AN ORIGINAL SIMULATION APPROACH FOR SEMICONDUCTOR MANUFACTURING

Mickaël Bureau
Stéphane Dauzère-Pérès
Claude Yugma

Ecole des Mines de Saint-Etienne
Centre Microélectronique de Provence Georges Charpak
Avenue des Anémones - Quartier Saint-Pierre
F-13541 Gardanne, FRANCE

Leon Vermariën

STMicroelectronics
77, avenue Olivier Perroy
Z.I. Peynier Rousset
F-13790 Rousset, FRANCE

ABSTRACT

This paper deals with the consistency of global and local scheduling decisions in semiconductor manufacturing. We propose an original simulation approach to deal with global scheduling of a semiconductor manufacturing facility (fab) and local area dispatching. The aim is to ensure that global objectives are met by dynamically adapting the local behavior during the simulation. After introducing the context of this study, the framework is presented. The first experimental tests performed with this original approach are promising. Some criteria are improved such as: cycle time, number of completed lots, etc.

1 INTRODUCTION

Work-In-Process (WIP) management is an important research subject for semiconductor companies [VF06, VS06]. Production scheduling in semiconductor wafer manufacturing is well-known to be one of the most complex of the industry [Kum94]. Processes are so long and diverse and machines so expensive that decisions must be validated before operational execution. This decision-making validation is traditionally performed by simulation. Decisions can be made at different levels:

- at the global level for the WIP management decisions of the whole fab,
- and at the local level for dispatching and scheduling decisions in various areas of the fab.

Coherence between medium-term level decisions (production planning) and short-term level decisions (production scheduling) is of primary importance as it allows reduction of cycle times and thus the increase of the performance of the workshops. Indeed, in semiconductor workshops, manufacturing a product requires the coordination of various resources such as machines, operators, transport means, storage spaces, etc. This justifies the need for consistency between decisions of different levels.

In the traditional approach, decisions are made in a hierarchical way and each decision level has its own models and resolution methods. It consists in, initially, simulating the start of the lots considered on the medium level term in order to determine critical resources and to set priorities of the lots at various stages. Then, local management on each resource or set of resources is performed at a very short term level to determine the assignment of products to resources and the order of products on these resources. This approach does not explicitly ensure that local decisions are consistent with global management strategies.

The aim of this article is to provide global management with a way to speed up or slow down various flows according to various strategies. Global objectives must be achieved by adapting (if necessary) the local behavior during simulation without restarting the simulation. Indeed, standard simulations [FLT06] use "what-if" analysis, i.e. executing a simulation and, in regard of its results, re-perform this simulation with adjusted parameters. Our approach to simulate is quite different. In this article, we propose an original approach to simulate WIP management in semiconductor manufacturing. The proposed approach enables us to interact dynamically in the simulation. The paper is organized as followed. In Section 2, we give and explain the strategy adopted to ensure consistency between decision levels. In Section 3, we present simulation results based on the chosen strategy and we conclude and give some further research directions in Section 4.

2 ADOPTED STRATEGIES

For global management, some levers of action are required to drive the production or program level [VF06]. The idea we propose consists in the possibility to speed up or slow down some way of production. If on the one hand, during the production we realize that certain global objectives will not be achieved then we can decide to increase the speed of production. If, on the other hand, for an unspecified reason we realize that there is a need to slow down production we make it. The levers of action are the priorities which are given to the lots. Indeed, all the lots do not have the same priority. Certain lots have higher priority than others [EPS92, RSH01]. More a lot has a high priority more quickly it will have to be produced. To deal with this consistency between decision levels, we use an original simulation approach. Based on a standard and widely used in semiconductor manufacturing simulation tool (AutoSched AP [Bro02]), an extension was developped called Stop&Go (see Figure 1).

