Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

ON THE IMPORTANCE OF OPTIMIZING IN SCHEDULING: THE PHOTOLITHOGRAPHY WORKSTATION

Abdoul Bitar
Stéphane Dauzère-Pérès
Claude Yugma

Department of Manufacturing Sciences and Logistics
Ecole des Mines de Saint-Etienne – CMP
CNRS UMR 6158 LIMOS
F-13541 Gardanne, FRANCE

ABSTRACT

This paper analyzes the impact of scheduling decisions on the capacity of a semiconductor manufacturing workstation. The study was conducted on real industrial data of a well-known bottleneck workstation, namely photolithography, which includes various complex constraints. The results of our numerical experiments show the importance of an effective optimization algorithm and how it impacts capacity, i.e. the cycle times of lots and thus the ability to schedule more lots. Additional computational results illustrate that, when the problem complexity is reduced by ignoring setup times, the impact of determining optimized schedules is also reduced.

1 INTRODUCTION

Semiconductor manufacturing processes are probably more complex than in any other industry (Gupta et al. 2006). They correspond to multistage processes with re-entrant flows, which include multiple steps such as polishing, diffusion, film deposition, *photolithography*, implant (doping), etc. (Mönch et al. 2011) For each product type, and depending on the technology, a silicon wafer goes through hundreds of process steps over a period of a few weeks. Scheduling these wafers is a complex task due to the large number of products and machines involved and to numerous complex constraints.

In this paper, we focus on the *photolithography workstation* which is generally a bottleneck area in semiconductor manufacturing facilities (also called fabs) and contain the most expensive tools. In photolithography, *wafers*, often grouped in lots of 25, have to be processed on non-identical parallel machines. The process consists in transferring an integrated circuit pattern on the wafers. To perform this operation, some photo resist must be put on the wafer which is exposed to an ultraviolet light. An auxiliary resource, i.e. a specific *mask* (or reticle), is necessary to shape the pattern on the wafers. Hence, the process can only be started if both the lot and the right mask are available. The parallel machine scheduling problem is further complicated by the fact that a mask must be on the machine for the duration of the process and that there is usually only one mask per process step of a given product. Moreover, depending on the product (also called family) of the lot, the machines need a specific configuration linked to the temperature of the machine settings to run a process. Switching a machine from one configuration to another requires a setup time related to lowering or increasing the temperature. Finally, each machine is eligible (called qualification in semiconductor manufacturing; Johnzén, Dauzère-Pérès, and Vialletelle 2011) for only a limited set of families (i.e. it cannot process the lots of other families).

Chiou and Muh-Cherng (2014) and Yan et al. (2011) study various optimization aspects in the photolithography area. Kock et al. (2011) and Morrison (2011) contributed to tool modeling in this workstation.

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

Generally, when scheduling lots in photolithography, the mean cycle time is one of the most common criteria to minimize. The goal of our study is to analyze the impact of the efficiency of scheduling algorithms on production capacity (in terms of cycles times which is directly related to the number of lots that can be processed). We used The IBM developed methodology called *Operating Curve (OC)*. OCs have been used in semiconductor manufacturing for some time as a method to benchmark productivity and to manage the trade-off between cycle time and throughput. An operating curve helps to evaluate the workstation capacity. Tirkel (2013) studies the factors contributing to production variability, and evaluates the influence of variability on Cycle Time in a semiconductor manufacturing system. The paper demonstrates the significant effect of variability on Cycle Time, and indicates that it can exceed the effect of utilization. It explains how reducing Cycle Time by decreasing variability is more effective than by decreasing utilization. Diaz et al. (2005) study the impact of masks in photolithography scheduling. As far as we know, no existing work deals with comparing operating curves with different scheduling algorithms. In addition, our study is made on different versions of the photolithography scheduling problem (the original one and a simplified version). Some papers such as the ones of Karmarkar (1987) and Dauzère-Pérès and Lasserre (2002), which are not related to semiconductor manufacturing, discuss the impact of scheduling on production planning.

The remainder of the paper is organized as follows. In Section 2, we formalize the photolithography scheduling problem. Section 3 explains how the experimental tests have been conducted. Numerical results on different industrial instances and with different scenarios are presented and discussed in Section 4. We conclude and give some perspectives in Section 5.

