CARNEGIE MELLON UNIVERSITY

YIELD FORECASTING

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY
in
ELECTRICAL AND COMPUTER ENGINEERING

by

PRANAB KUMAR NAG

Pittsburgh, Pennsylvania
April 17, 1996

Yield Forecasting

Abstract

Success of modern semiconductor manufacturing has been achieved through a number of key innovations in the areas of IC design, manufacturing, testing and failure analysis. Research and development in each of these areas have grown to such a level of complexity - reflecting the complexity of today's industry - that the interdependence among these areas has largely been side-tracked. But the performance of a semiconductor industry is not only dependent on the advancements in these individual areas but also on their interactions.

One of the significant detractors of cost in a modern manufacturing line is yield loss due to contamination and the time required to ramp-up the yield to profitable levels. Yield loss and its rate of change with time in a manufacturing line is determined by the various attributes of fabrication, product, testing and failure analysis and their interactions which determine yield learning rate. This research attempts to gain an understanding of the nature of this inter-relationship by addressing the problem of predicting yield as a function of time for a multi-product manufacturing line.

In this thesis, first the process of contamination related yield learning as it happens in a manufacturing line is presented. Then a methodology to predict yield learning curves for a multi-product manufacturing line is proposed. A suite of models has been developed which capture the primary factors determining yield learning rate. The methodology and models have been implemented in a discrete event simulator - Y4 (Yield Forecaster). Through a series of simulation experiments, estimates of performance parameters like cycle time, yield, test escapes, and learning rate are presented to illustrate some of models individually. Then another series of experiments are presented to illustrate the applicability of Y4 in performing cost-revenue trade-off studies for a variety of situations. Through these experiments, it is concluded that more attention must be devoted to characterizing those attributes of product and failure analysis that determine the ease of diagnosis. But more importantly, the interrelationship between manufacturing entities should be characterized well in order to be able to and determine cost benefits of making improvements in the design objectives.

Acknowledgments

First of all, I would like to thank my advisor, Prof. Wojciech Maly, for introducing me to a field of research which allowed me to look at engineering from a broad perspective. His own relentless efforts to gain deeper understanding of his subject has been an inspiration for me to question and perfect every argument that I came across.

I would like to thank Prof. Stephen Director, who in spite of his busy schedule, had time to provide comments on and critique the technical contents of this research. I am very thankful to Prof. Hank Walker who has been following this work from its very beginning and has provided invaluable feedback all the way. Dr. Hermann Jacobs, with whom I had the opportunity to work as Siemens, provided many needed encouragements. He has been instrumental in providing the necessary technical data without which this research would have been incomplete.

It was unfortunate that Dr. Charles Stapper from IBM could not make it to the oral presentation of this work (I earnestly hope that he recovers from the car accident which has left him paralyzed). But I will be forever indebted to him for providing me the opportunity to work with him in the summer of 1991. It was then that this project took its initial shape and that experience provided me with the confidence to embark on such a monumental task.

I would like to offer my appreciation to the member companies of Semiconductor Research Corporation and Sematech for their gracious financial support. In particular, I would like to thank Darius Rohan of Texas Instruments, Dallas, and Randy Hughes of Tyecin System, San Jose, for inviting me to talk to their co-workers in their respective companies. I would also like to thank Steven Brown, then at Sematech, for providing technical input.

It would have been nearly impossible to conduct this research without the support of Elaine Lawrence, Lynn Phillibin, Lyz Knight and Judy Bandola. Meeting the omnipresent critical deadlines would not have been possible without their fullest help in every respect. I learned a lot from my past and present colleagues, Phil Nigh, Tom Storey, Derek Feltham, Samir Naik, Anne Meixner, Jitu Khare, Anne Gattiker, Hans Heineken, Mathew Heath, and Charles Ouyang. I am thankful to them for listening patiently to my several practice talks and providing constructive criticisms on my presentations.

My personal friends, Suresh Konda, Stephanie Szakal, Eswaran Subrahmanian, Anna and Antonio Fevola, Sangeeta and Krishna Pendyala, Balu and Jyoti Sarma, Rajat Ghosh, Ajapa Mukherjee, and Harsh Baxi, provided a warm, homely atmosphere throughout my years of study at CMU. I want to express my special appreciation of little Kiernan, who, in all his wisdom, stated that I am now a *doctor of nothing*, putting my professional pride in the correct perspective from a universal view point.

Most of all, I would like it to be known that my lovely wife, Nandita, had been the greatest critique of whatever I did as a part of my research and personal adventures. Her ability to question why I did whatever I did helped to put my understanding of my life in proper perspective. The foundation of all my achievements was laid by my parents who unquestioningly supported my quest for realizing my professional dreams. The conclusion of my dissertation was followed by my sister, Kity, becoming a mother reminding me that life's natural wonders never cease in spite of major technical advances.

to

my nephew Indranil Bose

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Rapid growth in several key technological areas such as communications, transportation, computers, software and consumer electronics has been possible due to tremendous advances in the integrated circuit manufacturing technology [1]. In the last 25 years, the semiconductor industry has grown from the Intel 4004 chip containing 2300 transistors to the 9.3 million transistors of Digital 21164 [2, 3, 4] closely following Moore's Law [5]. This has been possible due to the equally rapid advancements in manufacturing process and IC design technology. Increasing demands for more powerful and smaller computing machines have fueled the need to manufacture such complex ICs. It is predicted that this need of the electronic industry for even more complex ICs is going to grow at an even faster rate. By the end of this century a state-of-the-art microprocessor may be expected to contain upwards of about 50 million transistors [6, 7].

From a cost point of view, such a growth was unhindered in the past mainly because manufacturers were able to maintain sufficient volume of production to ensure low cost per fabricated unit. Increasing demand for more IC products have, however, attracted more manufacturers making very similar products resulting in a highly competitive market. At the same time, the continuous drive towards smaller feature size on an IC and larger die size itself has caused an increase in the cost of manufacturing. Tough competition and increasing cost have thus made semiconductor manufacturing a risky venture.

To be more specific, the cost of a new VLSI fabrication line producing several different products using several hundred processing steps, is now estimated to be close to a billion dollars [8, 9]. Both the cost and complexity of manufacturing have been observed in the past to increase exponentially and there has been no indication that this trend is going to slow down [10, 11, 12]. To maintain a competitive edge, ICs must be precisely manufactured within tight tolerances. Thus, in order to keep the cost of manufacturing down one must ensure that no errors are made during any of the stages of producing ICs from design to packaging.

Manufacturing ICs without any errors is a complex task [13]. Whenever a new process or product is introduced, the manufacturing yield or the fraction of correctly manufactured ICs is usually low. One has to ensure not only that the processes and products are designed to be high-yielding but also that errors in manufacturing are eliminated as quickly as possible through continuous and timely improvements. This correction process is known as *yield learning*.

Rapid yield learning is key to manufacturing success [13]. High yield not only translates to lower cost per unit but also means that a larger number of ICs can be delivered in time to maintain a competitive edge. Therefore, one must be able to ramp up yield quickly using available resources efficiently. The rate of yield learning is a function of a number of inter-dependent attributes of IC design, fabrication process and the failure analysis facility. These attributes in turn depend on a large number of possible choices related to design style, products, equipment, technology, etc. Therefore, the technology to debug the manufacturing line needs to be more complex and advanced than the product technology to be effective and, hence, very costly. Optimum exploration of cost-revenue trade-offs requires adequate simulation and experimental models to be developed taking into consideration the yield learning process.

1.1 Yield Loss in IC Manufacturing

Figure 1.1 shows the sequence of stages in a manufacturing line leading to production of ICs from bare silicon wafers. These four stages are: wafer fabrication, probe testing, packaging and final testing. During wafer fabrication the IC structure is defined on the wafers each of which is tested during the probe testing stage. The wafers are diced and the dies which pass the tests are packaged and subjected to a final suite of tests before shipping. Errors or imperfect processing, which can occur during any of these stages, may lead to some or all ICs on each wafer to malfunction. Such malfunctioning ICs are detected at one of the two testing stages. In a manufacturing line most of the causes of yield loss occurs during the wafer fabrication stage. In fact, any yield loss observed at the subsequent stages is likely to have originated at the wafer fabrication stage. Some of the yield loss observed in the later stages could also be due to wafer handling problems.

