Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

ALLOCATING RETICLES IN AN AUTOMATED STOCKER FOR SEMICONDUCTOR MANUFACTURING FACILITY

Anna Benzoni
Claude Yugma

Mines Saint-Étienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
Department of Manufacturing Sciences and Logistics
880 route de Mimet,
F-13541 Gardanne, FRANCE

Pierre Bect
Alain Planchais

Department of Industrial Engineering
Department of Production Control
STMicroelectronics Rousset
190 Avenues Coq Monnet
F-13106 Rousset, FRANCE

ABSTRACT

This article addresses the problem of reticle allocation in a stocker of an existing photolithography workshop of a 200 mm semiconductor wafer manufacturing facility. A reticle stocker generally consists of two internal storage zones: the retpod, where reticles are stored with pods and have short retrieval times, and the carousel, a bare reticle stocker with longer retrieval time. The reticle is an auxiliary resource in photolithography workshop operations. Thus, if the right reticles are not stored in the right places in the reticle stocker, it can quickly become a bottleneck. The purpose of the article is to determine the right reticles to store inside the right place of the reticle stocker. This is a knapsack problem. Three heuristics considering the arrival of lots in the upstream steps of the photolithography workshop are proposed and tested on real instances.

1 INTRODUCTION

Semiconductor wafer manufacturing process can be described as a multi-step process with re-entrant flows (Monch et al. 2011). The processing is done layer by layer. Each layer requires several processing steps such as chemical deposition, etching and diffusion where the wafers are grouped in lots in order to undergo operations. A lot contains a maximum of 25 wafers (Monch et al. 2012). A lot moves through the semiconductor wafer fabrication (fab) plant for operations. Wafers are made up layer by layer and each lot can thus pass up to 40 times through the photolithography workshop. The photolithography process is one of the central operations in the production of wafers. Here, the process consists in transferring a diagram of an integrated circuit onto a wafer, on which a photosensitive resin has been applied, through exposure to ultraviolet radiation. To perform this operation, an auxiliary resource called reticle or mask (with an integrated circuit diagram) foments the pattern on the wafer (Monch et al. 2011). A reticle is associated with an operation of a given lot or product. The reticles are stored in a reticle stocker with limited capacity.

In a high-mix low-volume manufacturing environment, there are thousands of reticles, as one reticle is used for one process step and one product at a time. When a reticle is required, it is retrieved from the stocker and transported to the tool called stepper for processing the corresponding lot operation. Before being transported, the reticle is placed in a container, called pod to protect it from contamination.

In the 200-mm site of STMicroelectronics in Rousset the transport of lots and reticles have been upgraded from manual-based to an automated transport system. Hence, lots and reticles are transported by an OHT (overhead hoist transport) system from stockers to tools and vice versa, see (Ben-Salem et al. 2017). The photolithography area has more than 4000 reticles and three automatic reticle stockers. To give an idea of the complexity management, daily, an average of 1600 reticles is used on the stepper tools for lot processing. This generates complexity in determining to which stocker send the reticle and then in which compartment of the storage location to put it. In Figure 1, the reticle stocker is shown. Each stocker has 2 automated ports, where reticles are stored and retrieved by OHT vehicles (a load port dedicated to the entry and another one dedicated to retrieving the reticle). If one port is busy, the other port must wait for the adjacent port to complete its task before starting any other operation. The handling time of the reticle in the stocker (enter and remove) depends on reticle storage location inside the stocker: pod shelf (Figure 1, b) is dedicated to store empty and full pods, comprising approximately 60 places, including 20 places strictly reserved for storing pods with reticles. The time required to remove and enter a reticle in this compartment, called retpod, is approximately 15 seconds. In the second compartment, called carousel, which contains 1,500 places (Figure 1, c), the reticles are stored without pods. When a reticle is in the carousel area, as the reticle has to be coupled with a pod before exiting the stocker (and decoupled from the pod when entering), the storage and retrieval time is longer and it is about 1 minute. Consequently, depending on the