2 FORMALIZING THE STUDY

In this section, we summarize the description of the considered scheduling problem including the main constraints and the objective function. We consider a set of N jobs (lots) to schedule on a set of M machines using a set of ℓ auxiliary resources (also called masks) that are necessary to process the jobs on the machines.

The constraints of the problem are listed below:

- A job is processed once and only once on a qualified machine for this job.
- All jobs and all machines are available at time 0.
- There is no preemption, i.e. a job is not interrupted during its processing.
- A machine can only process one job at a time.
- Two jobs having the same required mask cannot be processed at the same time on two different machines. Indeed, masks are shared resources between jobs and processing a job on a given machine implies a *move* of the required mask towards the machine, if the mask is not already loaded in the machine.
- There are sequence and machine dependent setup times. This is due to the fact that there are job families and a setup time is required between two jobs from different families that are processed on the same machine consecutively.
- Only one mask of each type is considered: The model does not handle the case where more than one mask is available for the same family.

The processing time of a job on a machine is modeled using an $Ax+B$ model with data provided by our industrial partners. The objective function minimized in this study is the sum of the completion times, which is equivalent to the average completion time (or cycle time) of the jobs.

To solve the problem, a genetic algorithm is proposed in another paper (Bitar et al. 2014). The originality of this algorithm lies in the fact that its coding structure represents a dominant subset of solutions for our objective function. Furthermore, the proposed local search method, with its neighborhood operators, ensures that an optimal solution can be reached with the algorithm.

3 DESCRIPTION OF EXPERIMENTAL TESTS

As explained in the introduction, the photolithography workstation is often bottleneck in a semiconductor manufacturing facility. Hence, the large number of waiting lots and the numerous complex constraints make it a suitable workstation to analyze by comparing operating curves. This comparison is only performed by considering cycles times in photolithography and not in the whole factory. In addition, through our industrial partners, we have real data sets for this workstation. The genetic algorithm proposed in Bitar et al. (2014) to solve the problem considers three different criteria: Maximization of the number of produced wafers within a time horizon, minimization of the mean cycle time and minimization of the number of mask transfers between machines. As already mentioned, we focus in this paper on the mean cycle time objective and different settings of the genetic algorithm are used to build four different heuristics.

The most advanced version of the genetic algorithm is named Heuristic M. Three other heuristics, derived from Heuristic M with some modifications on the genetic algorithm key parameters, are named Heuristic 1, Heuristic 2 and Heuristic 3, respectively.

- Heuristic 3 is the least effective heuristic, but also the fastest. It is derived from Heuristic M by strongly restricting the search population size (which is set to 10) and the number of iterations without improvement before stopping the search of the solution, which is set to 10.
- Heuristic 2 has been designed by slightly increasing the values of the same key parameters than for Heuristic 3. The population size is set to 50 and the number of iterations is set to 100.
- These values are increased even more in Heuristic 1. The population size is set to 80 and the number of iterations is set to 150.
- In Heuristic M, the population size and the number of iterations are set to 200.

Note that even if the parameters of Heuristic M are larger than in Heuristic 1 and because these algorithms are stochastic, Heuristic 1 can sometimes produce better results than Heuristic M.

Real industrial data sets are used to conduct the experimental tests. Three sets of data (Instances 1, 2 and 3) corresponding to three different production periods have been extracted. The maximum number of considered lots is 560.

Each displayed result is the mean cycle time obtained on an instance. The number of machines is 15 and the number of families is 5. The processing times are between 30 minutes and 90 minutes and setup times are between 0 and 5 minutes. For each figure the X-axis represents the number of lots and the Y-axis the mean cycle time. Then, for each heuristic, a curve is built to represent the relationship between the mean cycle time and the number of lots that has been scheduled. This pattern is done twice:

- The first set of experiments corresponds to the original scheduling problem,
- The second set of experiments corresponds to a simplified version, where no setup times are considered (i.e. all lots are considered to be in the same family).

In the following section, the experimental tests are presented and discussed.

4 NUMERICAL EXPERIMENTS

4.1 Operating Curves for the General Problem

Figure 1 shows the results of the four scheduling algorithms on Instance 3. The average cycle time determined with heuristic M increases slower with the number of lots than the average cycle time determined with the other heuristics. Significant differences can be observed between each pair of heuristics. The average cycle time for the maximum number of lots is close to 400 for Heuristic M while it is twice as large for Heuristic 1, larger than 1,000 for Heuristic 2 and larger than 1,350 for Heuristic 3. Heuristic 3 is particularly ineffective, even with less than 100 jobs.

