Development of an Optimal Sampling Strategy for Wafer Inspection

Raman K. Nurani *, Ram Akella **, A. J. Strojwas
Carnegie Mellon University, Pittsburgh PA 15217
*** and University of California at Berkeley, Berkeley, CA 94720*
R. Wallace
KLA Instruments Corporation, San Jose, CA 95161
M. G. McIntyre, J. Shields, I. Emami
Advanced Micro Devices, Austin TX 78741
*** Supported by the Sloan Foundation and * Semiconductor Research Corporation*

Abstract

This paper presents a methodology for the development of an optimal sampling strategy for defect inspection, which is crucial for yield management of state-of-the-art technologies. This requires understanding of the defect-yield relationship and yield reducing process drift models. Further, the sampling plan is based on the trade-offs between the costs of sampling and of defective dies. Our methodology is demonstrated using data from different fabs.

1 Introduction

Comprehensive defect inspection techniques are of crucial importance in yield management for state-of-the-art technologies such as 0.35μm. In-situ particle monitoring and bare wafer inspection provide information on equipment cleanliness. Patterned wafer laser scanning and digital image processing techniques are employed to provide quick feedback on deposited film qualification as well as real time defect information on product wafers. To maximize the efficiency of wafer inspection it is necessary to develop an optimal sampling strategy.

This paper will discuss a methodology for the development of such a strategy, which should specify the number of inspected lots, the number of inspected wafers per lot, the number of inspected dies per wafer along with the spatial distribution of these dies, critical layers and a range of defect sizes per layer. In general, several sampling strategies are required to cover the different phases of technology development and manufacturing. In the initial ramp-up phase it is important to extract the equipment characteristics in terms of particle generation as well as sensitivities of a particular product to defects at the most critical levels. This information can be used to develop a yield model which takes into account defect density, clustering and size distribution for each critical level. In this yield learning phase, the so-called short-loop experiments, in which specially designed test structures are employed to focus on a particular process step (e.g. metal 1 patterning), are extremely valuable. The information gathered in these short-loop experiments can be then translated into the requirements for the full-flow monitoring and diagnosis. Correlation studies must be done to relate the probe yield losses to the defects detected via in-line inspection. As a result of these experiments an in-line inspection and sampling strategy can be developed for monitoring and for statistical process control in volume manufacturing to provide relevant information with a significant impact on the throughput and manufacturing cost. The sampling plan is based on trade-offs [2] between the costs of sampling and of defective dies, as well as on the limited capacity of inspection equipment.

In this paper we first demonstrate a methodology for in-line defect inspection based yield prediction using the data gathered in commercial IC fablines. Understanding the defect-yield relationship is necessary since defect monitoring is a surrogate for yield monitoring. Next we discuss the components of sampling strategy and present a defect sampling methodology framework. Then we illustrate the development of a sampling methodology using the data collected from a volume production fab. All the necessary data analysis techniques used to determine the defect distributions are presented as well. The paper concludes with a brief discussion on the economics of sampling.

2 Wafer inspection technology

To develop a comprehensive yield management system, a number of different approaches need to be integrated. These approaches range from the in-situ particle monitoring using equipment such as HYTs to wafer inspection to the more detailed defect source analysis methods (e.g. SEMs with de-layering). In this paper we will focus on the in-line wafer inspection. To assess the usefulness of wafer inspection, several key performance parameters must be established: speed of inspection, the ability to

Authorized licensed use limited to: KLA Tencor. Downloaded on October 30,2023 at 22:18:31 UTC from IEEE Xplore. Restrictions apply.