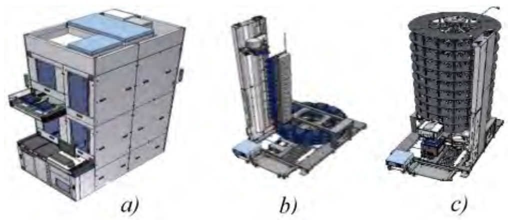

Figure 1: a) External overview; b) POD Fixed Shelf and POD Carousel Shelf; c) Carousel reticles

storage compartment inside the stocker, the delay for storing or retrieving a reticle can considerably vary. This leads to an increase in the processing time of the stocker. As a result, the throughput of the stocker can slow down and become the bottleneck of the automated material handling system. Thus, depending on the stocker capacity and the use of the reticles, it is important to determine the right reticles to store in the retpod. More precisely, the interest is to maximize the number of missions (a mission is the task consisting in storing or retrieving a reticle in the stocker) in retpod compartment compared to the missions in the carousel compartment to decrease stocker processing time. In other words, the right reticles performing the greatest number of missions must be chosen to be stored in the retpod compartment. For example, it is relevant to store a reticle with many wafers to process in the future (in a given period) in the retpod compartment, as it is supposed to be retrieved soon. Besides, an analysis of the current situation of reticle stocker shows that only 8% (low value) of missions are coming from the retpod compartment. This is due to poor management of reticles meaning that the right reticles are not placed in the right compartment of the stocker. The current management of reticles must be improved by increasing the number of missions in the retpod compartment and by avoiding that the stocker becomes bottleneck. The purpose of this article is to efficiently manage the stocker by determining the reticles to be assigned to the retpod.

The remainder of this article is organized as follows: In Section 2, we present related works on managing reticle storage. The reticle allocating problem is formalized and a solution approach is provided in Section 3. The experimental results are shown in Section 4 and finally, some conclusions and future perspectives are drawn in Section 5.

2 RELATED WORKS

Ben-Salem et al. (2016) addressed the problem of designing an AMHS for reticles in a photolithography area. An agent-based simulation model is implemented to determine the reticle storage capacity, the size of the vehicle fleet, and the vehicle dispatching rules. What-if scenarios are used to determine the storage capacity of the reticles. Ndiaye et al. (2016) studied the management of an Automated Guided Vehicle (AGV) for the reticles in cohabitation with Overhead Hoist Transport (OHT) system for the transport of lots. Through the use of a discrete-event simulation model, different layouts, dispatching rules and sizing of the fleet of vehicles are tested while focusing on some key performance indicators as the travel time, throughput, and vehicle utilization. The objective is to determine the right number of vehicles, with the best dispatching policies. Murray and Miller (2003) studied two rules for reticle transport and two storage strategies: centralized and distributed strategies. The authors show that the distributed strategy allows having shorter reticle delivery times. Yamagishi (2003) proposes one of the first OHT-based automated transport for reticles. The study shows that OHT considerably improves the production performances of photolithography. Also, they study a relevant problem: the reticle stocker is divided into two different stages, where the most frequently used reticles are stored in a defined stage, and the other part of the storage is connected via an elevator. Tamehiro (2005) proposes a "push" approach management coupled with a reticle dispatch algorithm to solve the bottleneck problem at the load ports of a bare reticle stocker. In their study, the bare reticle stocker and the pod stocker are divided. The idea is to anticipate reticle requests when it is required on tool equipment, by predicting the wafer lot dispatching. But in this work, no details are given on the times linked to internal reticle compartments. It appears from the literature that most of the works does not specifically deal with reticle storage behavior, where the access times to different compartments are taken into account to determine which reticles to place in the different compartments of the stocker.

We see that the reticles storage problem considered is close to the storage problem encountered in warehouses. Therefore, similar research can be encountered in warehouse management problems. Warehouses often have two areas: the front area, with limited capacity, and where items can be easily picked up; a reserve area to accommodate bulk storage and this does not allow easy access. The goal is to replenish the front area and store the items in the other area. The problem of determining the items to be assigned to the front zone to balance replenishment and preparation has been studied in the literature. Hackmann et al. (1990) develop a heuristic procedure to decide which items to allocate to automated storage/retrieval and the quantity to be stored to minimize the costs of picking and replenishment. Gu et al. (2010) proposes a Branch and Bound algorithm to solve the problem. One can also find different variants of the problem proposed by van den Berg et al. (1998), Heragu et al. (2005) and Walter et al. (2013). A notable difference between warehouse work and ours is that items leaving the warehouses do not come back, whereas in this study reticles can come back into the stocker.

We aim to determine the right reticles to store in the retpod compartment, while taking into account the arrival of lots requiring the reticles, to maximize the number of missions (tasks of storing or retrieving a reticle) in the retpod compartment.

3 PROPOSED APPROACH

We tackle a problem where lots arrive dynamically in photolithography and we must decide which reticles to consider in the retpod to maximize the number of missions (retrieve or store reticles). We write $\mathcal{R} = \{R_1 \dots R_n\}$ the set of reticles that are outside the stocker. Each reticle can occupy one place located in the retpod compartment if it is stored.