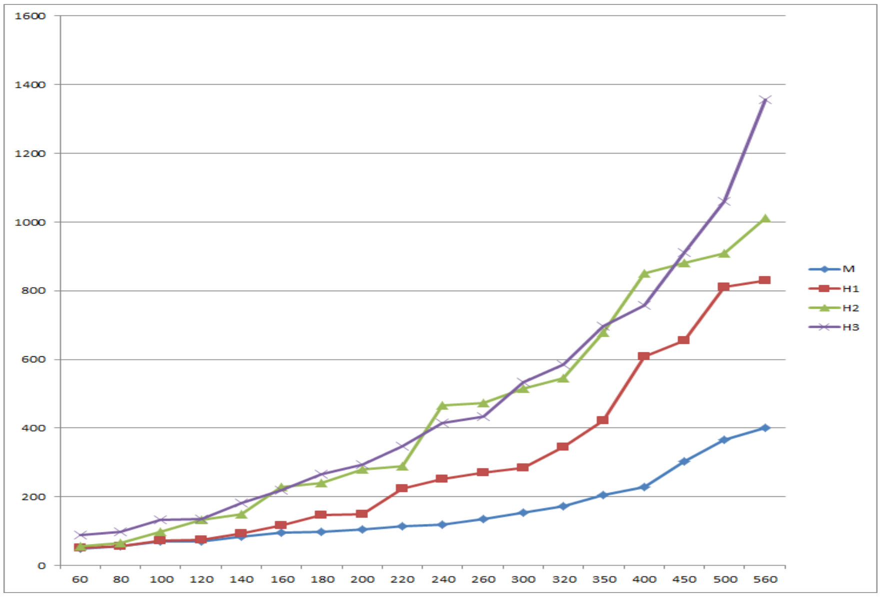

Figure 1: Operating curves with four scheduling algorithms of different quality on Instance 3.

Figures 2 and 3 show the results obtained on Instances 1 and 2, respectively. Heuristics 1 and *M* are very close, and Heuristic 1 is actually sometimes slightly better than Heuristic *M*. For Instance 3, Heuristics 2 and 3 give poor results, although Heuristic 2 is competitive up to 240 jobs but quickly worsens with more than 300 jobs. Heuristic 3 is particularly ineffective.

Figures 1, 2 and 3 illustrate the fact that the more effective the scheduling algorithm, the lower the mean cycle time. Thus, significant gain in production capacity can be obtained by designing and implementing advanced scheduling algorithms. Moreover, the robustness of an optimization algorithm that determines good schedules for any instance is interesting, whereas a less effective scheduling method might give good solutions in some cases (as in Instance 1 for Heuristic 2) but poor solutions in other cases (as in Instance 3 for Heuristic 2).

Table 1 details the results obtained on three different instances. Note that, depending on the instance, Heuristics 1 and 2 are more or less effective but the hierarchy between the four heuristics remains the same.

4.2 Operating Curves for the Problem Without Setup Times

In this section, we simplify the scheduling problem by ignoring setup times between families. Figure 4 shows that, for Instance 3, the gaps between Heuristics 1, 2 and *M* are reduced compared to Figure 1. On the other hand, Heuristic 3 (yellow curve) has not changed much since it is too bad to get very different results in this case. The three other algorithms give different results compared to the case with setup times. They all obtain better solutions and one can even note that the lines for Heuristics *M* and 1 (red line and green line) intersect.

In Figure 5 associated with Instance 1, the gaps between Heuristics 1, 2 and *M* have not narrowed compared to Figure 2. This illustrates again the robustness of optimization algorithms that remain effective independent of the instance. The average cycle time remains below 300 for the maximum number of lots with Heuristics 1 and *M*, while it is larger than 750 for Heuristic 2 and is close to 1,400 for Heuristic 4.

In Figure 6, operating curves for Instance 2 are very similar to those for Instance 1.