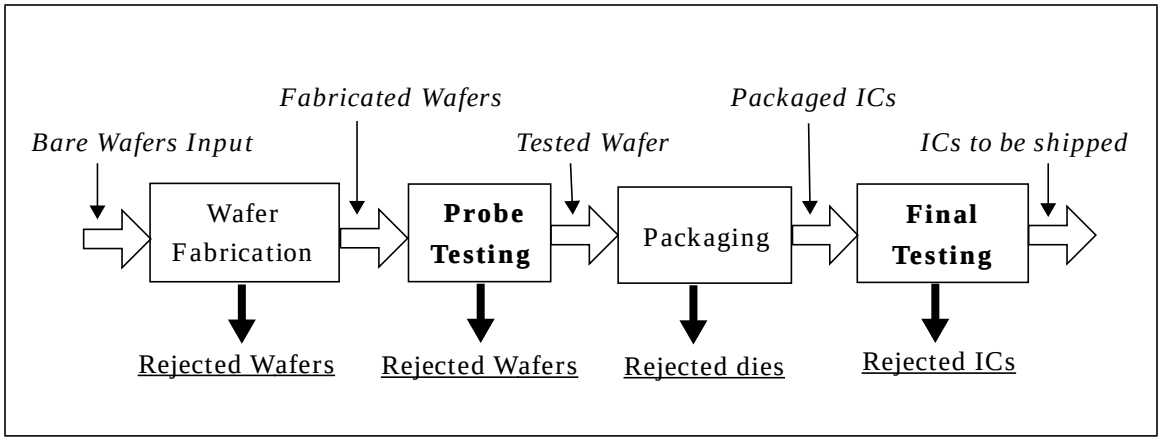

Figure 1.1 Flow of a manufacturing process.

The reasons for yield loss during the wafer fabrication process can be classified as shown in Figure 1.2. The two main classes are throughput yield loss and die yield loss. This classification does not include systematic yield problems related to design errors, and only yield problems due to random events in manufacturing are considered.

Throughput yield loss, as the name suggests, is the difference between the input rate and output rate of wafers during the fabrication stage. This difference can be due to wafers being rejected because of misprocessing or mishandling. Misprocessing can happen because of equipment failure, incorrect sequencing of wafers, etc. Mishandling of wafers by the operators can lead to wafer breakage, gross defects on the wafers, etc. In a modern manufacturing line throughput yield loss is usually very low because most of the steps are automated.

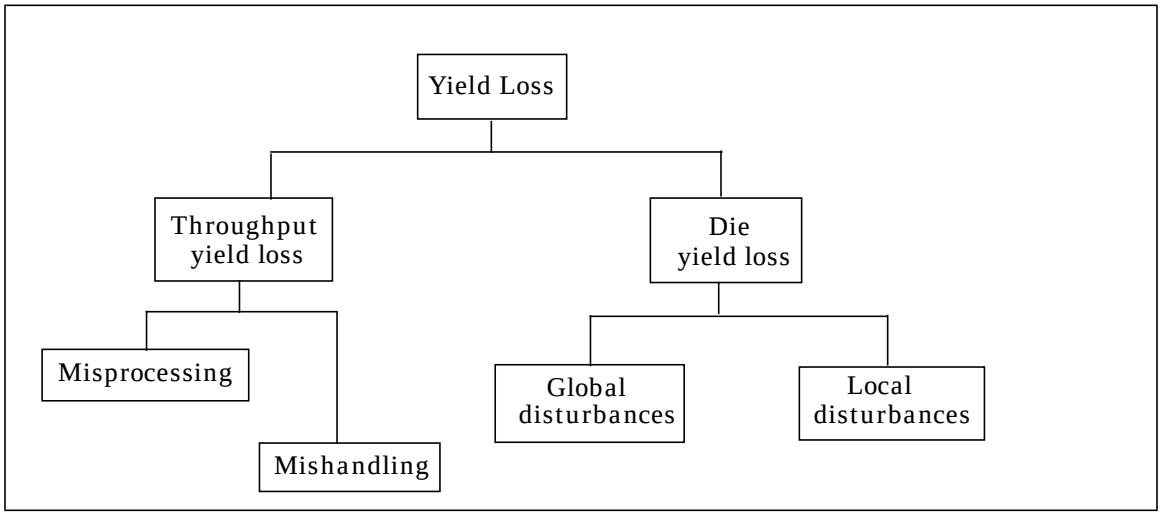

Figure 1.2 Types of yield loss in fabrication process.

Die yield loss is defined as the fraction of the total ICs manufactured which are defective. The disturbances leading to die yield loss can be further classified into two types: global and local disturbances [13, 14]. Global disturbances are those which affect entire wafers in such a way that all or most of the dies fail to meet acceptance criteria. These could be random variations in equipment settings, material properties or errors in masks [15]. More specifically, variations in gas pressure, temperature settings, dopant concentrations, etc., are examples of global disturbances. The effect of such variations can be uniform across the entire wafer surface or can be non-uniform. For example, if the etching time for contacts is insufficient then one can expect the con-

tact resistance to be higher than nominal for all dies on a wafer. Alternatively, if there is an etch rate variation across the wafer surface then one might observe higher than nominal contact resistance in only some of the dies. As in the examples above, global disturbances usually affect the electrical properties of the transistors and interconnects leading to variation in performance parameters such as speed and power consumed by the ICs. A failure is said to occur when performance parameters are outside accepted limits. Such failures are more commonly referred to as *parametric failures* [13].

Unlike global disturbances, local disturbances affect portions of the wafer surface whose dimensions are comparable to those of IC features such as transistors, contacts, etc. The local deformations manifest themselves as small regions of extra or missing material in the IC structure and are referred to as *spot defects* [16, 17, 18]. Spot defects can occur in any of the conducting, semiconducting or insulating layers of the IC and may lead to alteration in the topography of the intended circuit. For example, a spot defect can cause a short between two or more electrically unconnected nodes or, a break in an electrical path, etc. Such topographic changes in the circuit alter the intended functionality of the circuit and therefore the resultant circuit failures are referred to as *functional failures*, or *faults* in ICs [16].

Spot defects leading to functional failures are caused by the presence of *contamination* from various sources during fabrication of wafers. Such contamination are particulate in nature comprised of solid particles or liquid droplets. These particles may be present in the materials used for processing, generated by the equipment, or may even be something airborne.

From a manufacturing cost perspective, it is important that the yield loss resulting from such diverse causes be as little as possible [19, 20]. During the early stages of manufacturing - the prototyping stage - the yield is low because of both global and local disturbances. At this point, the focus is on correcting or controlling the global dis-

turbances which cause entire wafers to fail. This is achieved through observing electrical parameters of fabricated dies, measuring in-line parameters (such as dimensions of deposited material), etc., and subsequently correcting process settings to produce the desired results. This is referred to as statistical process control (SPC) [15, 21, 22].

Once the fabrication line is stabilized from the point of view of global disturbances, the focus is shifted towards correcting yield loss due to local disturbances. During this stage - the yield learning stage - failed dies are analyzed and corrective actions are taken to control the level of contamination. This stage is also accompanied by an increase in the volume of production. Time domain changes in yield at this stage have a substantial impact on the cost of manufacturing and the accrued profits. This research focuses on the defect limited yield learning for a manufacturing line.

Eventually, the rate of yield learning decreases as the yield approaches 100% and this is the high volume stable manufacturing stage. Any semiconductor manufacturing operation would like to reach this stage as quickly as possible since the bulk of the profits are realized in this period. Figure 1.3 shows an example average yield vs. time curve illustrating the three stages of manufacturing described above.

Time domain changes in yield could also be the result of an inherent change in the nature of the disturbances, but it is mainly due to the deliberate continuous improvements made in the design and to the process. The rate of yield learning could have been slower or faster (shown as dashed curves in Figure 1.3) depending on how quickly one is able to remove the process problems. A slower rate of yield learning can result not only in loss of revenue but may also lead to losing the market to other competitors. A higher rate of yield learning may require a more costly and complex contamination control strategy. Understanding this cost-revenue trade-off is a necessity in decision making.

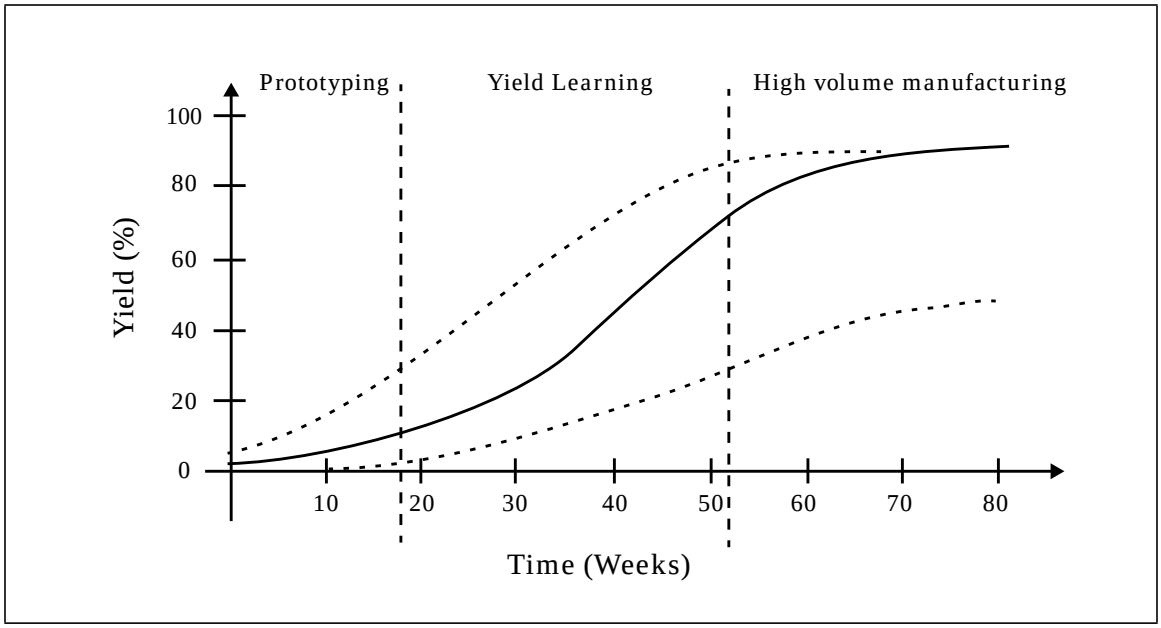

Figure 1.3 Yield vs. time curves.