The static version of the problem can be seen as a set of reticles where each one has a weight of 1 (1 location) and a profit value (corresponding to the number of lots to process with the considered reticle), and the objective is to determine the reticles to store in the retpod so that the total weight is less than or equal to the retpod capacity (C the total number of available places) and the total profit is maximized. This problem is a well-known problem called Knapsack problem (Assi and Haraty 2018).

Due to the dynamic of the fab (vehicles congestion, down tools, etc.), we face a high uncertainty about the arrival of reticles at the entrance of the stocker. Then, it is complicated to predict with enough confidence the arrival time of reticles in the stocker. For this reason, we have decided not to consider reticle arrival times. Each of the reticle R_i has an incoming activity $L_{i,k}$ which corresponds to the number of processed lots with the reticle R_i while considering lots that are from operation (current operation of Photolithography 1) to the (k upstream operations). In other words, k refers to the operation step number before the photolithography step (which is 1). This gives the upstream position of the lot related to the photolithography. Hence, depending on the position of the lot (step number k) the reticle can wait for a

longer or a shorter time.
To evaluate the reticle waiting time we define two factors:

1. Lot priority $P_{i,k}$. This is the maximum priority at operation k for lots associated with reticle R_i.
2. Lot flexibility index $F_{i,k}$. This is a ratio corresponding to the number of qualified machines executing the recipe (these are all the settings required on a machine to operate) linked to the reticle R_i over total authorized machines. In other words, it is a ratio that gives the percentage of machines for each operation k available for processing lots, associated with a given reticle R_i.

Knapsack problem is known as an NP-hard problem (Wilbaut and Hanafi 2008). A well-known heuristic for the knapsack problem resolution is to rank the items from the highest to the lowest value according to the ratio *"profit/weight"* and assign the items to knapsack until the capacity is filled (for more details see Hackmann et al. (1990)).

To address the problem, we have derived from the same algorithm from Hackmann et al. (1990) three heuristics to evaluate:

1. Heuristic 1 (H1). The reticles are classified according to the objective function divided by the place occupied by the reticle in the reticle zone.

$$\sum_{k=1}^{k=K} \frac{(P_{i,k} \times F_{i,k}/k)}{1} \quad \forall R_i$$

   The value 1 indicates that one place is occupied by a reticle (It is the cost when considering the well-known Knapsck heuristic).
2. Heuristic 2 (H2). It is exactly the same objective function proposed, but the number of lots $L_{i,k}$ allocated to the reticle at each operation is considered.

$$\sum_{k=1}^{k=K} \frac{(P_{i,k} \times F_{i,k} \times L_{i,k}/k)}{1} \quad \forall R_i$$

3. Heuristic 3 (H3). The third heuristic is a slight variation of the knapsack heuristic. In this version only the closest photolithography previous operation with $L_{i,k} > 0$ is considered and the weight of the reticle corresponds to the operation k in which the lots are attending is considered. This means that the further the previous operation considered is from photolithography, the greater the weight of the reticle is.

$$\frac{P_{i,k} \times F_{i,k} \times 1}{k} \textit{if } k \textit{ minimum index with } L_{i,k} > 0 \; \forall R_i$$

The objective function is obtained by the combination of the flexible index for reticle R_i at operation k and the highest priority between lots assigned to reticle i at operation k: operation k gives a penalty to reticle gain, depending on the “distance” between the operation and the photolithography step. To solve the problem while integrating the arrival of lots, a rolling horizon of T is considered. At each iteration of T, an algorithm is refreshed and the reticles currently outside the stocker are taken into account.

4 EXPERIMENTAL TESTS

We carried out experimental tests on the proposed heuristic using real data provided by the company. We briefly describe the data and analyze the results.

4.1 Data analysis

An analysis of the extracted data showed that between all the external reticles, an average of 34% could be considered for the storage assignments of retpod. The rest were unable to be considered for quality reasons or because they did not return to stocker in the reporting period. Due to the uncertainty of certain process operations, it is necessary to define the retpod capacity C for each period T, so that the available retpod locations are not exceeded. In addition, the capacity of the retpod can be adjusted by applying Little's law (Little and Graves 2008). We note that if we assign to the retpod compartment 25% of reticles better classified according to heuristics, between the outside reticles (see section 4.2), the average residence time in stocker is between 5 and 5.30 hours and 92% of occupied places is reached. We decided to take the capacity of retpod 25% of the number of the reticles outside from stocker. With this capacity, we expect to fill up retpod zone.

4.2 Heuristics results

The three heuristics H_1, H_2 and H_3 were tested on data extracted from different days of non-consecutive weeks of production (same number of lots produced per week). There are approximately 13,000 lots per week.