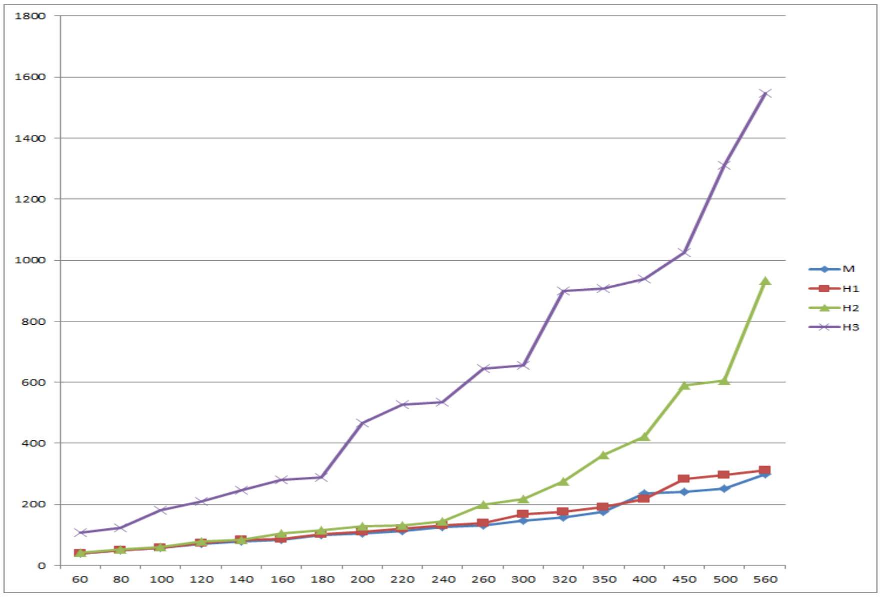

Figure 2: Operating curves with four scheduling algorithms of different quality on Instance 1.

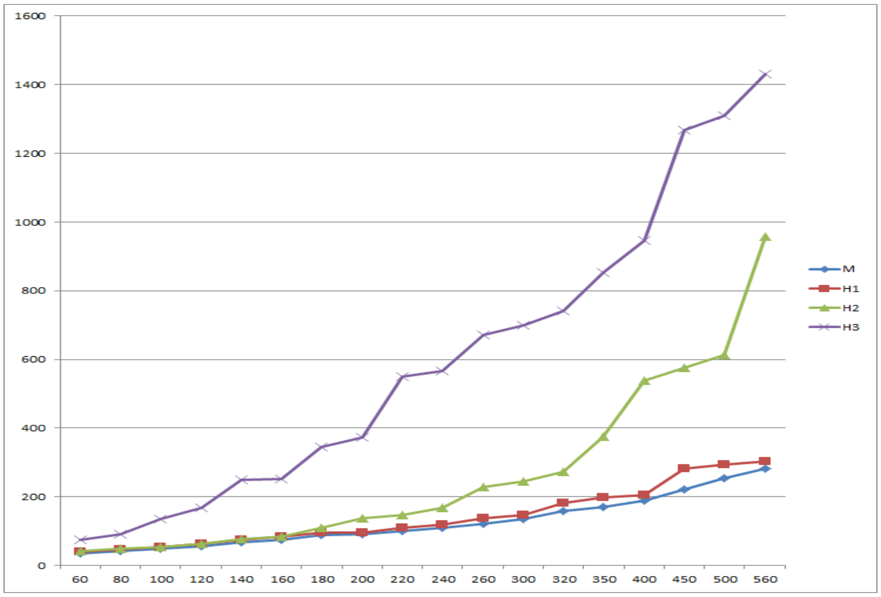

Figure 3: Operating curves with four scheduling algorithms of different quality on Instance 2.

Table 1: Mean cycle times obtained on three industrial instances.

	Instance 1				Instance 2				Instance 3			
WIP	M	H1	H2	H3	M	H1	H2	H3	M	H1	H2	H3
60	39	40	42	108	35	40	42	75	49	51	55	89
80	49	50	53	123	42	47	48	91	56	55	65	99
100	58	59	61	180	48	54	54	136	69	72	99	132
120	71	73	78	210	55	62	62	169	71	74	133	136
140	78	85	84	246	68	75	78	249	83	94	150	181
160	85	88	106	282	75	84	85	251	96	117	229	219
180	100	102	115	289	88	95	110	344	98	147	241	265
200	104	110	129	466	91	96	137	374	104	149	280	294
220	114	122	131	527	101	109	147	550	114	223	290	347
240	125	130	145	534	109	118	167	566	119	251	466	414
260	131	138	200	646	121	138	228	672	135	270	472	434
300	148	167	217	656	136	148	244	700	154	285	516	534
320	158	175	275	900	159	183	272	741	172	345	545	584
350	176	192	363	908	171	199	375	854	206	421	678	696
400	237	218	423	939	188	206	539	945	229	609	850	758
450	242	283	591	1,026	222	281	575	1,267	304	655	881	911
500	253	296	606	1,312	255	293	612	1,310	367	810	909	1,060
560	299	311	932	1,546	283	302	958	1,430	402	829	1,012	1,357

