism for global-local scheduling in semiconductor manufacturing facilities". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1768–1773. Washington, DC: Institute of Electrical and Electronics Engineers, Inc.

Chung, J., and J. Jang. 2009. "A WIP balancing procedure for throughput maximization in semiconductor fabrication". *IEEE Transactions on Semiconductor Manufacturing* 22(3):381–390.

Dabbas, R. M., and J. W. Fowler. 2003. "A new scheduling approach using combined dispatching criteria in wafer fabs". *IEEE Transactions on Semiconductor Manufacturing* 16(3):501–510.

Fordyce, K., D. Dalton, B. Gerard, R. R. Jesse, R. Sell, and G. G. Sullivan. 1992. "Daily output planning: Integrating operations research, artificial intelligence, and real-time decision support with APL2". *Expert Systems with Applications* 5(3–4):245–256.

Ham, A. M., and M. Cho. 2015. "A practical two-phase approach to scheduling of photolithography production". *IEEE Transactions on Semiconductor Manufacturing* 28(3):367–373.

Jung, C., D. Pabst, M. Ham, M. Stehli, and M. Rothe. 2014. "An effective problem decomposition method for scheduling of diffusion processes based on mixed integer linear programming". *IEEE Transactions on Semiconductor Manufacturing* 27(3):357–363.

Knopp, S., S. Dauzère-Pérès, and C. Yugma. 2017. "A batch-oblivious approach for Complex Job-Shop scheduling problems". *European Journal of Operational Research* 263(1):50–61.

Leachman, R. C., J. Kang, and V. Lin. 2002. "SLIM: Short cycle time and low inventory in manufacturing at Samsung Electronics". *Interfaces* 32(1):61–77.

Lee, B., Y. Lee, T. Yang, and J. Ignisio. 2008. "A due-date based production control policy using WIP balance for implementation in semiconductor fabrications". *International Journal of Production Research* 46(20):5515–5529.

Lee, Y. H., and B. Lee. 2003. "Push-pull production planning of the re-entrant process". *The International Journal of Advanced Manufacturing Technology* 22(11-12):922–931.

Lee, Y. H., J. Park, and S. Kim. 2002. "Experimental study on input and bottleneck scheduling for a semiconductor fabrication line". *IIE transactions* 34(2):179–190.

Li, S., T. Tang, and D. W. Collins. 1996. "Minimum inventory variability schedule with applications in semiconductor fabrication". *IEEE Transactions on Semiconductor Manufacturing* 9(1):145–149.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011. "A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations". *Journal of Scheduling* 14(6):583–599.

Sadeghi, R., S. Dauzere-Pérès, and C. Yugma. 2016. "A Multi-Method Simulation Modelling for Semiconductor Manufacturing". *IFAC-PapersOnLine* 49(12):727–732.

Varadarajan, A., and S. C. Sarin. 2006. "A survey of dispatching rules for operational control in wafer fabrication". *IFAC Proceedings Volumes* 39(3):715–726.

Yugma, C., S. Dauzère-Pérès, C. Artigues, A. Derreumaux, and O. Sibille. 2012. "A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing". *International Journal of Production Research* 50(8):2118–2132.

Zhou, Z., and O. Rose. 2010. "A pull/push concept for toolgroup workload balance in wafer fab". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, 2516–2522. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, Z., and O. Rose. 2011. "A composite rule combining due date control and WIP balance in a wafer fab". In *Proceedings of the 20011 Winter Simulation Conference*, edited by S. Jain, R. Creasey, J. Himmelspach, K. White, and M. Fu, 2085–2092. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, Z., and O. Rose. 2012. "WIP control and calibration in a wafer fab". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 2007–2018. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

FELICIEN BARHEBWA-MUSHAMUKA is a Ph.D. student in the Department of Manufacturing Sciences and Logistics at École des Mines de Saint-Étienne, France. He received his M.S. degree in Industrial engineering and operations research from the École des Mines de Saint-Étienne, France, 2017. His Ph.D subject is about Novel optimization approaches for global fab scheduling. His email address for these proceedings is felicien.barhebwa@emse.fr.

STEPHANE DAUZERE-PERES is Professor at the Center of Microelectronics in Provence (CMP) of Mines Saint-Étienne in France and Adjunct Professor at BI Norwegian Business School in Norway. He received the Ph.D. degree from the Paul Sabatier University in Toulouse, France, in 1992; and the H.D.R. from the Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at the Massachusetts Institute of Technology, U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. He has been Associate Professor and Professor from 1994 to 2004 at the Ecole des Mines de Nantes in France. His research interests broadly include modeling and optimization of operations at various decision levels (from real-time to strategic) in manufacturing and logistics, with a special emphasis on semiconductor manufacturing. He has published 75 papers in international journals. He has coordinated multiple academic and industrial research projects. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in 2013.His email address is dauzere-peres@emse.fr.

CLAUDE YUGMA is professor in the Department of Manufacturing Sciences and Logistics at the Provence Microelectronics Center of the École des Mines de Saint-Étienne. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003 and the HDR from Jean Monnet University of Saint-Étienne ; He was a Postdoctoral student at the École Nationale Supérieure de Génie Industriel, Grenoble, from 2003 to 2004 and from 2005 to 2006 at the Provence Microelectronics Center. His research includes scheduling problems in production and logistics, scheduling simulation in semiconductor context, etc.. His email address is Yugma@emse.fr.