1.2 Yield Learning Curve as a Decision Tool

In the past, learning curves have been used widely to prepare cost reduction programs, forecast price, and set product development goals in several industries such as automobiles, airplanes, steel, chemical, etc. [23, 24]. In 1936, the concept of a learning curve was first defined with the observation that the man-hours required to assemble an airplane declined by 20% each time the unit production doubled [23]. In this paper, data from 15 industries were presented to illustrate the cost reduction obtained with increasing production volume. It was also observed that semiconductor industry consistently demonstrated a higher rate of learning during the period 1973-1978 due to increase in volume.

The cost of manufacturing in the semiconductor industry is, however, determined to a greater extent by the rate of increase in yield rather than reduction in man-hours. Increasing capacity of production can reduce the unit cost of production only to a certain extent [25]. In the semiconductor industry the choice of technology, equipment, product, design style, etc. is very diverse and so are the number of ways yield loss can

occur. The choices regarding techniques and equipment available for correcting causes of yield loss are similarly large. Each cross-section of choices must be evaluated from the yield learning perspective to be of any use in the decision making process. From Figure 1.3, it is clear that the actual nature of the yield learning curve is of great strategic interest.

Besides using yield learning curves to choose the appropriate resources, one can also use them to evaluate the operating strategies of a manufacturing line. For example, one can specifically dedicate certain products, including specially designed test structures, to aid in analyzing the causes of failures. Appropriate sampling rates of wafers for analysis can be estimated for achieving better yield learning rates. Problem areas can be isolated and resources can be reassigned to deal with specific types of yield loss efficiently.

The success of a manufacturing line also depends on timely alterations of resources and products. In this case, one can use yield learning curves to aid in deciding when to introduce new products in the line which could also be simply smaller or more advanced versions of the existing products. In this way one can take advantage of a partially “debugged” line to achieve a higher rate of learning for a new product. Timely introduction of new equipment and technology can also be judged based on the estimated yield learning curves. Maximizing the use of existing equipment and technology is of concern given that product and technology life cycle is usually only a few years.

1.3 Need For a Methodology For Yield Forecasting

In the past, some methods were developed based on mapping the yield learning curves of past products and technology onto new ones [25, 26, 27]. In these methods the rate of learning was assumed to be known a priori or was assumed to be easily obtainable by extrapolating from past yield data. Such assumptions are valid in cases

where a manufacturing line is dedicated to a single product like DRAM, and, extrapolating is useful to some extent as shown by [26]. In another method, instead of modeling yield, the defect rate is modeled as a function of time [26, 28]. In both of these methods neither the yield learning process is considered nor do they describe the learning rate in terms of physical attributes of yield loss mechanisms, products, testing strategy, failure analysis strategy, etc.

A yield predicting methodology based on such attributes can be used to address several important issues in a manufacturing line. When a new product is introduced in a line one would like to know the length of time to reach a particular value of yield. If the length of time is unacceptable then one would like to know means to shorten it. There can be a number of available choices but one must be able to quantify the cost effectiveness of each of the options. The options could be alternate product design, cleaner equipment or more failure analysis resources. But more often than not the best solution can be a combination of a number of options. For example, one can use an easily diagnosable product like memories to bring the yield up to a certain level quickly. Then the second product can be introduced at a predetermined time to take the maximum advantage of higher initial yield and learning rate. But such quantifications require that a yield prediction methodology be developed taking into account the inter-relationship between domains which are traditionally considered in isolation.

Using yield learning curves to aid in the decision making process is appropriate only when combined with cost estimates or cost learning curves. As a simple example, Figure 1.4 shows cumulative manufacturing cost and revenue versus time curves. The intersection, point A, represents the point in time when the manufacturing line starts to make a profit. If the product is replaced by another design which is easier to diagnose then the revenue curve could look like the one shown by the dotted line. Time to profit (point B) is shorter than the previous situation. However, if the manufacturing

cost increases simultaneously (shown by dashed line) then the time to profit could be as shown by point C. The worst case situation is when cost of manufacturing is high and yield learning rate is low as represented by point D on the graph. In this case, the risk of losing business to the competition is very high.

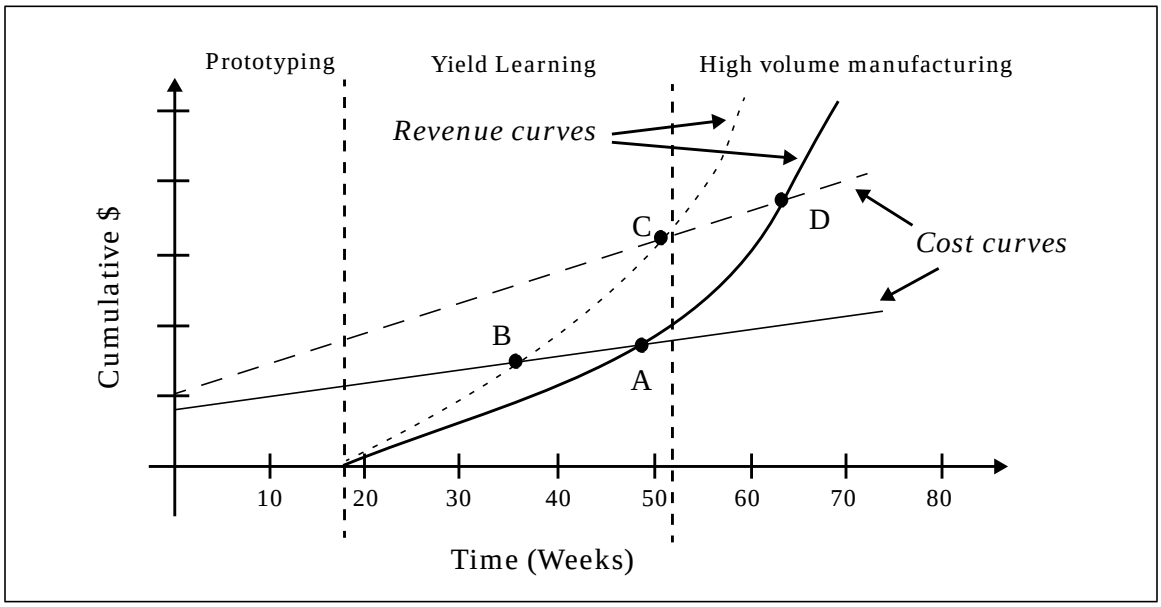

Figure 1.4 Impact of yield learning on *time-to-profit*.

None of the methods developed earlier have the capability to estimate cost learning curves. Thus there is a need for developing a methodology which takes into consideration the yield/cost learning process in such a way that its is useful for evaluating manufacturing strategies.

1.4 Research Goals

The goal of this research is to develop a methodology to predict yield and cost as a function of time for a given multi-product manufacturing line. Such a methodology should take into consideration the essential elements which govern defect related yield and its rate of change with time in a modern semiconductor manufacturing line. These elements of a manufacturing line are: product design attributes, timing and

operational attributes of wafer fabrication, efficiency and accuracy of failure analysis, the nature and effect of corrective actions, and contamination properties. Such a methodology should be combined with appropriate models describing the characteristics of these elements which are consistent with the methodology to forecast yield. The models must closely reflect observed phenomena and take into account the complex interactions between various manufacturing attributes. Appropriate cost models should be developed to enable cost-revenue trade-off studies.

