Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

SIMULATION RESULTS AND FORMALISM FOR GLOBAL-LOCAL SCHEDULING IN SEMICONDUCTOR MANUFACTURING FACILITIES

Mickaël Bureau
Stéphane Dauzère-Pérès
Claude Yugma

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

Leon Vermariën
Jean-Bernard Maria

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

ABSTRACT

This article deals with an approach for managing scheduling in semiconductor manufacturing facilities. The proposed approach ensures consistency between global and local scheduling decisions, by ensuring that global objectives are met through dynamic adaptation of the local behavior. This approach is validated by simulation.

After describing the context and the framework of the approach, we introduce the formalism and present first simulation results obtained on real data of the fab of STMicroelectronics, Rousset. The experimental tests are promising since substantial improvements are obtained on criteria such as cycle time, number of completed lots, etc.

1 INTRODUCTION

Semiconductor manufacturing is more complicated than other types of manufacturing in terms of the number of machines, the different capabilities of machines, the huge number of manufacturing steps (several hundreds), the reentrant nature of process flows, etc. Wafer manufacturing systems present formidable challenges in their modeling, scheduling, simulation, analysis and control.

For scheduling (or dispatching) decisions, a hierarchical approach is generally adopted for semiconductor manufacturing facilities or fabs. This hierarchical approach divides the operational level in global and local levels. It consists of, initially, simulating the start of planned lots at the global level, corresponding to a short-term horizon, in order to determine critical resources and to fix priorities on the lots at the various manufacturing stages. Then, resources or sets of resources are locally managed at the local level, corresponding to real-time horizon, to determine the assignment of lots to resources as well as the sequence of lots on these resources. In this context, ensuring consistency between decision levels means that strategies and global objectives defined at the global level should be followed at the local, with some degree of flexibility.

In the literature, many papers deal with global fab management (Pillai et al. 2004) and some others deal with restricted areas (Arisha et al. 2004) or specific problems (Moench and Habenicht 2003). There are few articles that consider explicitly the interactions between global and local decision levels. This is actually given as a potential research direction in Varadarajan and Sarin (2006). This is why we developed a Work-In-Process (WIP) framework to answer the need for consistency. WIP management can be described as the management of lots and tasks that are, or are not, allocated to waiting lines (or waiting queues, or work lists). WIP management is a crucial research subject for semiconductor manufacturing companies (Vialletelle and France 2006).

The paper is organized as follows. In Section 2, we describe the simulation approach called Stop&Go which ensures consistency of decisions between global and local scheduling levels. In Section 3, we formalize these objectives and parameters. Then, in Section 4, we present simulation results that use this framework. Section 4 concludes and gives further research directions.

2 FRAMEWORK

To deal with consistency problems and to fill the gap between simulation and operational management, we based our development on a global and a local framework. This framework is illustrated in Figure 1. The figure indicates the different parameters and results needed and taken into account at global and local levels.

Global management needs lever of actions to drive the execution at product or program level. The idea is to speed up or slow down flows according to current strategic priorities. At the global level, parameters dealing with priorities are calculated and used to indicate the management strategy

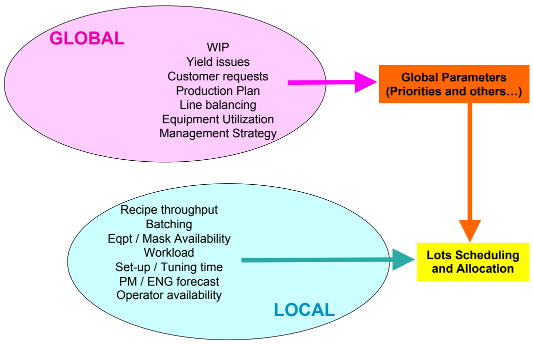

Figure 1: Global - Local Framework.

and the fab status. These calculations take into account the current WIP, customer request, the production plan, etc.

Starting from these global parameters, used at inputs of the local level, priority vectors are assigned to each group of lots according to various criteria such as position in the flow, workload, recipe throughput, technology, next process step, product, customer, lateness, etc.