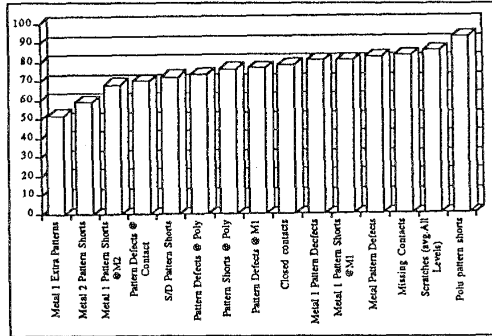

Figure 1: List of Defects with High Killing Rates

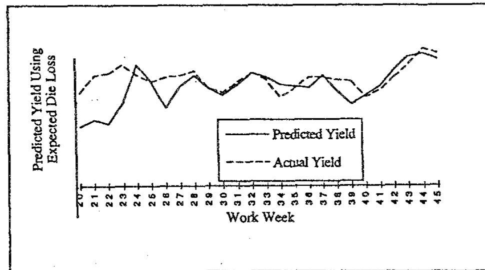

Figure 2: Correlation of Actual Yield to Expected Yield

inspect entire die area, defect capture rate and the ability to inspect defect types (both particles and process defects) for all layers and process steps.

The primary technologies available for wafer inspection include laser scanning, Fourier filtering and digital image processing. Laser scanning systems such as Tencor Surfscan 7600 are capable of a very high inspection rate due to large spot sizes. Fourier filtering systems such as IQ-165m employ real time scanning of the Fourier image, and with the pixel size of 1 μm are also capable of providing high speed inspections. Digital image processing systems such as KLA 2100 Series which utilize smaller pixel sizes and significant effort in image acquisition and processing are slower than the previous two methods. However, for the pixel size of 0.6 μm x 0.6 μm, even these systems can inspect the entire wafer below 10 minutes; this performance is acceptable for in-line monitoring (i.e., it would not impact the overall cycle time significantly). Both laser scanning and digital image processing techniques are capable of entire die inspection, while Fourier filtering is limited to repetitive patterns which makes it useful for memory arrays only. In terms of sensitivity, digital image processing techniques are superior and have proved capable of detecting defects well below 0.2 μm for arbitrary geometries. While the digital image processing techniques can detect all defect types on all layers and for all process steps, laser scanning systems are limited to light scattering particles and may have difficulties for certain steps (e.g. resist patterns) or layers (e.g. metalization layers with micro-roughness). In this paper, we focus on the defect information collected using KLA wafer inspection tools.

3 In-line inspection based yield prediction

The initial data for the sampling strategy development has been collected at the AMD manufacturing facility in Austin, TX. Four critical post etch steps were selected: Poly, Contact, Metal 1 and Metal 2. Wafers were inspected on the KLA 2130, the defects were reviewed and classified on the KLA 2606, and the results were stored in the KLA 2550 database. The passing and failing die coordinates at sort were matched to the in-line defects detected at the critical levels. Finally the kill rate values were determined for each defect type (see Figure 1).

It was determined that 44 percent of the die with the defects found via in-line inspection failed sort. Based upon this study, formulas for Expected Yield Loss (EYL) for each defect type were derived as a function of in-line defect density, area scanned and the ratio of die tested to die scanned . The correlation of the in-line detected defects to the actual die yield is quite good. This is demonstrated in Fig. 2 where the predicted and actual yield values per work week are shown. This type of information is essential to determine the sampling strategy for in-line wafer inspection; it permits the sampling strategy to focus on defect monitoring as a surrogate for yield monitoring.

4 Sampling Strategy

In this section we discuss the components of a sampling strategy, different analysis methods, as well as determinants of sampling strategies (such as the nature of the

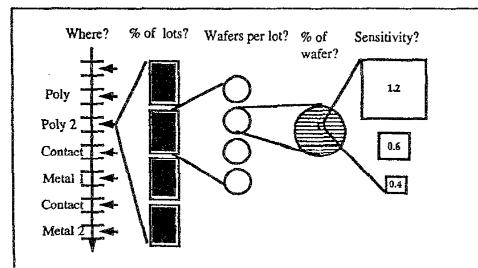

Figure 3: Decisions for Sampling Strategy

Authorized licensed use limited to: KLA Tencor. Downloaded on October 30,2023 at 22:18:31 UTC from IEEE Xplore. Restrictions apply.

line, whether it is a development or a high volume production line). Here, we focus on the use of inspection for the purpose of detecting yield reducing process drift. Figure 3 describes the ingredient decisions of a sampling plan and indicates the critical levels that most successful fabs use. Table 1 shows current best practice in terms of the values of sampling percentages. This translates into having one KLA 21XX series in-line wafer inspection tool per 750 wafer starts per week.

