Int. J. Production Economics 160 (2015) 213–219

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

A mathematical programming approach for optimizing control plans in semiconductor manufacturing

Justin Nduhura Munga [a,b,*], Stéphane Dauzère-Pérès [a,1], Claude Yugma [a,2], Philippe Vialletelle [b,3]

[a] Ecole des Mines de Saint-Etienne – Centre of Microelectronics in Provence, Department of Manufacturing Sciences and Logistics 880 avenue de Mimet, F-13541 Gardanne, France

[b] STMicroelectronics – Centre Commun de Microélectronique de Crolles, Department of Industrial Engineering 850 rue Jean Monnet, F-38926 Crolles, France

ARTICLE INFO

Article history:
Received 14 May 2012
Accepted 4 November 2014
Available online 13 November 2014

Keywords:
Control plan
Semiconductor manufacturing
Dynamic sampling
Risk
Defectivity
Integer linear programming

ABSTRACT

In a globally competitive environment, sustaining high yield with a minimum number of quality controls is key for manufacturing plants to remain competitive. In modern semiconductor manufacturing facilities, with the moves to ever smaller geometries and the variety among products to be run concurrently, designing efficient control plans is becoming increasingly complex. Since a 100% of inspection is neither feasible nor interesting because of the cost and reliability of each control, dynamically identifying the right product to inspect is one of the keys to achieve high yield and reduce the cycle time. However, when control parameters are over- or under-estimated, a dynamic sampling static sampling strategy can lead to poor results. In this paper we propose an integer linear programming approach to optimize the use of inspection capacity through dynamic sampling. The goal is to determine two key parameters (called warning limit and inhibit limit) that are related to the resulting level of risk and the available inspection capacity. The model has been implemented on a commercial solver and tested using actual industrial data. Results show that the overall risk can be strongly reduced without any additional capacity.

1. Introduction

Semiconductor manufacturing is characterized by up to 700 processing steps resulting in a significant cycle time of more than two months. This significant cycle time, combined with the reduction in device sizes, leads to a high complexity when designing control plans for sustaining high yield (Vits et al., 2006). On one hand, a control plan aims at identifying defective products and yield detractors throughout production (Bassetto and Siadat, 2009). On the other hand, a control operation is considered as a non-added value (Bunday et al., 2007) and, therefore, each time a control operation is performed on a lot, the cycle time of the lot is increased with consequences on the final product costs. To stay competitive, companies have to provide pricing power against competitors. This implies that companies have to be able to sustain high yield with a minimum number of control operations.

Several works have been conducted on sampling techniques in order to minimize the number of controls without increasing the risk on the production (Boussetta and Cross, 2005) (Purdy, 2007). Compared to static sampling, dynamic sampling techniques are more suitable for modern semiconductor plants (Mouli and Scott, 2007). This is particularly true in high-mix semiconductor plants where more than 200 products are run concurrently. However, the complexity is such that it is not always easy or feasible to implement a dynamic sampling strategy because of the factory specificities, production constraints, or IT infrastructure. Depending on the production environment, the efficiency of a dynamic sampling policy will directly be linked to input parameters. If these input parameters are not optimally set, wrong sampling decisions can lead to very poor results. This is the case when some parameters such as process or product criticality are determined mainly based on the experience of engineers. This is also the case when some expected levels of risks are directly set by production managers without considering the effective and available capacity for

* Corresponding author at: Ecole des Mines de Saint-Etienne – Centre of Microelectronics in Provence, Department of Manufacturing Sciences and Logistics 880 avenue de Mimet, F-13541 Gardanne, France. Tel.: +33 4 7692 6065.
E-mail addresses: justin.nduhura27@gmail.com, nduhura@emse.fr (J. Nduhura Munga), dauzere-peres@emse.fr (S. Dauzère-Pérès), yugma@emse.fr (C. Yugma), philippe.vialletelle@st.com (P. Vialletelle).
[1] Tel. +33 4 4261 6660.
[2] Tel. +33 4 4261 6664.
[3] Tel. +33 4 3892 2290.

http://dx.doi.org/10.1016/j.ijpe.2014.11.004
0925-5273/

inspection. Consequences are the inability for dynamic sampling to take relevant decisions regarding the lot or product to inspect. Some lots may be prioritized because of the product criticality but never inspected because of the lack of inspection capacity.

In this paper, we propose an integer linear programming (ILP) approach to optimize the use of inspection tools through dynamic sampling. The approach consists of determining two key parameters that are related to the resulting levels of risk and the available inspection capacity. These two key parameters are called warning limit (WL) and inhibit limit (IL) and were introduced in (Dauzère-Pérès et al., 2010) for dynamically sampling lots in front of inspection steps. WL corresponds to the limit above which the situation starts to become critical in term of control, and IL is the limit above which the production should be stopped if a control operation is not performed. In terms of wafers, IL represents the maximum number of wafers that can be run on a production tool between two control operations, and WL is an alarm taking into account the cycle time between a production tool and the next inspection tool. The study has been performed for the defectivity area, where control operations are done on wafers in order to monitor the contamination of production tools. Hence, all production tools in the different processing stages of semiconductor manufacturing (etching, implantation, photolithography, chemical mechanical polishing (CMP), etc.) are controlled in defectivity. Results indicate that WL and IL values drive the sampling policy. If the WL and IL values are underestimated, a significant number of lots will be sampled since the objective is to stay below IL. However, because of the lack of inspection capacity, not all sampled lots will be inspected. Consequences are increased cycle times for the sampled lots that are stopped at the inspection step for "nothing". Similarly, if the WL and IL values are overestimated, the number of sampled lots will be reduced but the risk will be significant, and the inspection capacity will be wasted due to inspection of lots without added value. By using the WL and IL values obtained with our ILP model, we show that an overall risk reduction can be obtained without additional inspection capacity.