1.5 Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 discusses the general nature of the yield learning process in a manufacturing line. The various attributes of manufacturing elements are discussed and presented in this chapter. Based on this discussion a general methodology to predict yield learning curves is derived and presented in Chapter 3. This methodology is shown to be suitable for simulation and basic requirements for simulation are also discussed. Chapter 4 presents the simulation models which are mainly derived from existing models after making certain simplifying assumptions. In some instances new models have been developed to describe the inter-relationship between manufacturing attributes. Chapter 5 describes the general organization of Y4 - software which implements the yield forecasting methodology and models. Some basic results are presented in Chapter 6 to illustrate the general capabilities of Y4 in mimicking well understood phenomenon. Chapter 7 deals with more involved simulation experiments which illustrate the relevancy of the models towards developing manufacturing strategies. Future work in several directions are suggested in Chapter 8 and conclusions in Chapter 9.

References

[1] “A Survey of the Computer Industry - The third age”, *The Economist*, pp. 2-22, Sept. 17, 1994.

[2] S. Mazor, “The History of the Microcomputer - Invention and Evolution”, *IEEE Proceedings*, vol. 83, no. 12, pp. 1601-1608, Dec, 1995.

[3] C. R. Barett, “Microprocessor Evolution and Technology Impact”, *1993 Symposium on VLSI Technology - Digest of Technical Papers*, pp. 7-10, Kyoto, May 1993.

[4] T. R. Halfill, “Intel’s P6,” *Byte*, vol. 20, no. 5, pp. 42-58, April 1995.

[5] G. Moore, “VLSI: Some Fundamental Challenges,” *IEEE Spectrum*, vol. 16, p. 30, 1979.

[6] *The National Technology Roadmap for Semiconductors* - Semiconductor Industry Association, 1994.

[7] J. D Meindl, “The Evolution of Solid State Circuits: 1958-1992-20??”, *1993 Int. Solid State Circuits Conference (ISSCC) - Commemorative Supplement*, pp. 23-26, Feb, 1993.

[8] *Mid-Term 1995 Status and Forecast of the IC Industry*, Integrated Circuit Engineering Corporation, Scottsdale, AZ, 1995.

[9] G. D Hutcheson and J. D. Hutcheson, “Technology and Economics in the Semiconductor Industry”, *Scientific American*, pp. 54-62, Jan. 1996.

[10] W. Maly, “Prospects for WSI: A Manufacturing Perspective,” *IEEE Computer Magazine*, vol. 25 no. 4, pp. 58-65, April 1992.

[11] W. Maly, “Cost of Silicon Viewed from VLSI Design Perspective,” *Proceedings of the 1994 Design Automation Conference*, pp. 135-142, 1994.

[12] S. Director and W. Maly and A. J. Strojwas, *VLSI Design for Manufacturing Yield Enhancement*, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, MA 02061, 1990.

[13] W. Maly, “Computer-Aided Design for VLSI Circuit Manufacturability,” *Proceedings of the IEEE*, vol. 78 no. 2, pp. 356-392, February 1990.

[14] W. Maly and A. J. Strojwas and S. W. Director, “VLSI Yield Prediction and Estimation: A Unified Framework,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-5, no. 1, pp. 114-130, January, 1986.

[15] W. Maly and A. J. Strojwas, “Statistical Simulation of the IC Manufacturing Process,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 1 no. 3, pp. 120-131, July 1982.

[16] W. Maly, "Modeling of Lithography Related Yield Losses for CAD of VLSI Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 4 no. 4, pp. 166-177, July 1985.

[17] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated Circuit Yield Statistics", *Proceedings of the IEEE*, vol. 71, no. 4, pp. 453-470, April, 1983.

[18] T. L. Michalka, R. C. Varshney and J. D. Meindl, "A Discussion of Yield Modeling with Defect Clustering, Circuit Repair, and Circuit Redundancy," *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, no. 3, pp. 116-127, August 1990

[19] .B. T. Murphy, "Cost-Size Optima of Monolithic Integrated Circuits," *Proceedings of the IEEE*, vol. 52, no. 12, pp. 1537-1545, December 1964.

[20] R. B. Seeds, "Yield, Economic, and Logistic Models for Complex Digital Arrays," *IEEE International Convention Record*, pp. 60-61. March 1967.

[21] R. W. Berger and T. H. Hart, *Statistical Process Control, A Guide for Implementation*, The American Society for Quality Control, Milwaukee, 1986.

[22] P. K. Mozumder, C. R. Shyamsundar, and A. J. Strojwas, "Statistical Control of VLSI Fabrication Processes: A Framework", *Trans. on Semiconductor Manufacturing*, vol. 1, no. 2, pp. 62-71, May 1988.

[23] J. A. Cunningham, "Using the learning curve a management tool", IEEE Spectrum, pp. 45-48, June, 1980.

[24] C. J. Teplitz, *The learning curve desk book: a reference guide to theory, calculations, and applications*, Quorum Books, 1991.

[25] D. Dance and R. Jarvis, "Using Yield Models to Accelerate Learning Curve Progress", *IEEE Trans. on Semiconductor Manufacturing*, vol. 5, no. 1, pp. 41-46, Feb, 1992.

[26] C. H. Stapper and R. J. Rosner, "Integrated Circuit Yield Management and Yield Analysis: Development and Implementation", *Trans. on Semiconductor Manufacturing*, vol. 8, no. 2, pp. 95-102, May 1995.

[27] V. Ramakrishna and J. Harrigan, "Defect Learning Requirements", Solid State Technology, pp. 103-105, Jan, 1989.

[28] D. R. LaTourette, "A Yield Learning Model for Integrated Circuit Manufacturing", *Semiconductor International*, pp. 163-170, July, 1995.

Chapter 2

Yield Learning in VLSI Manufacturing

In this chapter, a background on the process of yield learning in a manufacturing line will be presented. The discussion will be based on two important interdependent issues of a manufacturing line. One issue is the physical components which make up a manufacturing line and their relevant attributes. The second issue is the causes of yield loss and the nature of dependence of manufacturing line attributes on the yield loss mechanisms.

2.1 Organization of Manufacturing Process

In Figure 1.1 on page 3, the structure of a manufacturing line was shown from the perspective of a linear flow of wafers from input to the point where IC's are ready to be shipped. From a yield learning point of view, however, one has also to factor in the role of *failure analysis* as shown in Figure 2.1. After the probe and final testing stages, wafers with defective dies and packaged defective IC's are sampled for failure analysis. After the causes of observed failures are detected and diagnosed, certain *corrective actions* need to be taken. Corrective actions are taken so as to remove or reduce causes of yield loss in the fabrication stage.

One can further simplify the view of a manufacturing line as shown in Figure 2.2 since such a view is sufficient to illustrate the main attributes of yield learning [1]. Here, the manufacturing line is shown to consist of three phases: wafer fabrication, probe testing and failure analysis. Wafers are processed in a sequence of steps defined by the process recipe. At each step, a unique piece of equipment is used, and a specific

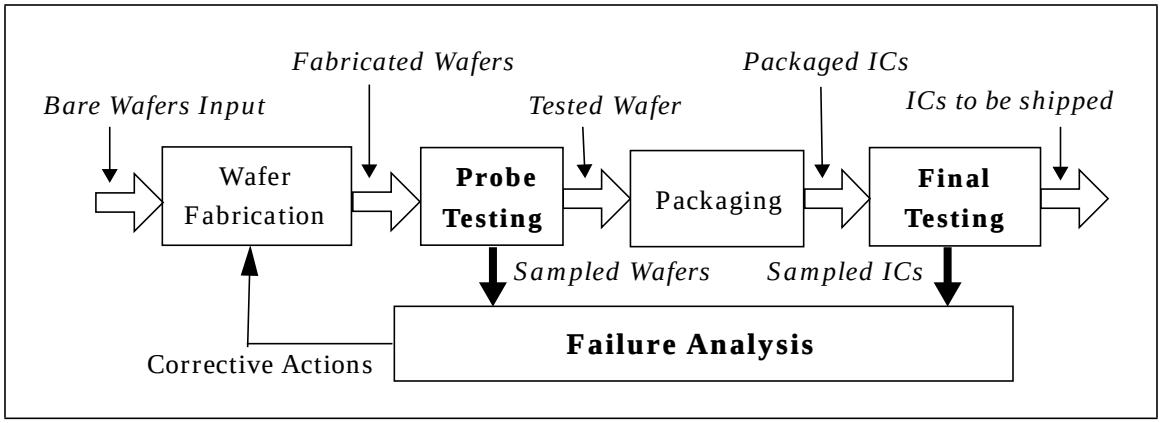

Figure 2.1 Structure of Manufacturing line - yield learning perspective.

layer of the IC defined. *Disturbances* can be introduced at each of these steps resulting in a less than ideal environment for processing the wafers.