The rolling horizon T is fixed at 30 minutes. This time is considered because it represents the average duration for a reticle to process a lot on the stepper. Furthermore, in the algorithm, we look at the state of the upstream steps of the photolithography. We have chosen to go back until the fourth $k = 4$ upstream operations. It is not necessary to consider further operations because we would face too much uncertainty about the arrival of lots. Also, we cannot take into account lots currently being processed in photolithography because we do not have information on the state in which lots are (currently on the machine, awaiting processing, already processed, etc.).

Four different indicators have been defined to measure the performances of the versions of algorithms:

1. Percentage of retpod missions,
2. Average reticle residence time in the retpod,
3. Surplus of occupied retpod places after one day of ramp up (compared with the available ones),
4. Algorithm error: Reticles stocked in retpod, but with a residence time higher than 5.3 hours.

Table 1: Results of the 3 heuristics on different production weeks

	Wk 1 (Nov 2019)			Wk 2 (Dec 2019)			Wk 3 (Jan 2020)			Wk 4 (Feb 2020)		
INDICATORS	H1	H2	H3	H1	H2	H3	H1	H2	H3	H1	H2	H3
Missions (%)	21%	21%	22%	21%	21%	21%	21%	20%	22%	18%	17%	21%
Reticles RT(h)	4:05	**4:04**	4:29	4:24	**4:19**	4:37	5:21	**5:18**	5:39	5:29	**5:11**	5:53
Error (%)	22%	**21%**	23%	22%	**21%**	23%	31%	**31%**	32%	32%	**30%**	34%
Surplus places (%)	7%	**0%**	10%	12%	**7%**	23%	22%	**12%**	28%	2%	**0%**	23%

Table 1 summarizes the results. We notice that all algorithms return almost the same percentage of retpod missions. We see some differences in the other indicators: the heuristic H_3 involves a longer average reticles residence time in the stocker, with the consequence that too many available places would be necessary for the retpod compartment (the maximum occupied places of the retpod are calculated after a transitional day). Indeed, the heuristic error is much higher than the error obtained with the other considered heuristics. The other two algorithms have fairly similar values, even if the second, maintaining a good average of retpod missions, provides a minor error and requires only a slight increase in the places available for the retpod. The most notable differences are between the different periods studied: if the error in the second two weeks is around 30%, in the first weeks, the error is around 20%. The results obtained are not so surprising: in the last weeks we have noted a high value of hold lots. Lots can be kept hold for reasons of non-quality specifications. When many lots are in hold condition, the proposed algorithm will present a more emphasized inaccuracy. The results obtained show that our algorithm can greatly improve the management of the reticles by reducing their processing times: addressing much more reticles with a short inactivity time (residence time) to the retpod compartment, we can see that the average missions time decreases. The proposed algorithm can make the percentage of retpod missions increasing from 8% to 20%: this great improvement means a reduction in average missions time of about 6 seconds, with a consequent reduction of one stocker processing time on the day duration of more than 1 hour.

However, for obtaining the entire gain, a slightly increase in the number of retpod places in each stocker is required to absorb the retpod surplus places highlighted in the results of experimental tests (see Table 1).

5 CONCLUSIONS AND POTENTIAL RESEARCH DIRECTIONS

The paper addresses the problem of the allocation of reticles inside the storage compartment in a 200 mm semiconductor wafer manufacturing facility. A reticle stocker consists generally of two internal storage zones: the retpod with short retrieval time and the carousel with longer retrieval time. Not assigning the right reticles to the right compartments can lead to bottleneck issues. We have studied the allocation of reticles inside the stocker (retpod), taking into account the continuous arrival of lots. The problem can be modeled as an adaptation of the classic Knapsack problem. We have proposed three different policies to determine the allocation of the reticle in the stocker. The experimental tests carried out on real data show a significant improvement in the management of the stockers in terms of storage or retrieval of reticles in the retpod compartment.

A next step of the study is to improve the algorithms by considering reticle arrival times at the entrance of the stocker: to do so, it will be necessary an in-depth analysis of industrial instances on reticle processing time and reticle transport time from equipment to stocker. Another point to explore is to couple the internal management of the reticle stocker with vehicle management. Indeed, when a reticle is ready to be retrieved from the stocker loading port, an event is submitted to the AMHS controller with the request to send the vehicle to retrieve the reticle. During this waiting time for unloading the reticle, the stocker is not authorized to perform another task. Therefore, anticipating the vehicle travelling to the stocker is relevant to be considered.

ACKNOWLEDGMENTS

This study has been done within the framework of a joint collaboration of STMicroelectronics in Rousset, France, and the center of Microelectronics in Provence (CMP) of the Ecole des Mines de Saint-Etienne (EMSE) in Gardanne France. The authors would like to thanks the ANRT (Association Nationale de la recherche et de la Technologie) which has supported this study.

