Computers & Operations Research 53 (2015) 301–308

Contents lists available at ScienceDirect

Computers & Operations Research

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

Setting defect charts control limits to balance cycle time and yield for a tandem production line

Miri Gilenson [a], Michael Hassoun [b], Liron Yedidsion [a]

[a] Faculty of Industrial Engineering and Management, Technion, Haifa, Israel
[b] Department of Industrial Engineering and Management, Ariel University, Ariel, Israel

ARTICLE INFO

Available online 3 June 2014

Keywords:
Statistical Process Control
Yield
Cycle time
Semiconductors
Markov chains
Queuing theory

ABSTRACT

Control limits in use at metrology stations are traditionally set by Yield requirements. Since deviations from these limits usually trigger a machine's stoppage, the inspection design has a direct impact on the station's availability, and thus on the product cycle time (CT). In this research we formulate a trade-off between the expected values of the CT and the die Yield. Based on the impact of the inspection's control limits on both performance measures, we formulate the CT to Yield Pareto-optimal set for a tandem production line.

We consider a semiconductor production line in which production stations are afflicted by a defect deposition process and immediately followed by an inspection step. First we study the impact of the upper control limit on both expected values of Yield and CT on a single station. Then, we extend our result to a tandem production line and present an optimal greedy algorithm that provides the Pareto-optimal set of Upper Control Limit (UCL) values for the line. The obtained model enables decision makers to knowingly sacrifice Yield to shorten CT and vice versa.

1. Introduction

The semiconductor industry is characterized by numerous cutting edge production technologies making perfect quality virtually impossible, and a certain portion of produced devices inevitably fails the functional tests that conclude their fabrication process. The portion of functional devices at the end of the process, denoted as “die Yield”, is a crucial indicator of the technology health.

Another trait of this challenging industry is the very long production spans, or as it is called in the fab jargon “Cycle Time” (CT). The race for ever smaller, higher density devices from the same silicon wafer, and the periodic upgrades in devices' size drive the industry through very fast obsolescence cycles, making CT a crucial managerial performance measure. In this research we articulate one of the several trade-offs existing between CT and Yield, and find an optimal working point between them.

The connection binding Yield and CT can fall into two broad categories. The first one, considered among others by Wein [28] and Cunningham and Shanthikumar [11], implies that a prolonged stay in the line exposes wafers to more particle contamination, thus reducing their Yield. In the second case, the mechanism is just the opposite. The constant process monitoring necessary to ensure a required level of Yield takes its toll on the product's CT. Both the additional time spent by the wafers at metrology stations and the machine stoppages following an out-of-spec monitor slow down the work-in-process (WIP) flow, as it has been well described in Colledani [7], and Colledani and Tolio [8] and [9].

Traditionally, the semiconductor industry emphasizes quality over CT. Yield engineers usually set quality control requirements based on targeted Yield, basically leaving the industrial engineers to struggle for the best possible CT under these requirements. Memories have a much shorter product life cycle (about one and a half years) compared with CPUs (about three years). For such ephemeral products, the typical CT of three months represents a huge chunk of their short life. Accordingly, new CPU usually see their release postponed by as long as half a year after their qualification to stabilize their process and reach higher Yield levels, whereas manufacturers of memory products usually settle for low Yield in order to release their products earlier to the market. Following this practice, researchers have started to analyze the trade-off between CT and Yield.

Meyersdorf and Yang [24] as well as Khetan et al. [19] present some aspects of the Yield–CT trade-off without quantifying them. More recently, Tirkel et al. [27] proposed a dynamic monitoring policy instead of the traditional constant sampling. Such dynamic policies are not new, but were so far used to improve Yield [12]. In a similar vein, Goren and Rabinowitz [13] suggest a model for efficient integration of Yield and CT under a combined in-line inspection and repair policy. They suggest random inspection and repair times, as well as finite queues, while analyzing a queuing network model performance with the decision variable being the inspection rate.

http://dx.doi.org/10.1016/j.cor.2014.05.011
0305-0548/

In this research, we propose to study how one of these Yield-driven decisions, namely setting the value of the control limits for defect charts at the monitor, affects both the Yield and the CT of a product. By doing so we depart from the traditional approach that sets control limits based solely on quality needs, and measure how they also impact CT.

Setting control limits to optimize Yield and other parameters is not a new concept and has mostly been applied in economic design of control charts. Ho and Case [16] present a review of these economic models and techniques. To our knowledge, however, none of these models takes production speed into consideration.

The mechanism binding Yield and CT through the control limit values is a simple one: more stringent control limits will obviously improve quality at the expense of more frequent machine stoppages, thus impeding the WIP flow. To study the impact of the control limits on both Yield and CT, we first analyze their impact on each measure separately. In the framework of this paper we consider production steps rather than machines for each item. In order to analyze the trade-off between CT and Yield more precisely we use a somewhat simplistic model consisting of a tandem production line, rather than a reentrant job-shop which is typical of the semiconductor industry. The production stages are assumed to be independent of each other with regard to quality performance which allows for a straightforward calculation of the expected Yield. Furthermore, we assume here only one type of countable defect, namely particle contamination. Yet, the suggested model is generic enough to enable its application in a broad range of quality control scenarios.