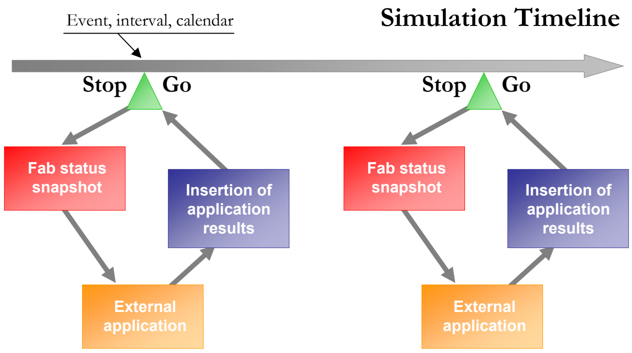

Figure 1: Stop&Go approach

This approach is summarized in five stages as follows:

1. Stop simulation temporarily,
2. Take the state of the Work-In-Process in the fab,
3. Start an external application which calculates new global priorities,
4. Insert these new updated values in the model,
5. Continue simulation with the updated values.

The global objectives to optimize are the following:

- Maximize Activity, i.e. the number of lots manufactured by the machines. The number of completed lots gives an approximation of this parameter.
- Minimize Cycle Time, i.e. the time between the start and the completion of a lot.
- Improve WIP linearity, i.e. variation of the number of lots in the various stages. Because it involves bad management of resource utilization and the losses of performance.

The main objective we first focus on is the improvement of the WIP linearity. As explained before it consists in maintaining the same level of WIP in every area of the fab over time. This WIP linearity is particularly important for bottleneck stations. Indeed these stations always need to have WIP to avoid idleness.

To deal with WIP linearity, we use the notion of block. A block can be seen as a group of steps of the semiconductor manufacturing process, that is to say a subpart of the route of each product. These blocks are logically determined, not in terms of tools or technology steps, but in terms of activity. In our case study we divided the routes in 10 blocks.

The goal is to obtain a good WIP balancing in these blocks and to minimize the WIP variation in each block to achieve WIP linearity. For this, we use WIP levels targets for each block, as well as a tolerance interval. This interval corresponds to a percentage of the target under or above of which we consider that we are still within the target. Then, when the WIP in a block does not fit the target, we try to react and correct this deviation by changing the priority of the lots in this block. A more precise description of the algorithm used is provided in Section 3.1.

To take the calculated priorities into account, a specific rule has been created, that sorts waiting lots by increasing priority. Indeed, lots with the lowest priority are considered as the most important ones. In case of equal value, this rule behaves as a First-In First-Out rule.

3 SIMULATION RESULTS

First, we want to prove the feasibility of the Stop&Go approach. To do so, we performed tests on a small but realistic model with only two products. We showed that the modification of global priorities throughout the simulation has an impact on WIP management as illustrated in Figure 2. This figure shows the simulation time versus the WIP quantity. We can see the WIP evolution of the two products during the simulation. The vertical lines indicate Stop&Go interventions to change global priorities of products. For example, we can see that at the beginning of the simulation both products have comparable WIP levels since they have the same priorities. At the first vertical line product B is accelerated and product A is slowed down. Therefore the WIP of the accelerated product decreases and the other one increases. The second vertical line changes this behavior and the last one puts back both products at the same priority. This example, although simple, shows the impact of global priorities on local WIP management.

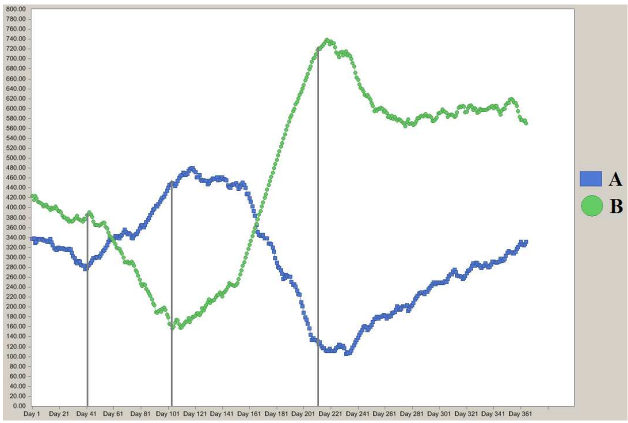

Figure 2: WIP evolution in a test simulation with global priority changes

We have tested the approach on real fab data. The following part presents the model and parameters used for the tests and the first results obtained with these data.