The current simulation solution consists in fixing global results, i.e. lot priorities, and to simulate the local behavior in order to determine bottlenecks or identify problems. The proposed approach is guided by global parameters that can and should be updated. Indeed if, during the simulation, we note that the objectives are not or will not be reached, it is necessary to dynamically adapt the simulation parameters. Moreover, this corresponds to the reality of the fab where decisions will be made if it is observed that objectives will not be achieved. Managers ensure that production is driven to the right direction. Based on this principle, we propose a simulation approach named Stop&Go. It allows interactions between global and local levels and can be described 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 parameters,
4. Insert the new updated values in the simulation model,
5. Continue simulation with the updated values, and go to Step 1 after some time (or when an event appears).

The goal is to ensure that global parameters used at local level help to satisfy the global strategy. Indeed if, for example, the global strategy is to ensure that there is no difference in WIP production on some manufacturing stages and a machine goes down, then we have to slow down production for arriving lots in order to balance the workload on that machine. On the opposite, if an important customer requests some products rapidly, the corresponding lots have to be accelerated. This approach, that appears to be original, intends to get the simulation results closer to the reality of the fab and thus to provide more relevant results.

This approach was not covered in standard simulation tools, as AutoSched AP that has been used for our experiments. Figure 2 presents the functioning of a standard simulation tool and what has been added to implement the Stop&Go approach.

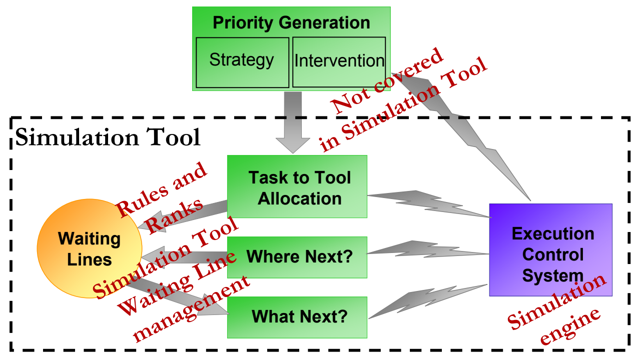

Figure 2: Simulation tool functioning and addition for Stop & Go approach.

In Figure 2, the “Execution Control System” triggers the different scenarios and the external part, “Priority Generation”, corresponds to the new feature of the Stop&Go approach. It corresponds to the engine, i.e. the core of what is described in this paper. We also have three scenarios that correspond to what happens when a machine or a lot are looking for the next activity:

- “What Next?” happens when a machine becomes free and is looking for work. It will pick a lot in its waiting line.
- “Where Next?” happens when a lot has finished its process on a machine and is looking for its next step (going to a tool or a waiting line).
- “Task to Tool Allocation” corresponds to the sorting of Waiting Lines, according to rules and ranks.

The additional part, “Priority Generation”, is dynamically providing global parameters that influence “Task to Tool Allocation”. This is the original part of our simulation approach.

Before discussing experimental tests, we describe a formalism for the priority generation mechanism.

3 FORMALISM

The fab production system $\mathscr{S}$ is constituted by heterogeneous machines $\mathscr{M} = \{M_k | k = 1, \ldots, m\}$. A set of jobs (lots) $\mathscr{J} = \{J_i | i = 1, \ldots, n\}$ has to be executed in the fab system during a time period $[1, T]$. Each job J_i has several parameters:

- w_i: the number of wafer of this job. The standard lot size is 25 wafers, so we have $\forall i,\ w_i \leq 25$.
- r_i: the release date of the lot.
- O_i: the set of all operations (as well called route) followed by the lot. This route has q_i operations and corresponds to the set $O_i = \{o_{ij} | j = 1, \ldots, q_i\}$.
- $\mathscr{M}_{ij} \subseteq \mathscr{M}$: set of machines that can perform operation j of job i.
- $o_{ij}^k = 1$ if operation j of job i is assigned on machine k, 0 otherwise.
- p_{ij}^k: process time of operation j of job i on machine k.
- t_{ij}: start time of operation j of job i.

A route can be divided in B "blocks". Blocks correspond to a logical separation that allows to have intermediate controls on lots manufacturing. Thus we determine intermediate steps (b_r) to divide route, what can be formalized as follows:

$$\begin{aligned}
&Block_1 = \{o_{ij} | j = 1, \ldots, b_1\} \\
&Block_2 = \{o_{ij} | j = b_1 + 1, \ldots, b_2\} \\
&\quad\vdots \\
&Block_B = \{o_{ij} | j = b_r + 1, \ldots, q_i\}
\end{aligned}$$

In our experiments presented in the following section, $B = 10$, i.e. all routes are divided in ten blocks, that have been provided by industrial engineering.