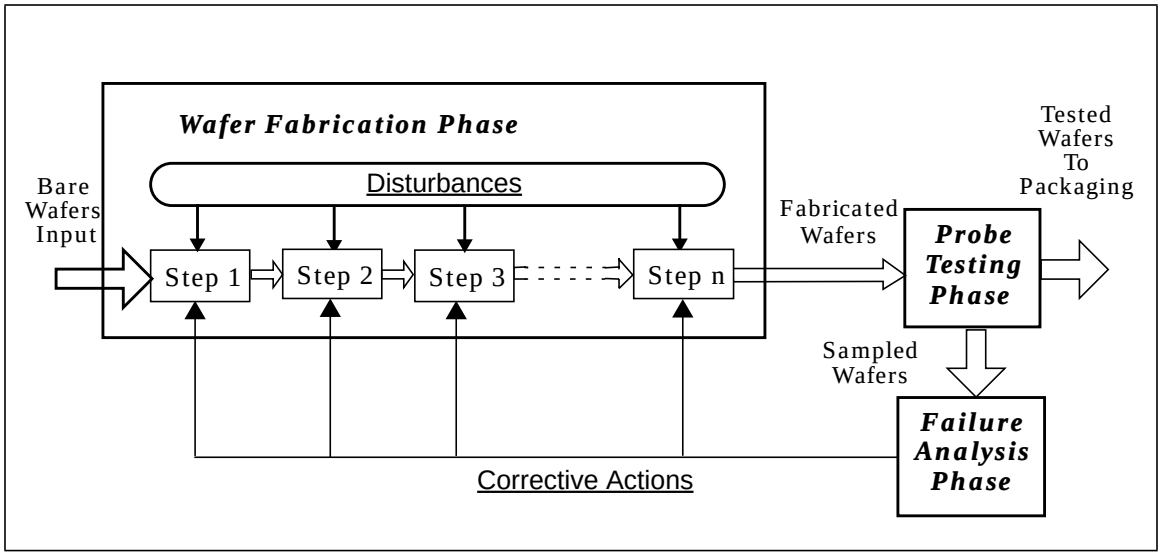

Figure 2.2 Simplified structure of a manufacturing line.

After completing the fabrication process, every die on each wafer is subjected to probe testing to detect faults. The tested wafers are diced and the functionally acceptable die are packaged and tested further. A fraction of the tested wafers are selected to perform failure analysis. During failure analysis a fraction of the defective die on the sampled wafers are carefully analyzed in order to detect the dominant cause of

failures. Based on this analysis, corrective actions are taken on the piece of fabrication equipment found responsible for the observed failures. Let us now take a closer look at each of these three phases along with the reasons for yield loss and the effect of corrective actions.

2.2 Wafer Fabrication Phase

IC fabrication can be viewed as a process of moving wafers in groups (called lots) through a sequence of equipment as defined by the process recipe. A number of factory attributes must be considered as shown in Figure 2.3. *Physical organization* refers to manner in which equipment, storage areas, etc. are located. Process recipes impose a *conceptual organization* on the fabrication line since defining IC layers requires a series of steps to be performed that may belong to different physical partitions of the line. Wafers are differentiated by the *product* they belong to and their characteristics also determines the operation of a fabrication line. *Operating personnel* have the responsibility of sequencing and scheduling of wafers through the equipment. And lastly, a number of *operating rules* are used to aid in scheduling.

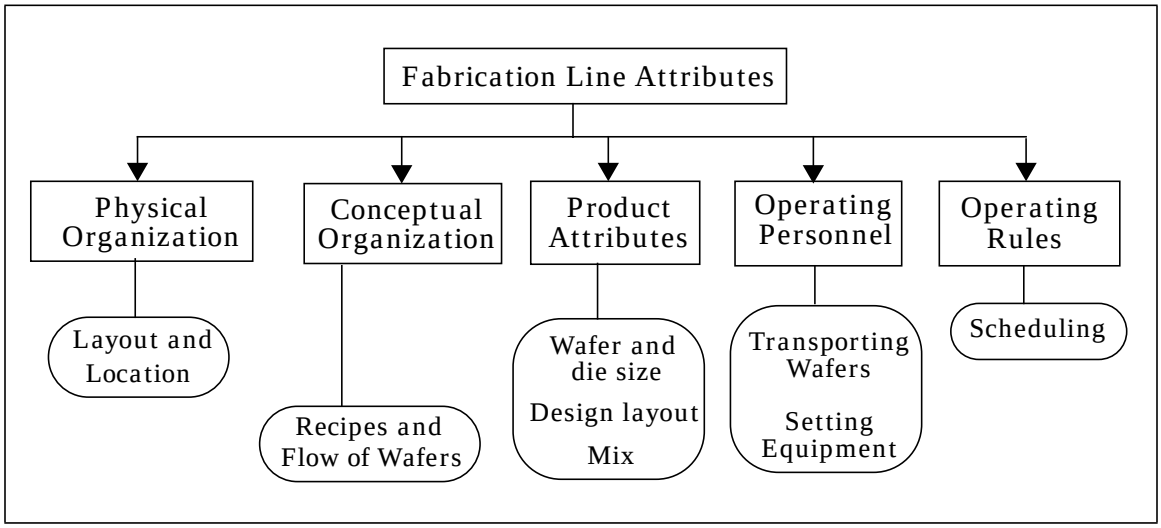

Figure 2.3 Wafer fabrication attributes.

2.2.1 Workstations, Equipment and Storage Areas

At the top level, the fabrication line is physically organized as *work-areas* where related process steps like lithography (resist-spin, expose, bake, etch) are performed [2, 3]. As shown in Figure 2.4, each work-area is further divided into *work-stations* which are, in turn, composed of a number of pieces of *equipment*, all generally capable of performing the same step such as oxidation, diffusion, etc. Each workstation is associated with a *storage* or *stocking area* where the wafers (lots) are temporarily stored. Broadly, there are two kinds of equipment that a wafer encounters during fabrication: *processing* and *measuring* (*metrology*) equipment. Processing equipment actively alters the surface of the wafers by depositing, oxidizing, etching, etc., defining the IC structure and its electrical properties. The capacity, or the number of wafers that can be processed in a piece of equipment is not necessarily the same as the lot size (number of wafers in a lot is usually about 24-25 wafers). Some are *batch equipment*, like furnaces, which can process usually about 100 or more wafers at a time. Some equipment like steppers can process only a single wafer at a time and some others like resist spin-on equipment can process a few wafers at a time. Since all the wafers in a lot are moved in a single group there are several rules which are applied to maintain this organization. These rules will be discussed later.

Metrology equipment gathers data about the wafer such as layer thickness, width, and undesirable formations of features on the surface. Equipment measuring undesirable features on wafers are also known as *particle* or *defect monitors* [4]. The data so gathered may or may not be used further to control the properties of the line. Note that not all wafers have to undergo such measurements. Depending on a particular fabrication line policy only a fraction of wafers may be examined. The control of wafer flow depends on the process specifications and will be discussed in the next section.

Occasionally, the equipment used in wafer fabrication may malfunction or break down. In this case, the particular piece of equipment may have to be taken off-line and

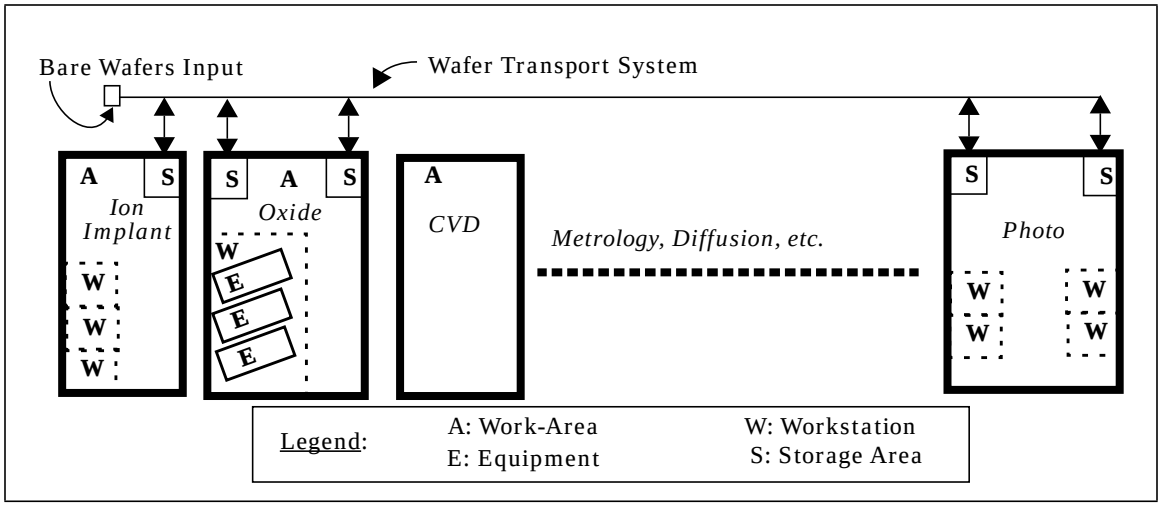

Figure 2.4 Work-Areas, Workstations, Equipment and Storage Areas.

undergo repair. The amount of time a given piece of equipment is available for processing is known as *uptime*. Not all of the uptime is devoted to processing and some *idle time* is thus inevitable. The fraction of the uptime that a piece of equipment is devoted to processing is known as *utilization* of the equipment and it serves as an indicator of factory performance [3].