In contrast to the Yield, determining the expected CT is a bit more complex since consecutive stations are linked to one another through WIP flow variability. We address this question with the help of a queuing network (QN) approximation.

Using QN to model semiconductor production systems is not a new idea. Chen et al. [6] was one of the first to use a QN model, rather than simulation, in this field. Hopp et al. [18] developed a capacity design tool for semiconductor facilities that makes the use of QN approximations and optimization routine. Connors et al. [10] present a performance evaluation model of semiconductors' systems based on an open QN.

Due to the dependencies between consecutive servers, a closed-form expression for CT in a tandem QN is usually out of reach. In this research, we adopt a well-known $G/G/1$ QN approximation (see [3] and [17]) to describe our system.

In Section 2 we analyze the impact of the control limits on Yield and CT for the single station case and present the trade-off between the two measures. Then, in Section 3, we consider a multi-station tandem line and extend our previous results by finding the Pareto-optimal set of control limit values that optimize the balance between Yield and CT for the whole production process.

2. Single station system

In the framework of this paper we model a tandem production line (see Li and Meerkov [22]). All items (wafers/lots) arrive at the first station and leave the line at the last station; therefore the system is an open QN. Stations are connected by exactly one input and one output. In this section, we consider one of the stations in isolation, model it, and study the impact of the control limits on both Yield and CT. We make use of a station index for all variables and parameters, since they are needed in later sections. As illustrated in Fig. 1, each production station in the network is followed by a metrology step in which the items are examined for defects through the use of SPC (Statistical Process Control) charts, and the decision whether or not to let the station continue producing is taken. The capacity of the metrology stations is infinite and there are no queues forming in front of them. Each station behaves as a single first-come-first-serve waiting line with a single server.

Tens to hundreds of microelectronic devices are built layer upon layer on the silicon base. Defects appearing at any stage of the process may, or may not, destroy the device's functionality. In our model, defects are device killers, independent of their exact location. However, the definition of a defect, regarding its size or any other characteristic, can be different at each station (in practice, certain operations are more sensitive than others).

2.1. The impact of the control limit on CT

In a tandem production line, the average item's (either a wafer or a lot) arrival rate is equal at all stations and we denote it as λ. The service duration (processing time) is a constant t_m. At the metrology station, part of each item's surface is sampled. Let us denote the sample area for station m by A_m; that is the proportion of sampled dice on the item. When the number of defects on the sampled area exceeds a predefined Upper Control Limit for station m (UCL_m), the station is said to be Out Of Control (OOC), and production is interrupted. Otherwise, the station is said to be In Control (IC). The all-target value for defects is obviously zero; therefore, we disregard here any type of lower control limit. Without any loss of generality, we assume the number of defects added to the sampled area of a specific item at process step m, denoted as x_m, to be a Poisson process with parameter μ_m. This assumption may correspond better with a particle contamination process, yet it is generic enough to be relevant to other types of defects as well. The station is described as a two-state station, and its defect deposition rate can either be low $\underline{\mu}_m$ or high $\overline{\mu}_m$: $\overline{\mu}_m > \underline{\mu}_m$. Although some machines witness a strong bimodal defect deposition process that would justify such an assumption, it is mostly adopted for simplicity reasons and should be released in future extensions of this work.

The probability for a monitor to exceed the control limit can be obtained by

$$P(OOC_m) = 1 - P(x_m \le UCL_m) = 1 - \sum_{k=0}^{UCL_m} \frac{(\mu_m)^k e^{-\mu_m}}{k!},$$

where $\mu_m \in \{\underline{\mu}_m, \overline{\mu}_m\}$. The inspection process is subject to errors. We denote by α_m the probability that a monitor exceeds UCL_m when the defect deposition rate is low (type 1 error), and by β_m the probability of a monitor to remain below the UCL_m when deposition rate is, in fact, high (type 2 error). We consider a sample to be IC if $x_m \le UCL_m$ and OOC otherwise. Accordingly

$$\alpha_m = P(x_m > UCL_m | \underline{\mu}_m) = \sum_{k=UCL_m+1}^{\infty} \frac{(\underline{\mu}_m)^k e^{-\underline{\mu}_m}}{k!} = 1 - \sum_{k=0}^{UCL_m} \frac{(\underline{\mu}_m)^k e^{-\underline{\mu}_m}}{k!}; \quad (1)$$

$$\beta_m = P(x \le UCL_m | \overline{\mu}_m) = \sum_{k=0}^{UCL_m} \frac{(\overline{\mu}_m)^k e^{-\overline{\mu}_m}}{k!}. \quad (2)$$

We model the evolution of a single station over time with four states (also depicted in Fig. 2):

1. The defect deposition rate is low ($\underline{\mu}_m$) and the monitor indicates that the process is *IC*.
2. The defect deposition rate is low ($\underline{\mu}_m$) and the monitor indicates that the process is *OOC* (type 1 error).
3. The defect deposition rate is high ($\overline{\mu}_m$) and the monitor indicates that the process is *IC* (type 2 error).
4. The defect deposition rate is high ($\overline{\mu}_m$) and the monitor indicates that the process is *OOC*.