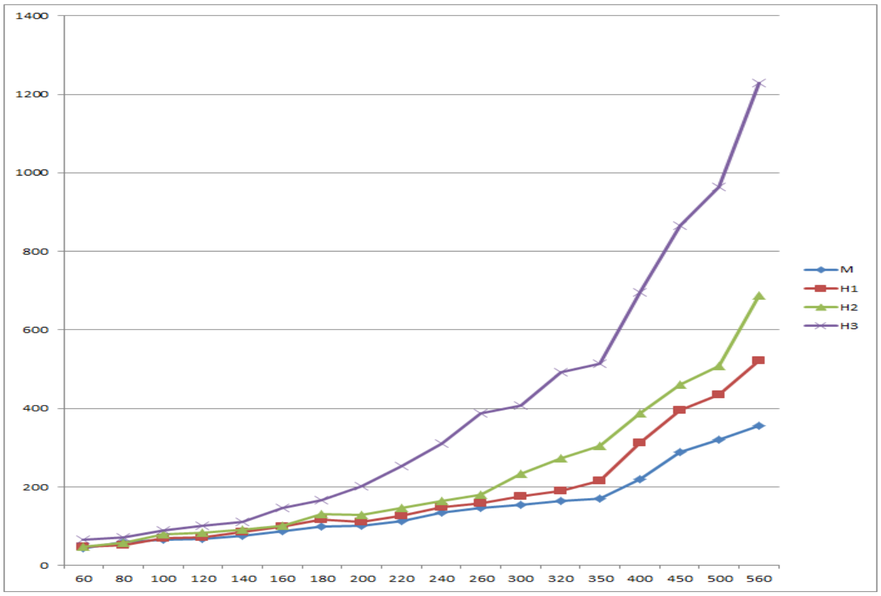

Figure 4: Operating curves with four scheduling algorithms of different quality without setup times on Instance 3.

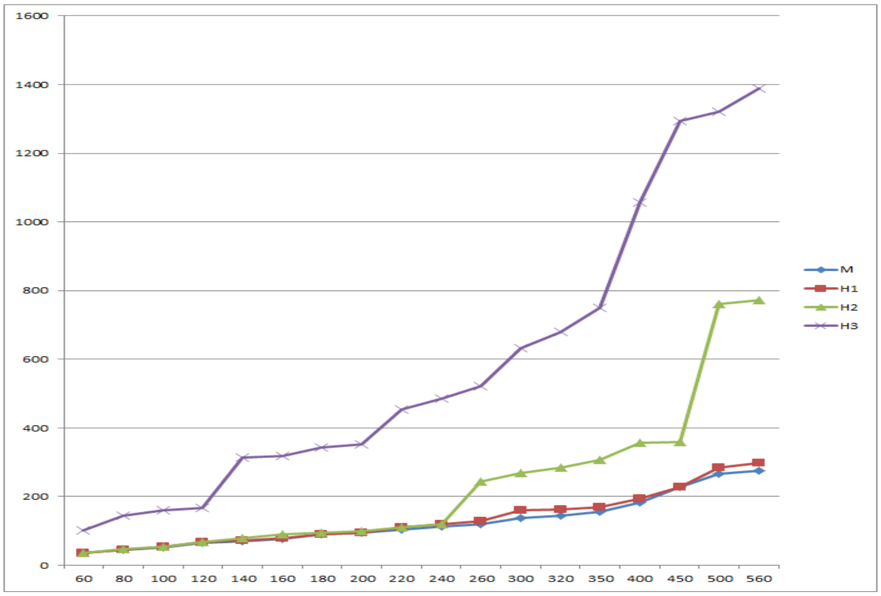

Figure 5: Operating curves with four scheduling algorithms of different quality without setup times on Instance 1.