3.1 Simulation Parameters

The model corresponds to an actual fab, with 230 products grouped in 22 technologies and following 134 different routes. The simulation time is 20 days.

The first tested algorithm of the external application is relatively simple. We took into account targets defined for real WIP management and tolerance intervals. Moreover we fixed 5 priority levels, from 100 to 500, with steps of 100. The way these priorities are fixed is shown in Figure 3.

First we simulate the normal behavior of the simulation tool, that is to say with every lot at the same priority level (300) and without any intervention. Then we made different simulations with varying parameters. The first parameter is the frequency of interventions. To evaluate the duration and the magnitude of interventions, we tested from one suspension per hour to one suspension every two days.

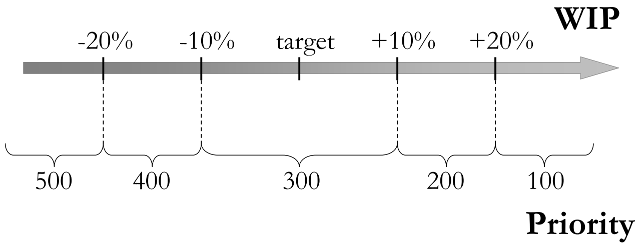

Figure 3: Procedure to fix priorities

To be more realistic, we consider the difference between the calculated priorities and the actual lots priorities. A lot priority cannot be modified by more than one step (100). For example in our case the external application gives a priority of 100. A selected lot has a standard priority of 300. So its new value will become 200, i.e. we change its value in the sense of the calculation but only by 100.

In the same realistic purpose and in regard to the first simulation results, we introduced another parameter, corresponding to the portion of modified lots in each block and at each suspension. For example with this parameter equals to 50%, we modify the priority of one lot over two and the other one keeps the same priority. Note that:

- with the portion parameter equal to 0% we find the results of the standard simulation tool,
- and even if a lot is selected to have its priority modified, the calculation can lead to the same priority results. Changes are only done if needed.

3.2 Results

Some of the results of the simulations described in the previous section can be summarized in the Table 1 to Table 4. The parameters are:

- Completed Lots: the number of finished lots during the simulation time. This is an indicator of the activity.
- Cycle Time: the average ratio between the theoretical cycle time and the observed cycle time of the completed lots. To improve the cycle time, this parameter has to be close to 1.

- Target: the number of blocks that were in the target at each daily suspension (10 blocks per 20 suspensions = 200 observations, except for the two-day suspensions where we only have 100 observations)

The tables present results for the different suspension frequencies and portions of modification. The first results column (Ref) always shows results of the simulation tool as reference. The *italic* values represent improving values compared to the simulation tool ones and the **bold** values indicate the best value for each parameter in the considered table. To compare the same values about Target we considered the values of each day suspension, that is to say we only took into account one value on the 24 ones available for the simulation with suspensions every hour. Nevertheless, for the simulation with suspensions every two days we analyze on half less values than for the others.

Table 1: Suspension every hour

	Ref	10%	20%	50%	100%
Completed Lots	**1283**	1115	1175	1152	1142
Cycle Time	2.54	**2.49**	*2.53*	*2.52*	*2.50*
Target	80	*101*	**128**	*118*	*126*

Table 2: Suspension every three hours

	Ref	10%	20%	50%	100%
Completed Lots	**1283**	1180	1148	1111	1176
Cycle Time	2.54	2.54	**2.46**	*2.49*	*2.50*
Target	80	*86*	*110*	*102*	**118**

Table 3: Suspension every day

	Ref	10%	20%	50%	100%
Completed Lots	1283	**1308**	1251	1252	1181
Cycle Time	2.54	**2.45**	*2.46*	*2.46*	2.55
Target	80	*106*	**110**	*93*	76

In complement to the highlighted values we can give the best value for each parameter:

- Completed Lots: 1308, suspension every day with 10% of modified lots
- Cycle Time: 2.43, suspension every two days with 50% of modified lots
- Target: 128, suspension every hour with 20% of modified lots

These results highlight the improvement made for the considered parameter, that is to say linearity of activity. Indeed, there are more occurrences of suspension that fit the target. The second parameter, cycle times are as-well significantly improved. More precisely, we notice that the more often we suspend the simulation, the less we have to change priority of lots to improve results. On the contrary with frequent suspensions, we can change a larger proportion of lots priorities.