2.2.2 Process Recipe and Flow of Wafers

A process recipe is defined as a sequence of steps to be performed to fabricate a product with a given technology such as CMOS, BiCMOS, DRAM, etc. Note that a single recipe may be shared by many different kinds of products and there may be more than one process recipe defined for a wafer fabrication line. A workstation is defined for each step of the recipe where a wafer lot can use any of the pieces of equipment belonging to the workstation to perform the process step. Each step is also associated with process specification like time required to perform the particular step, temperature and gas pressure settings, etc. In the case of metrology equipment, the specifications take the form of the parameter that needs to be measured (thickness, width, etc.) and the settings of equipment.

Two apparently different process recipes may in fact share the same steps and in these cases the specifications for the shared step must be the same. Thus, if the step uses batch equipment then several lots from different products can be possibly loaded into the same equipment. In practice, however, there may be operating rules in effect which do not allow such mixing of products. The important point to note here is that both the products can be affected in a very similar manner whenever there is some overlap in the process recipes. Figure 2.5 illustrates this overlapping of steps graphically showing that both work-station and specifications must match for steps to be considered equivalent. This is usually possible for processing steps of similar layers like metal interconnects.

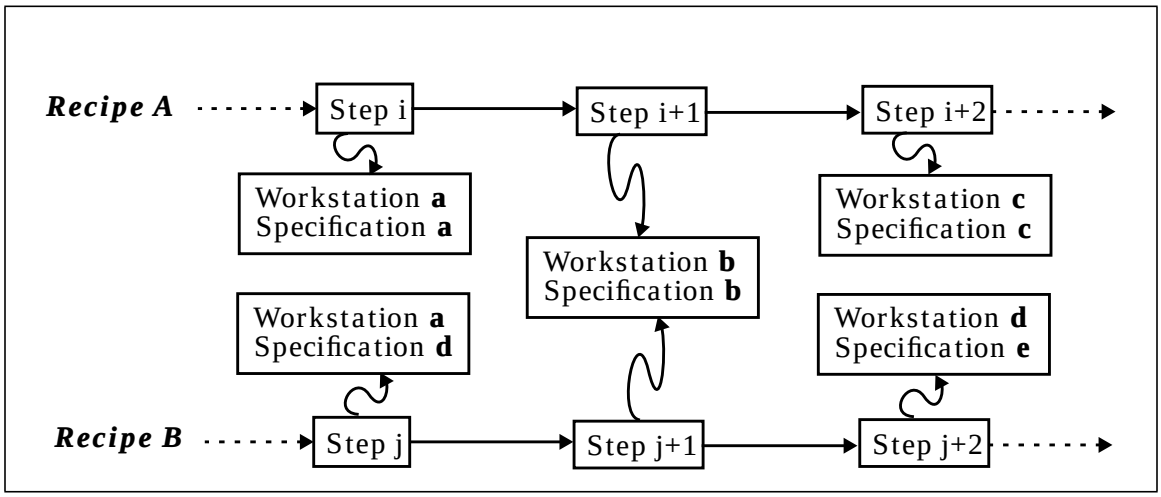

Figure 2.5 Overlap of process steps.

Quite often, several different steps of a particular recipe share the same workstation, a frequently occurring example being the photo-lithography step. Generally speaking almost all the mask exposure steps are conducted in the same workstation. Such steps are known as *re-entrant steps*, as is illustrated in Figure 2.6. Thus a wafer can be exposed to the same environment more than once in the same piece of equipment although a different layer is affected each time. In this case, although a work-

station is shared the process specifications are necessarily different since two unique layers are defined at each of the two steps. From a operational viewpoint, the scheduling task becomes complex and it will be discussed later.

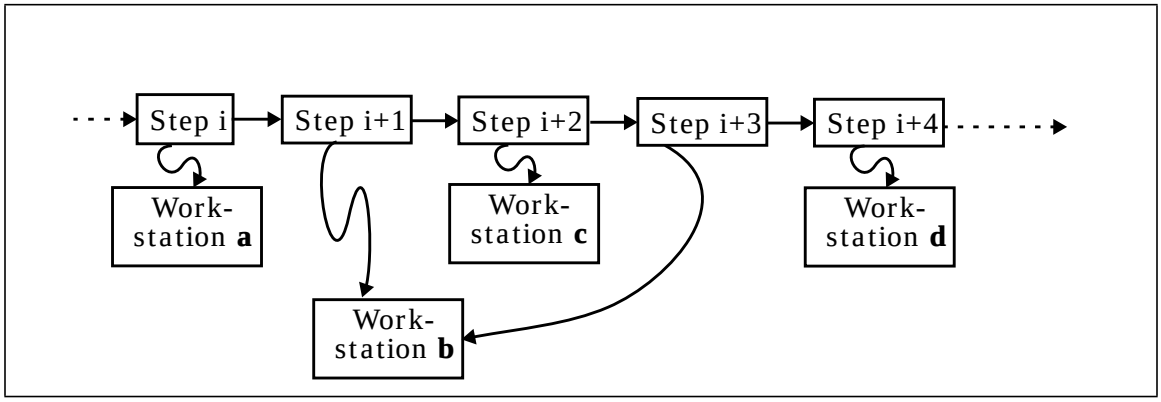

Figure 2.6 Re-entrant steps.

The control flow of wafers through the metrology step is different since not all wafers are sampled for in-line measurements (Figure 2.7). When the measured factor for the wafers sampled are within specifications then the entire lot is accepted and sent for further processing. Some wafers can be outside the acceptance limits but still within reasonable bounds so that they can be corrected. The corresponding lot is sent to appropriate steps for reworking in order to correct the problem. In the case when the observed problem cannot be corrected the all wafers in the lot are rejected.

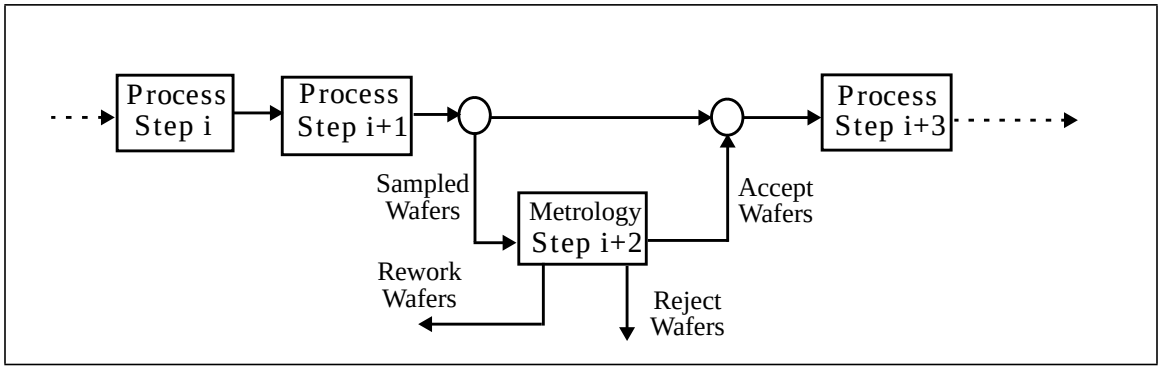

Figure 2.7 Wafer flow control in metrology step.

2.2.3 Product Attributes

A product is identified by two factors: the design of the IC and the process recipe that must be used to fabricate it. IC design is represented geometrically in terms of layout of the different layers (diffusion, gate, polysilicon, metal, etc.) and is physically translated into *masks* to be used during the photo-lithography steps. A typical CMOS process may require about 13 masks and a DRAM process may require about 17 masks.

From the point of view of technology, the other important attributes of products are *minimum feature size, die size* and *wafer size*. Feature size is defined by the minimum achievable length of the polysilicon over the gate oxide. Die size depends to a large extent on the aperture of the stepper and to a lesser extent on the complexity of the IC design and the minimum feature size. Wafer size is limited by the design of the equipment used for processing. A current state-of-the-art fabrication line can achieve 0.35μ gate-oxide length, about 1-2 cm^2 die size using 200 mm wafers. Die size and wafer size are important aspects of a fabrication line since the productivity of the line depends on them.

From an operational perspective, one has to consider the desired rate of production of a given product. This is determined by the input feed rate of wafers which is expressed as a constant number of *wafer starts per week* (WSPW). In a single product factory operating at a certain capacity, the WSPW value is also the *production volume* of the line. In a multi-product factory the production volume is the sum of the WSPW values for each product. The relative WSPW values of each product defines another important aspect referred to as *product mix*. As expected, operation of a multi- product and multi process factory is generally more complex and some of its aspects will be discussed later. A DRAM fabrication line is usually dedicated to a single product or different versions of the same product. In an ASIC (Application Specific IC) line, there can be hundreds of different products and several different process recipes.