The fixed global objectives, given by decision-makers, are the following ones:

- Linearity (L) consists of smoothing differences that could appear between WIP level of a block and its fixed target.
- Cycle Time (CT) corresponds to the time between the completion time of the last operation of the lot and its release date.
- Activity (A) corresponds to the global use of the fab. It can be considered and calculated differently, by the machine utilization rate or the number of completed lots.

Each objective can be formalized. First, for Linearity, assume $T_{b,t}$ as the fixed activity Target and $A_{b,t}$ as the observed Activity of block b, in the period $[t-1,t]$. Linearity is equal to:

$$\mathrm{L} = \sum_{t=1,\ldots T} \sum_{b=1,\ldots B} |T_{b,t} - A_{b,t}| \tag{1}$$

Note that target can be fixed over time and thus written T_b. For the observed Activity, we name by $\mathscr{O}^{b,t} = \{o_{ij}^k | o_{ij}^k = 1 \wedge o_{ij}^k \in Block_b \wedge t-1 < t_{o_{ij}^k} + p_{ij}^k \leq t\}$ the set of operation in block b in the period $[t-1,t]$, and calculate as follows:

$$A_{b,t=} \sum_{o_{ij}^k \in \mathscr{O}^{b,t}} w_i \tag{2}$$

Concerning Cycle Time, we name by $\mathscr{J}^T = \{J_i \in \mathscr{J} | t_{o_{iq_i}} + p_{iq_i}^k \leq T\}$ the set of finished jobs before period T. Average Cycle Time of completed lots is calculated by:

$$\mathrm{CT} = \frac{\sum_{J_i \in \mathscr{J}^T} (t_{iq_i} + p_{iq_i}^k - r_i)}{|\mathscr{J}^T|} \tag{3}$$

Then Activity has been approximated by the number of completed lots during the period T.

$$\mathrm{A} = |\mathscr{J}^T| \tag{4}$$

An other consideration of this parameter is the number of "moves" performed by machines, i.e. the number of wafers treated by machines. Then Activity becomes:

$$\mathrm{A'} = \sum_{k=1,\ldots m} \sum_{i=1,\ldots n} \sum_{j=1,\ldots q_i} o_{ij}^k w_i \tag{5}$$

These global objectives have to be combined and balanced to determine an objective function of the problem. In all cases, Linearity and Cycle Time have to be minimized and on the contrary Activity has to be maximized.

This formalism is a base to understand all the parameters that have an impact on results. It has been developed, with local objectives and constraints that are not yet used.

4 SIMULATION EXPERIMENTS

To test the concept, we performed various simulation experiments using the standard simulation tool AutoSched AP (Brooks-PRI Automation, Inc. 2002). We used a customized extension to control the external calculation of global parameters. In this study, global parameters are lot priorities. The simulation is suspended at regular time intervals and a file, representing a snapshot of the fab, is created. This file contains the list of lots in the fab, their current priority, current manufacturing step and, if they are being processed, their current machine. Then, we calculate the amount of WIP per block, compare it to the WIP level target of the corresponding block and fix new priorities, that are assigned to a portion of lots in each block. These new priorities are calculated as described in Figure 3. This figure shows examples of values for the priorities and target tolerances. We assume that lower priorities accelerate lots and the standard value is 300. Target tolerances indicate that under or above 10%, we consider that WIP level fits the objective. Then between 10% and 20% we have to accelerate or slow down lots by changing their priorities to

200 or 400. Finally, with more than 20% gap from target, we make an important change on priorities, setting them to 100 or 500.

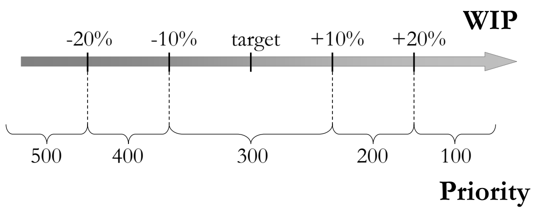

Figure 3: Example of priorities fixing.