This paper is structured as follows. Section 2 summarizes a literature review of sampling techniques in semiconductor manufacturing. The problem is described in Section 3. Section 4 presents the Integer Linear Programming (ILP) model we propose to compute the WL and IL values. Numerical experiments performed on actual data from company are presented in Section 5. We discuss the impact of the WL and IL values on a dynamic sampling policy, and assess the efficiency of values obtained with our ILP model. Section 6 concludes the paper and provides avenues for further research.

2. Literature review

Sample measurement inspecting for a process parameter is a necessity in semiconductor manufacturing, because of the prohibitive amount of time involved in 100% inspection while maintaining sensitivity to all types of defects and abnormality (Su et al., 2008). Moreover, a 100% inspection does not ensure 100% quality since, in semiconductor manufacturing, the inspection is never totally reliable and can easily introduce an error of almost the same order as the fraction of defectives (Pesotchinsky, 1987; Chien et al., 2007). Clearly, the development of sampling techniques is not recent in semiconductor manufacturing. However, much progress has been observed from static to dynamic through adaptive sampling techniques. A detailed review on sampling strategies and techniques in semiconductor manufacturing can be found in Nduhura-Munga et al. (2013). Some of the key references are discussed below.

Static sampling techniques consist of always selecting a fixed number of lots to inspect at different stages of manufacturing. This technique, widely used in the 90's because of the ease of implementation, is no longer suitable for modern semiconductor plants. For instance, Tomlinson et al. (1997) describe a study performed in an IBM plant to determine the optimal sampling plan for the poly etch module. The goal is to minimize both the risk for the product and the cost of inspection. Optimized sampling techniques for overlay measurements are discussed in Chien et al. (2001), and are validated through simulations using historical data from fabs. A discussion on the strengths and weaknesses of various sampling techniques for critical dimension (CD) measurement can be found in Elliott et al. (1999). Because static sampling means inspecting the same lots at all control operations, there is a problem of coverage in term of risk, and the factory dynamics is not taken into account (Venkateswaran and Son, 2006). Moreover, for the considered lots, there is a strong impact on cycle time and an increased risk of yield losses due to a larger number of steps to be performed and the significant time spent in front of each inspection step (Wang et al., 2006). The transition from static to adaptive sampling started in the second part of the 1990s. An example of an industrial deployment can be found in Williams et al. (1999). Adaptive sampling techniques are an evolution of static sampling techniques. Sampling rules are adjusted throughout production depending on the production state (Boussetta and Cross, 2005; Ho and Trindade, 2009). The main difference between the two techniques is that, in adaptive sampling, the number of lots or wafers to select is adjusted throughout production depending on the process state, see for example (Shanthikumar, 2007). An example can be found in (Prabhu et al., 1994), where a combined adaptive X chart is introduced that uses real-time information available from the process to make the control scheme proactive. In Kuo et al. (1997), simulation studies are performed to test and validate an adaptive sampling approach. Although adaptive sampling is generally much more effective than static sampling, its main drawback is the management of resources. By modifying the number of lots to inspect throughout production, the workload at the inspection step is no longer the same (Dauzère-Pérès et al., 2010).

Dynamic or smart sampling techniques are the more recent techniques being developed for modern semiconductor plants. They consist of selecting in real time the best lot or wafer to inspect depending on the production state, the inspection capacity, and the factory dynamics. Several indicators are analyzed in real time and no static rule is defined at the start of the production. All lots can be inspected and the decision of inspecting or not a lot is directly taken in front of the inspection step based on the gain brought by the lot. Dynamic sampling techniques are more suitable for modern and high-mix semiconductor plants (Purdy, 2007), but the problem is that it is not always easy to assess their efficiency. Analyzing in real time several indicators may lead to an increased complexity depending on the production environment. First, the significant amount of data to manage can lead to infeasible solutions depending on the factory environment. Second, the types of data or parameters to handle can impact the efficiency of a sampling technique. This is the case when some input parameters are based on the experience of engineers and not on actual data. Third, these techniques are still recent and authors often do not specify how lots can be dynamically selected and how a given technique can be implemented in practice. (Purdy, 2007; Holfeld et al., 2007) are the first authors who worked on industrial deployments. However, they do not provide details on the efficiency and complexity of these deployments. Lin et al. (2010) point out three main benefits related to developing a dynamic and intelligent sampling system in semiconductor manufacturing: sampling stability, satisfactory coverage of in-line products, and comprehensive inclusion of process tools. Also, and although no industrial assessment is reported, Sun and Johnson (2008) propose a scoring algorithm based on weighted objectives to determine the optimal wafer sampling for maximum coverage.

Dauzère-Pérès et al. (2010) introduce a dynamic sampling algorithm that has been tested and validated with actual data from different semiconductor manufacturing facilities (Yugma et al., 2011).