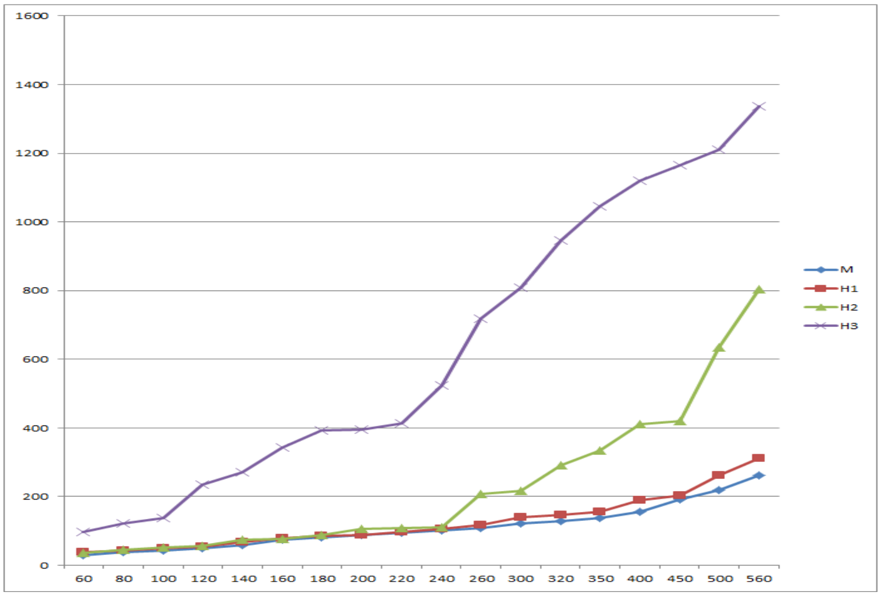

Figure 6: Operating curves with four scheduling algorithms of different quality without setup times on Instance 2.

Table 2: Mean cycle times obtained on three industrial instances without setup times.

	Instance 1				**Instance 2**				**Instance 3**			
WIP	M	H1	H2	H3	M	H1	H2	H3	M	H1	H2	H3
60	36	37	37	101	30	39	37	98	44	48	47	65
80	45	46	47	144	38	43	46	123	59	52	58	71
100	52	55	55	161	44	49	52	137	66	70	79	90
120	66	67	67	168	51	55	56	234	67	72	84	101
140	71	72	79	314	60	69	75	271	75	85	92	111
160	76	79	91	319	75	79	77	343	88	98	101	147
180	91	91	95	344	82	87	89	393	98	116	131	166
200	94	96	99	352	88	88	107	395	100	111	128	202
220	103	110	111	455	95	97	109	413	113	126	147	254
240	112	119	119	485	101	106	110	524	134	149	164	311
260	120	130	244	521	109	118	208	717	147	158	179	388
300	138	161	268	633	123	141	216	809	155	175	233	407
320	144	162	285	679	129	146	291	946	165	189	272	492
350	156	169	307	749	138	155	335	1,046	170	215	304	513
400	183	195	357	1,056	155	190	411	1,119	219	312	388	696
450	229	229	360	1,293	193	203	421	1,166	289	395	461	866
500	267	284	761	1,320	220	261	634	1,211	320	435	508	965
560	276	298	773	1,389	262	312	803	1,336	355	522	688	1,228

Table 2 details the results on three different instances. It is is interesting to see that, depending on the instance, the impact of reducing the complexity of the scheduling problem differs. However, overall, the production capacity increases since average cycle times are lower than in Table 1 and the differences between the scheduling algorithms have decreased.

5 CONCLUSION

In this paper, experimental tests were conducted on real industrial data to analyze the importance of using an effective algorithm when scheduling lots in a complex workstation of a semiconductor manufacturing facility. The main conclusions are that:

- An effective optimization algorithm helps to reduce cycle times of lots, and thus to increase the workstation capacity,
- Simplifying the problem by ignoring setup times reduces the impact of the quality of the scheduling algorithm,
- The impact of the scheduling algorithm can change significantly from one instance to another.

An extension of our work could be to consider more industrial instances to determine whether an algorithm is sensitive to a specific type of additional constraints. It could then be interesting to classify these instances according to the type of constraints that affects the operating curve the most.

For future research, to analyze the impact of our scheduling algorithm compared to dispatching rules on factory cycle times, it would be interesting to evaluate its performance in a full factory model that integrates stochastic aspects such as machine breakdowns. The computing time of schedules is also an important factor in this analysis. Experiments to study the trade-off between solution quality and computing time have yet to be made.