Priorities that correspond to global parameters are sent back to the simulation via a file that contains a list of lots and their new priorities. All lots that are not in this file keep the same priority. Finally the simulation restarts where it was stopped, with the updated values. To take priorities into account while simulating, a simple dispatching rule has been used. It consists in sorting lots according to the priorities and applying the FIFO rule in case of equal priorities.

Tests were performed on actual fab data. The model has 230 products grouped in 22 technologies and following 134 different routes. A route is a subset of manufacturing stages. Each route is divided in 10 "blocks", logically determined to approximately have the same level of activity and WIP. We fixed five priority levels, from 100 to 500, corresponding to industrial practices (see Figure 3).

The results are summarized in Figures 4 to 6 for each objective, presented in Section 3. These figures show results of 41 simulation runs. The first one corresponds to a simulation without any priority intervention and constitutes our reference results. Then, we ran simulations with two varying parameters:

- Frequency of suspension: This parameter is the interval between two priority recalculations. It corresponds to the different curves with suspension every hour, every three hours, every day and every two days.
- Percentage of modified lots: This parameter is introduced to moderate the impact of the priority recalculations. Indeed, at each simulation suspension, we do not need to modify the priority of all lots in a block. We only modified a random portion of these lots. The variation of this parameter is presented on the horizontal axis, from 10% to 100%, with 10% steps.

The four values corresponding to frequencies of suspension and the ten values corresponding to the percentages of modified lots plus the reference simulation provided us 41 results, described in the following part.

Simulation execution times go from 4 minutes to 30 minutes according to the frequency of suspension. These values are reasonable, but can not really be compared as this feature can not be reproduced in the standard simulation tool. Nevertheless, we assume that the external application, that only takes 2 or 3 seconds could be accelerated to improve execution times.

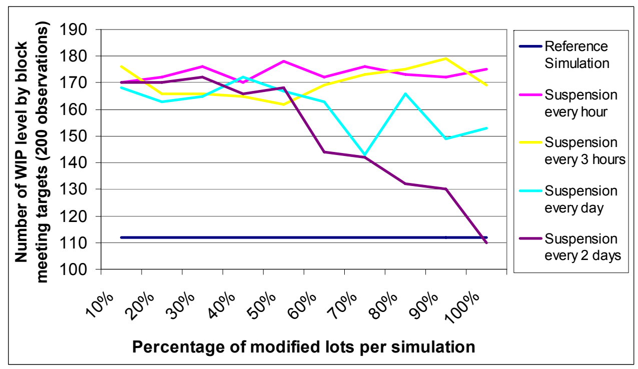

Figure 4: Number of WIP level by block meeting targets

Figure 4 corresponds to our main objective in these experiments. Due to software limitations, we could not immediately obtain activity data. Therefore, Linearity has been evaluated via WIP levels. The calculation is then comparable to the activity one. Indeed, we fixed the targets, determined by operational management and used in the actual fab, and compared them to the observed WIP level by block, trying to minimize the gap. As shown in figure 4, results are quite satisfactory, with an average improvement of more than 46%. However, looking at simulations with suspensions every day or every two days, the linearity becomes worse when a higher portion of the lots is modified at each suspension. This can be explained by the fact that a calculated priority does no longer make sense two days later and may lead to counterproductive results if it is applied to every lots. On the contrary, uptading very often priorities, leads to a very good control, independently of the portion of modified lots.

Figure 5 shows an estimator of cycle times, called Xtheormax and corresponding to the ratio between observed cycle time and expected, or theoretical, cycle time. This ratio has to be minimized. However, the analysis of results is not easy. Indeed, cycle times are always worse than the reference simulation for simulation with suspension every hour, but better for simulation with suspension every days and from 60% to 90% of modified lots. But these results are based on completed lots and thus values are not calculated on the same basis, what can explain these gaps. Still, results, even worse than reference simulation, remain acceptable.