Table 4: Suspension every two days

	Ref	10%	20%	50%	100%
Completed Lots	1283	**1298**	*1292*	1260	1201
Cycle Time	2.54	*2.46*	*2.47*	**2.43**	*2.50*
Target	80	**53**	*46*	*44*	30

Nevertheless the last parameter, activity, is almost always worse than with the standard simulation. This leads to another possible problem with the good results of cycle times. Indeed we improved cycle times but on a fewer number of finished products. Thus we can suppose that the remaining lots in process will have a worse cycle time and will decrease this parameter.

4 CONCLUSIONS

The results have shown that this approach looks promising but still needs adjustments of some parameters and improvement of objectives. For example the parameters taken into account at each suspension could be more numerous. The exported fields could be not only the priority but as well the time a lot spent in its current operation or queue. Likewise, the updated parameters could be the priority and a flexibility degree of this priority.

As explained in Section 3, we still encounter some problems with the loss of activity while using the Stop&Go extension. This suggests maybe to change the algorithm and to make some modifications according to different parameters, as already suggested.

Finally, this analysis of the first obtained results is a first step to develop a complete and automatic framework to deal with consistency between global strategies and local WIP management.

ACKNOWLEDGMENTS

This work was supported by STMicroelectronics and the research project Rousset 2003-2008, financed by the Communautée du Pays d'Aix, Conseil Général des Bouches du Rhône and Conseil Régional Provence Alpes Côte d'Azur.

REFERENCES

[Bro02] Brooks-PRI Automation, Inc. *AutoSched AP 7.2 User's Guide*, October 2002.

[EPS92] B. Ehteshami, R. G. Petrakian, and P. M. Shabe. Trade-offs in cycle time management: Hot lots. *IEEE Transactions on Semiconductor Manufacturing*, 5(2):101–105, 1992.

[FLT06] G. Fleury, P. Lacomme, and A. Tanguy. *Simulation à événements discrets*. Eyrolles, 2006.

[Kum94] P. Kumar. Scheduling semiconductor manufacturing plants. *IEEE Control Systems Magazine*, 14(6):33–40, 1994.

[RSH01] K. Rust, T. Stanley, and J. Hasenbein. Modeling hotter hot lots. volume Proceedings of the 2001 International Conference on Semiconductor Manufacturing Operational Modeling and Simulation, 2001.

[VF06] P. Vialletelle and G. France. An overview of an original wip management framework at a high volume / high mix facility. volume Industrial supplement and discussion papers, pages 89–92. INCOM, May 2006.

[VS06] A. Varadarajan and S.C. Sarin. A survey of dispatching rules for operational control in wafer fabrication. volume 2, pages 715–726. INCOM, May 2006.

AUTHORS BIOGRAPHIES

MICKAËL BUREAU is PhD Student at the Ecole des Mines de Saint-Etienne - Centre Microélectronique de Provence. His e-mail address is ⟨`bureau@emse.fr`⟩.

STEPHANE DAUZERE-PERES is Professor at the Ecole des Mines de Saint-Etienne - Centre Microélectronique de Provence. His e-mail address is ⟨`dauzere-peres@emse.fr`⟩.

CLAUDE YUGMA is Assistant-Professor at the Ecole des Mines de Saint-Etienne - Centre Microélectronique de Provence. His e-mail address is ⟨`yugma@emse.fr`⟩.

LEON VERMARIËN is a Corporate Industrial Engineer at STMicroelectronics Rousset. His e-mail address is ⟨`leon.vermarien@st.com`⟩.