Parameter	Pilot	Production
Wafer Inspection Steps	20-24	10-12
% of Lots	100%	100%
Wafers per Lot	3-5/25	2/25
% of Wafer	100%	100%
Sensitivity	0.2μm	0.3μm

Table 1. Best Practices for 0.5μm Process

Current industrial practice is ad hoc and a result of historical evolution. As might be expected, the sampling effort is considerably more during the pilot or yield ramp mode, compared to the final production phase. Notice that percentage of lots inspected is the one area where the sampling percentage is equally high during the yield ramp and production phases; intuitively, this enables process drift tracking. Similarly, a greater variation in the defect density requires an increased rate of sampling.

Classical sampling design is typically based on the assumption that there are random variations in measures such as defect counts [4]. Control limits for SPC (Statistical Process Control) and sampling plan are then developed such that any change in the random variation can be detected with minimum average cost (including those costs due to searching for assignable causes, undetected out of control states, and sampling). However, the presence of additional spatial noise has been observed in many fabs. This is due to causes such as clusters or spatial defect variations resulting from process variations or the introduction of particles. These make it difficult, if not impossible, to use SPC based on the total defect count effectively [1].

Next we describe our research which is based on statistical analysis and which considerably enhances current practice. In particular, we describe our use of automated clustering to set up enhanced SPC approaches; we reduce the "noise" due to clusters so as to focus on the remaining "random" defects, rather than the traditionally monitored total number of defects. The resulting random defects, after removing the clusters, are assumed to be closer to the independent, identically distributed random variables used for analysis purposes than the original unprocessed defects. The resulting diagnostic and sampling plans are thus more accurate and effective. The measures of systematic variations such as the

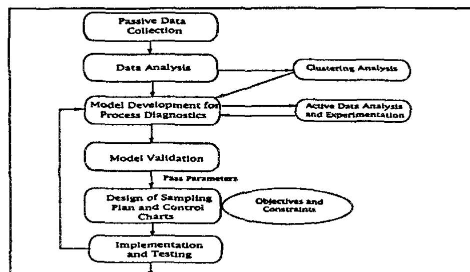

Figure 4: Defect Sampling Methodology Framework

number and type of clusters need to be monitored separately. Further, classification of the clusters enables the identification of process and other sources of defects, together with the attendant process improvement. Additionally, there is considerable debate about the use of defect counts versus defective die counts for monitoring purposes. We have observed that while defective dies are useful for yield prediction, defect counts are more useful for diagnostics.

Figure 4 outlines the basic steps of our approach for developing a sampling plan, as well as the associated monitoring and diagnostic policy. Passive data on defect information (count, size and class) with appropriate sensitivity setting and die yields is collected. The defect data are pre-processed by removing the systematic spatial variations, such as those due to clusters, so as to reduce the noise in the defect data. Statistical analysis (such as ANOVA) is used to understand the inter-wafer and inter-lot variations. An appropriate static or dynamic model is fitted to the pre-processed data. More generally, defect count process could be modeled as a stable defect level with small random variations, step jumps/excursions, or linear drift, or a combination of all these. Additional data, either historically available, or available through active experimentation, is used to validate the models. An estimator-detector, such as a static SPC or dynamic SPC (e.g. a Kalman Filter) is developed based on the process model, for future real-time use. The residual errors are computed and a hypothesis testing procedure is then used for the out-of-control state determination based on a given sampling plan. This model is then used along with the cost trade-off model to determine the optimal sampling plan. The last step, implementation and testing, requires testing residuals and updating process and drift parameters.

5 Illustrative Examples

We now describe two examples of fab data, demonstrate the specific applications of the approach outlined above in these cases, and discuss the results. We have used ag-

Authorized licensed use limited to: KLA Tencor. Downloaded on October 30,2023 at 22:18:31 UTC from IEEE Xplore. Restrictions apply.