The results on the last objective, Activity, are shown in Figure 6. These are the worst results since only few simulations have more completed lots (suspension every one or two days and 10% to 30% of modified lots). In all cases, we also get a loss of machine utilization percentage

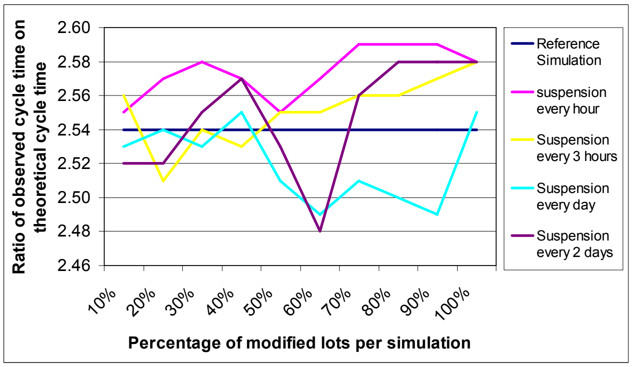

Figure 5: Average Cycle Time of completed lots

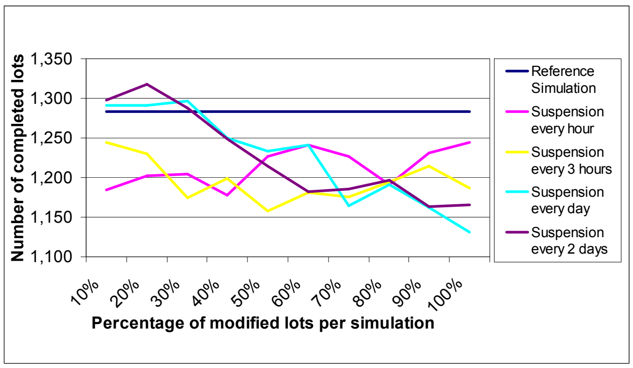

Figure 6: Completed Lots

(not presented here) that goes from 69.09 to 72.56 compared to 72.67 in the reference simulation. The formalization of activity can change according to the adopted point of view, i.e. machines or lots. The first formalization adopted, and presented in the previous section, is based on completed lots and results correspond to this formalization. Then we observe that for simulations with a high ratio of modified lots at each suspension, results are better with frequent suspensions. On the contrary, results are better when few lots are modified at each suspension but suspensions only occur every one or two days. These results are currently analyzed and their explanation and improvement is one of the main research perspective.

Globally, these result show a positive impact on the considered objective, Linearity. But this improvement is made at the detriment of the others, in particular Activity. The next step is then to complete the framework, with significant improvements of each objective, so that the fab manager could decide what are the most important objectives, using weights for instance, and ensure that local scheduling follow the defined strategy.

5 CONCLUSION AND PERSPECTIVES

This article presented a simulation approach for managing scheduling decisions in semiconductor manufacturing facility. The principle of the approach consists in dynamically interacting during the simulation by updating parameters if necessary. After describing the different steps of the proposed approach, we formalized the different parameters and objectives that have to be managed, and also performed simulation tests to validate the approach. The simulation results highlighted improvements of the considered objective (average improvement of 46% for Linearity), but a need to balance with other objectives. Indeed, Cycle Time is stable with an average increase of 0.3% but Activity decreases by more than 5% on average.

The main perspectives are to pursue our work on the formalism of the approach, and to link it with the simulation behavior to better understand the mechanisms involved and further develop the approach. Moreover, the approach has to be tested on more data sets.

ACKNOWLEDGMENTS

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

REFERENCES

Arisha, A., P. Young, and M. El-Baradie. 2004. A simulation model to characterize the photolithography process of a semiconductor wafer fabrication. *Journal of Materials Processing Technology*:2071–2079.

Brooks-PRI Automation, Inc. 2002, October. *Autosched ap 7.2 user's guide*. Brooks-PRI Automation, Inc.

Moench, L., and I. Habenicht. 2003. Simulation-based assessment of batching heuristics in semiconductor manufacturing. 1338–1345. Winter Simulation Conference.

Pillai, D. D., E. L. Bass, J. C. Dempsey, and E. J. Yellig. 2004, August. 300-mm full-factory simulations for 90- and 65-nm ic manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 17 (3): 292–298.

Varadarajan, A., and S. Sarin. 2006, May. A survey of dispatching rules for operational control in wafer fabrication. Volume 2, 715–726. INCOM.

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

AUTHOR 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 Associate-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>.

JEAN-BERNARD MARIA is an Industrial Engineer at STMicroelectronics Rousset. His e-mail address is <jean-bernard.maria@st.com>.