ACKNOWLEDGMENTS

This work is supported by the ENIAC European Project INTEGRATE. We thank all contributors from STMicroelectronics who helped us in our work with valuable information and discussions.

REFERENCES

Bitar, A., S. Dauzère-Pérès, C. Yugma, and R. Roussel. 2014. "A Memetic Algorithm to Solve an Unrelated Parallel Machine Scheduling Problem with Auxiliary Resources in Semiconductor Manufacturing". *Journal of Scheduling*. To appear.

Chiou, C.-W., and W. Muh-Cherng. 2014. "Scheduling of multiple in-line steppers for semiconductor wafer fabs". *International Journal of Systems Science* 45 (3): 384–398.

Dauzère-Pérès, S., and J.-B. Lasserre. 2002. "On the importance of sequencing decisions in production planning and scheduling". *International Transactions in Operational Research* 9 (6): 779–793.

Diaz, S., J. W. Fowler, M. E. Pfund, G. T. Mackulak, and M. Hickie. 2005. "Evaluating the Impacts of Reticle Requirements in Semiconductor Wafer Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 18 (4): 622–632.

Gupta, J., R. Ruiz, J. W. Fowler, and S. J. Mason. 2006. "Operational planning and control of semiconductor wafer production". *Production Planning and Control* 17 (7): 639–647.

Johnzén, C., S. Dauzère-Pérès, and P. Vialletelle. 2011. "Flexibility Measures for Qualification Management in Wafer Fabs". *Production Planning and Control* 22 (1): 81–90.

Karmarkar, U. S. 1987. "Lot sizes, lead times and in-process inventories". *Management Science* 33 (3): 409–418.

Kock, A. A. A., C. P. L. Veeger, L. F. P. Etman, B. Lemmen, and J. E. Rooda. 2011. "Lumped parameter modeling of the litho cell". *Production Planning and Control* 22 (1): 41–49.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011. "Scheduling Semiconductor Manufacturing operations: Problems, Solution Techniques, and Future Challenges". *Journal of Scheduling* 14 (6): 583–599.

Morrison, J. R. 2011. "Multiclass flow line models of semiconductor manufacturing equipment for fab-level simulation". *IEEE Transactions on Automation Science and Engineering* 8 (1): 81–94.

Tirkel, I. 2013. "The effectiveness of variability reduction in decreasing wafer fabrication cycle time". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3796–3805. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yan, B., H. Y. Chen, P. B. Luh, S. Wang, and J. Chang. 2011. "Optimization-based litho machine scheduling with multiple reticles and setups". In *Proceedings of the 2011 IEEE Conference on Automation Science and Engineering*, 114–119. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ABDOUL BITAR is a PhD student at the *École Nationale Supérieure des Mines de Saint-Etienne* (EMSE) in France since October 2012. He obtained his master degree at Paris VI University (France), in Artificial Intelligence, Operational Research and Decision. His email address is bitar@emse.fr.

STÉPHANE DAUZÈRE-PÉRÈS is Professor at the Center of Microelectronics in Provence (CMP) of the EMSE. 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 where he headed the team Production and Logistic Systems (about 20 members) between 1999 and 2004. He was invited Professor at the Norwegian

School of Economics and Business Administration, Bergen, Norway, in 1999. Since March 2004, he is Professor at the Ecole des Mines de Saint-Etienne, where he headed the research department Manufacturing Sciences and Logistics (SFL, about 20 members) from 2004 to 2013 and the CMP from 2013 to 2014. 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 more than 50 papers in international journals and contributed to more than 120 communications in conferences. Stéphane Dauzère-Pérès has coordinated multiple academic and industrial research projects, and also five conferences. His email address is stephane.dauzere-peres@emse.fr.

CLAUDE YUGMA is Associate Professor at the Center of Microelectronics in Provence (CMP) of the EMSE. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003. He was a Postdoctoral Researcher at the Ecole Nationale Supérieure de Génie Industriel, Grenoble, from 2003 to 2004 and from 2005 to 2006 at the CMP of the EMSE. His research interests broadly include modeling and optimization of operations at short and tactical decision levels with a special emphasis on semiconductor manufacturing. His email address is claude.yugma@emse.fr.