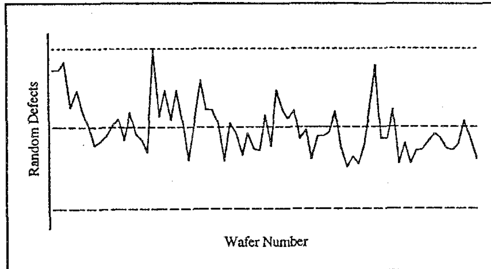

Figure 5: Random Defect Control Chart for Fab A

gregate defect count data for the purposes of the current analysis, although future work will use defect size and type information also. More details can be found in [3].

Fab A: Static Model and SPC Development

The data for this section has been collected at the AMD site mentioned earlier and corresponds to Post Etch Metal 1. The wafer maps are typically noisy in that the defects are not random, but have spatial distributions caused by process problems. We use automated clustering analysis to remove non-random defects. The remaining defects are assumed to be random, although causes which are currently unknown (and consequently unassigned) may remain. The standard Chi-square goodness-of-fit test reveal that the random defects conformed to the Poisson distribution. Figure 5 illustrates the random defect data with control limits obtained using the Poisson distribution. In this setting classical control charting and sampling plan schemes can be used. Also, we have developed a procedure for relating the sampling plan to process noise variability.

Fab B : Dynamic Model and SPC Development

Figure 6 below outlines the random defect trend. The trend indicates a stable defect level with small variations, step jumps and a linear trend with variations. We developed a dynamic model to fit the data, where the random

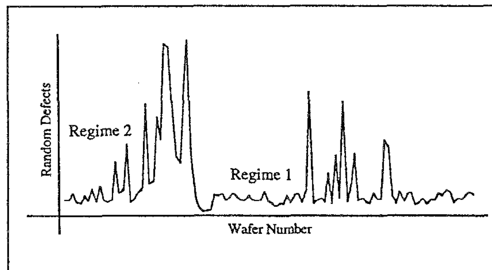

Figure 6: Random Defect Trend for Fab B

defect count variation with the wafer number shifts between two regimes. Regime 1 corresponds to random variation around a constant mean defect level. Regime 2 corresponds to an AR(1) model, i.e., an Autoregressive model of order 1, where there is a random variation around a linear trend in the defect count. We then use hypothesis testing, based on a window of x observations, to detect the onset of Regime 2, which corresponds to a process drift.

We develop the sampling plan by trading off the costs of sampling, being in undiagnosed out of control state, and search costs due to a in-control state mis-diagnosed as a bad state. The inter -wafer and -lot variances and the resultant estimator-detector error variances are also inputs to this model. The sampling frequency is derived from optimizing the costs/profits.

6 Economics

The purpose of this section is to show, using a simple model, that higher wafer inspection cost will be offset by the increased learning and subsequent defect reduction (Fig. 7). The vertical axis is not to scale. The model assumes an exponential decrease in ASP (average selling price) of each die with time and an average silicon cost of $4 per sq. cm. Fab X invests 3.2% of the cost of silicon for wafer inspection and learning efforts whereas Fab Y invests 1.6%. This results in an increased learning rate for Fab X. The rate of volume ramp is same for both the fabs. The benefit is equal to total revenues minus the total costs of processed silicon and of wafer inspection. Such economic analysis can be used to optimize the sampling plan.

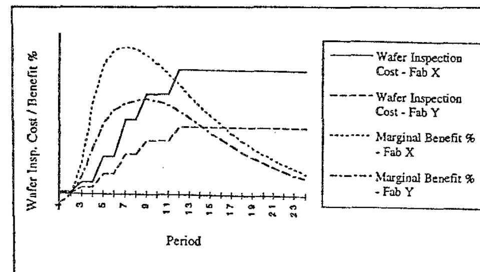

Figure 7: Economic Model

References

1) D. J. Friedman and S. L. Albin, *IEEE Transactions on Semiconductor Manufacturing.*, vol 4, 36-42, (1991).

2) T. J. Lorenzen and L. C. Vance, *Technometrics*, vol 28, 3-10, (1986).

3) R. K. Nurani and R. Akella, KLA Yield Management Seminar, (1994).

4) C. J. Spanos, *Proc. of IEEE*, vol 80, 819-830, (1992).

Authorized licensed use limited to: KLA Tencor. Downloaded on October 30,2023 at 22:18:31 UTC from IEEE Xplore. Restrictions apply.