2.2.4 Factory Personnel

Human assistance is necessary in a wafer fabrication line from time to time even though most of the operations in a modern facility are automated. Broadly, the role of factory personnel can be classified into four categories: wafer transportation, equipment operation, inspection and maintenance. Wafer transportation involves moving each lot from an output queue of one workstation to the input queue of the workstation for the next step in the process recipe for the product. When a piece of equipment is available for processing, the settings of the equipment (e.g., pressure, time, temperature, etc.) and the resources (e.g., chemicals, masks, etc.) may need to be changed. This is referred to as equipment *setup*. Two other functions of personnel are to *load* and *unload* the wafers in the equipment as it becomes necessary. Inspection of wafers at any metrology step, where human judgement is necessary for operation, can be performed by trained personnel. Maintenance personnels' main function is to inspect and repair various equipment in the event of malfunction or breakdown.

Fabrication line personnel can be viewed just like any other finite and limited resource of the factory. The efficiency of operations of the factory thus depends on the availability of the personnel at the right place in time. However, a large proportion of the functions performed by the personnel on a regular and repetitive manner is being automated to reduce errors.

2.2.5 Operating Rules

In order to understand the nature of operating rules, it is convenient to look at the timing of the wafer movement in the factory. The total time spent from the input of bare silicon to the completely fabricated wafer is an important factory performance indicator [3, 5]. This *cycle time* is composed of the actual time a wafer spends being processed (*raw processing time* or RPT), and the *waiting time* in between processing steps. Typically, RPT can range from 250 to 350 hours for a state-of-the-art process.

On the other hand typical cycle time can be 4 to 6 times the RPT value depending on the extent of automation in the line. The waiting time is primarily composed of three components: waiting for next available fabrication personnel to be transported to next equipment, in transit between the equipment, and finally waiting for the next equipment to become available for further processing. The secondary components of this waiting time are times spent in equipment setup, and loading and unloading of wafers. Figure 2.8 illustrates these timing components between two consecutive steps. The role of operating rules is to reduce and control the waiting time of the wafers in a factory [5, 6, 7, 8, 9, 10].

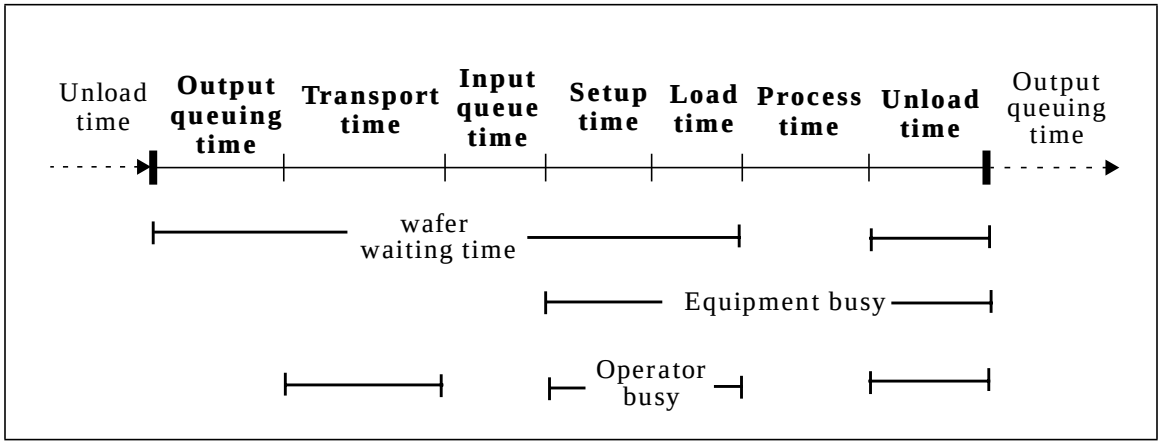

Figure 2.8 Timing of wafer movement between steps (not to scale).

Three kinds of operating rules are used in a factory: wafer release rules, scheduling rules, and rules for in-line metrology [2, 3, 6]. Wafer release rules control the way in which different products are released into the factory. The most common mechanism is to release products in single lots at an uniform rate. When production is ramped up, the input rate can also change with time. Quite often the input feed rate can be guided by other criteria such as expected due dates of the product, and inventory size. Inventory size as measured by the number of *work in progress* (WIP) may need to be kept

within a reasonable upper bound [11]. Excessively large WIP can cause waiting times to increase, causing cycle time to increase too.

Scheduling rules of wafers can be further classified as *load rules*, *setup rules* and *lot dispatch rules*. Load rule is applicable only for equipment whose batch size is greater than one lot. *Full load required rule* requires that whenever a piece of equipment is free it must be loaded up to its full capacity. This means that the equipment may have to wait until enough lots are available in the input queue. On the other hand, one can also employ the *partial load allowed rule* in which case equipment waiting time is reduced but this may decrease effective utilization [12].

Setup rules are required for re-entrant steps and multi-product factories where a workstation may be shared by different steps and products. Setup rules define when the operating conditions (settings) of a piece of equipment are changed. Ideally, one would like to minimize the number of *setup change-overs* since they cost time and introduce errors. Thus setup rules tend to select those lots which require exactly the same settings as the previous ones preferentially over others. However, this ad hoc rule can cause some lots to remain in the input queue for excessively long periods of time. In that case, exceptions to the rule are made based primarily on the waiting time of each lot in the queue.

Lot dispatch rules determine the ordering of the lots in an input queue. The simplest one is first-in first-out (FIFO) where the lot with the highest waiting time is given the highest preference. But such a rule may not always be optimal. Sometimes lots with smaller remaining processing times are given preference [2]. Due date of a particular product, and the size of the next queue may also be used as ordering criteria. By far the most important over-riding criterion is the presence of certain lots which are pre-assigned a high priority or *hot lots*. Where hot lots are present, these are given priority over any other lots and over-ride any other rules such as setup and load rules [13].

2.3 Yield Loss in Fabrication Process

As mentioned in Chapter 1 there are two kinds of yield losses one needs to consider namely, throughput yield loss and die yield loss. The causes of throughput yield loss can be further elaborated upon based on the discussion of wafer fabrication process presented in the last section. Most of the throughput yield loss can be attributed to mis-processing which can be due to equipment breakdown/malfunction during processing, incorrect equipment settings during a setup change-over, or lots being sent to a wrong workstation. These disturbances can result in an entire lot to be rejected. Individual wafers may break because of physical stress or improper handling. Lastly, any metrology step in the process may result in some or all of the wafers in a lot being rejected based on observed results. Reworking of wafers affects line capacity but usually leads to increased cycle times without affecting throughput.

In this section, the focus will be primarily on discussing die yield loss related to contamination. First, various sources of contaminants or particles during the wafer fabrication phase and their properties will be presented. Then the relationship between contamination, defects, and faults will be discussed. Finally, the effect of contamination rates on die yield will be presented and some methods of estimating yield will be analyzed.

2.3.1 Sources and Types of Contamination

Contamination or unwanted particles deposited on the wafer surface can come from a number of sources during fabrication. Broadly, they can be classified as originating from the environment, factory personnel, and equipment [14, 15, 16, 17, 18, 19, 20, 21, 22]. Particles in the environment can be introduced through the air supply system. Modern factories use a variety of schemes to eliminate these particles or reduce exposure of wafers to the ambient environment. A fraction of particles originate from the personnel handling the wafers. Currently, more and more factories are implementing

schemes to minimize human contact by automating wafer transport and equipment setup, loading and unloading of wafers. Therefore, in a modern factory, particles originating from equipment are the most dominant source of contamination since they are the most difficult to eliminate completely [14, 19, 23].

Particles in a piece of equipment can be introduced by many mechanisms such as [14, 24, 25]:

1. defective or leaky equipment introducing unwanted material,
2. repeated use of equipment without preventive maintenance causing material build-up inside chambers,
3. contaminated gases and chemicals used, and materials deposited on wafer surface that get dislodged and re-deposited.

Particles can be solid or liquid in nature and of arbitrary shapes and sizes ranging from sub-micron spherical objects to long strands of material covering a large area. During high temperature steps some of the particles may evaporate and some, depending on their adhesive properties, may get removed from the surface of the wafer in subsequent steps such as cleaning steps. Resistivity of the particles is another important aspect for consideration since they can directly alter local electrical connectivity. Thus, the physical, chemical and electrical properties of particles are important factors in determining the impact of contamination [24].

2.3.2 Contamination, Defects and Faults

Particles once deposited on the surface can lead to the formation of permanent features in the layers being defined in subsequent steps. These undesirable deformations or *spot defects* do not necessarily have the same shape and size as the original particles [14, 24, 26, 27, 28]. An example of *particle to defect transformation* is shown in Figure 2.9. In this figure, an opaque particle is assumed to be deposited on the photo-resist before the resist exposure step for defining the metal layer. The exposure step requires

a mask for the entire layer but only a portion of it is shown in the figure. The metal pattern after exposure, resist bake and etch, and metal etch is shown in the right hand side of the picture. The particle here leads to the formation of an extra metal defect as shown.