REFERENCES

Assi, M., and R. A. Haraty. 2018. “A Survey of the Knapsack Problem”. In *IEEE International Arab Conference on Information Technology (ACIT 2018)*. Werdanye, Lebanon.

Ben-Salem, A., C. Yugma, J. Pinaton, and E. Troncet. 2016, May. “AMHS design for reticles in photolithography area of an existing wafer fab”. In *Proceedings of 2016 Advanced Semiconductor Manufacturing Conference*, 110–115. Saratoga (USA).

Ben-Salem, A., C. Yugma, J. Pinaton, and E. Troncet. 2017. “A simulation-based approach for an effective amhs design in a legacy semiconductor manufacturing facility”. In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. Chan, A.D.Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E.Page, 3600–3611: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gu, J., M. Goetschalckx, and L. McGinnis. 2010. “Solving the forward-reserve allocation problem in warehouse order picking systems”. *Journal of the Operational Research Society* 61(6):1013–1021.

Hackmann, S. T., M. J. Rosenblatt, and J. M. Olin. 1990. “Allocating Items to an Automated Storage and Retrieval System”. *IIE Transactions* 22(1):7–14.

Heragu, S. S., L. Du, R. J. Mantel, and P. C. Schuur. 2005. “Mathematical model for warehouse design and product allocation”. *International Journal of Production Research* 43(2):327–338.

Little, J. D., and S. C. Graves. 2008. *Building Intuition: Insights from Basic Operations Management Models and Principles*, Chapter 5, 81–100. Springer Science + Business Media LLC, New york.

Monch, L., J. W. Fowler, S. Dauzère-Pérès, and O. Rose. 2011. “A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations”. *Journal of Scheduling* 14:583–599.

Monch, L., J. W. Fowler, and S. J. Mason. 2012. *Production Planning and Control of Semiconductor Wafer Fabrication Facilities: Modeling, Analysis and Systems*. Springer Science & Business Media.

Murray, A. M., and D. J. Miller. 2003. "Automated reticle handling: a comparison of distributed and centralized reticle storage and transport". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, 1360–1365: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ndiaye, M., S. Dauzère-Pérès, C. Yugma, L. Rullière, and G. Lamiable. 2016. "Automated transportation of auxiliary resources in a semiconductor manufacturing facility". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2587–2597: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Tamehiro, R. 2005. "New job classes for fully automated reticle and sorter dispatch". *IEEE International Symposium on Semiconductor Manufacturing*:59–62.

van den Berg, J., G. Sharp, A. Gademann, and Y. Pochet. 1998. "Forward-reserve allocation in a warehouse with unit-load replenishments". *European Journal of Operational Research* 111:98–113.

Walter, R., N. Boysen, and A. Scholl. 2013. "The discrete forward–reserve problem – Allocating space, selecting products, and area sizing in forward order picking". *European Journal of Operational Research* 229(3):585–594.

Wilbaut, C., and S. Hanafi. 2008. "A survey of effective heuristics and their application to a variety of knapsack problems". *IMA Journal of Management Mathematics* 19(3):227–244.

Yamagishi, T. 2003. "The World's First Automated Reticle Handling System Using OHT". *IEEE International Symposium on Semiconductor Manufacturing*:21–24.

AUTHOR BIOGRAPHIES

ANNA BENZONI is a Ph.D. student at Ecole Nationale des Mines de Saint-Etienne, France. She works as an Engineer in the Industrial Engineering Department at STMicroelectronics Rousset. She received the Mechanical Engineering master degree from Politecnico di Milano, Italy in 2017. Her Ph.D. subject is on Modeling and simulation of a heterogeneous vehicle in an Automated Material Handling Systems in semiconductor manufacturing. Her email address is anna.benzoni@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 H.D.R. 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 and scheduling and simulation in semiconductor manufacturing. His email address is claude.yugma@emse.fr.

PIERRE BECT is a an Industrial engineer at STMicroelectronics Rousset and he earned a Ph.D degree in computer science at the Grenoble institute of Technology in collaboration with Airbus Helicopter Marignane, France in 2013. Most of his work deals with data analysis and optimization algorithms development. His email address is pierre.bect@st.com.

ALAIN PLANCHAIS is senior project & program manager at STMicroelectronics Rousset, France. He earned a master degree in engineering of Electronics and Automation. With more than 35 years experience in semiconductor industry, he has gained various competences in international collaboration, innovation management, automation, Robotic and Automated Material Handling Systems. His email address is alain.planchais@st.com.