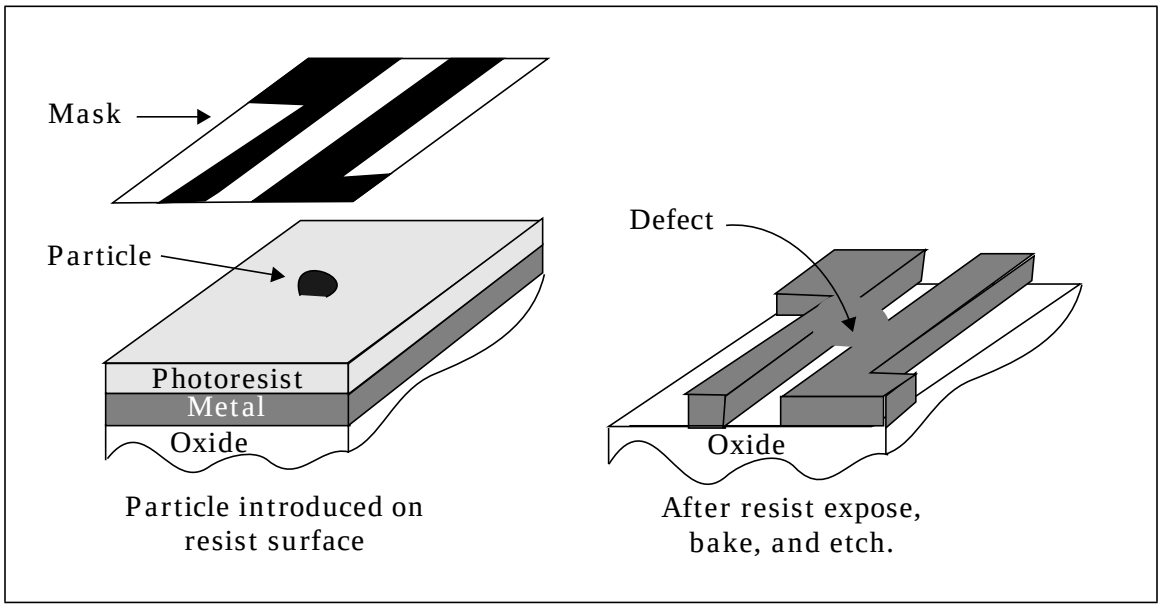

Figure 2.9 Particle to defect transformation.

Note also that the defect causes the two parts of the metal patterns to be connected to each other. These two metal patterns are most likely to be intended to be electrically disconnected and, thus, in this case the defect causes a *short* between two nodes of the intended circuit. A short can cause the circuit to malfunction under certain circumstances and in this case it is referred to have caused a *fault* - an example of *defect to fault transformation*. A different kind of particle could lead to a missing material defect which in turn could give rise to an open in a electrical net in the circuit. A more detailed description of this transformation of contamination to defects and ultimately to faults appears in [24, 25].

The type, size and location of a defect on the layout are the primary determining factors for a defect to cause a possible short or an open in a particular layer [29, 30, 31,

32, 33, 37]. There can be many defects in the IC, yet only a fraction of the defects actually may lead to a fault. In general, one can say that different defects can lead to different types of faults in a circuit depending on their type, size and location. In fact, defects of the same type (extra metal, for example) can actually lead to different faults. Moreover, for electrical nets which span a number of physical layers of an IC, one particular fault could be caused by many different types of defects.

This argument can be extended to the relationship between contamination and defects. Different sources of contamination may give rise to the same defect type. This happens when particles introduced at steps which logically define the same IC layer. For example, particles from any of the resist spin, expose, bake, etch, and metal etch steps could have lead to the extra metal defect. On the other hand, a particular type of particle could result in different types of defects. A particle deposited on the photoresist, as in Figure 2.9, could result in extra material defect in polysilicon, metal1, metal2, etc., depending on the step at which the particle has been deposited. Figure 2.10 summarizes the relationship between contamination, defects and faults as discussed above. Each directed edge in the figure represents a trace from the source of a type of particle to fault type.

The process of formation of defects, as presented above, is a simplification of a number of complex interacting phenomena occurring in a fabrication process [24] which, in reality, could produce defects in a number of fashions. A particularly important aspect is the class of defects formed from other defects in the IC structure. An example of such a transformation is shown in Figure 2.11. In this example, it is assumed that a small defect is first formed on top of the oxide layer. Then a metal layer is deposited on top of the oxide layer which conforms to the surface profile. The result is that a deformation forms in the metal layer which is larger than the original defect. Examples of resultant failures could be the formation of a possible electromigration site, opens in

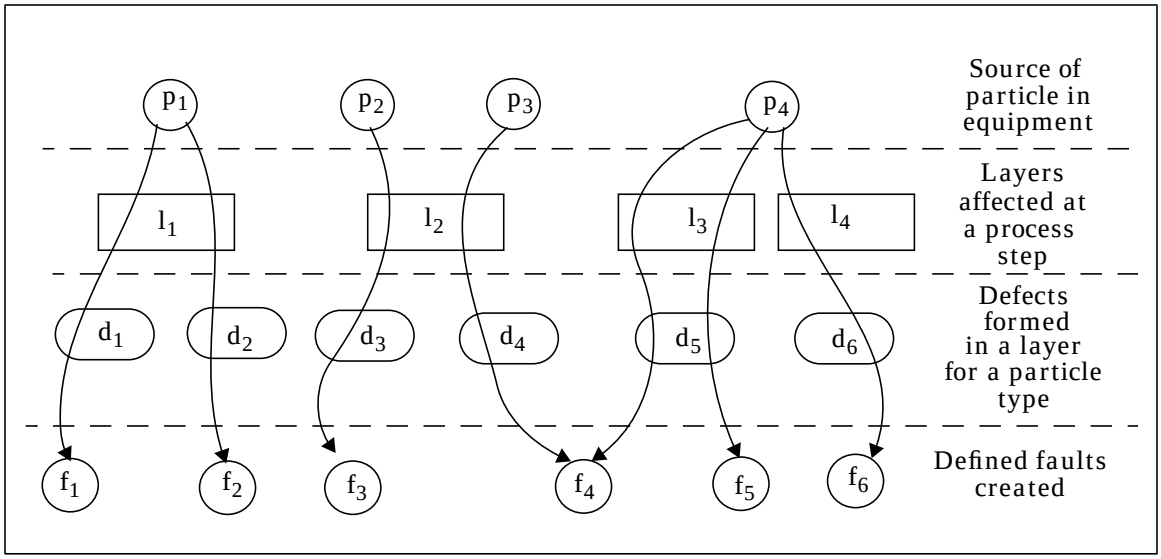

Figure 2.10 Relationship between contamination, defects and faults.

subsequent layers, etc.[24]. If the oxide layer is planarized [34, 35, 36] before depositing metal an unwanted contact can form.

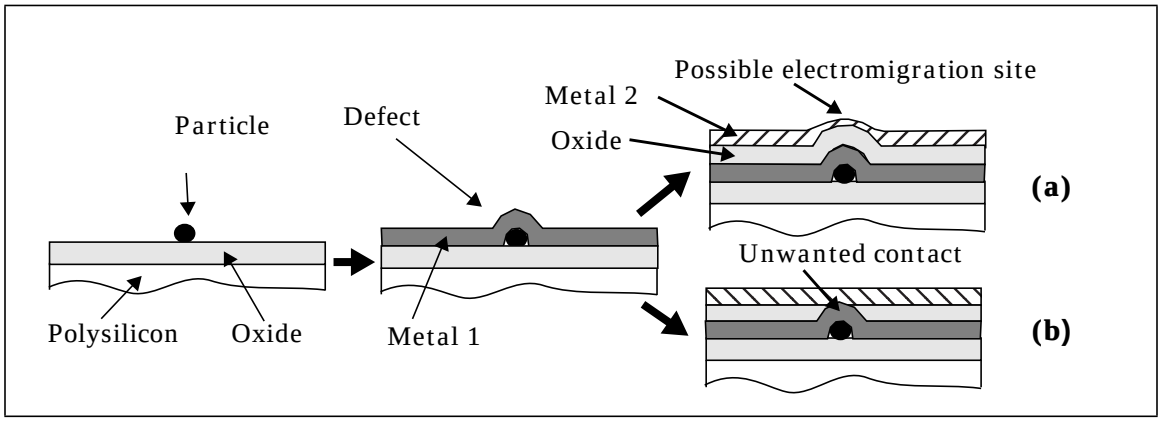

Figure 2.11 Defect propagation in IC's (a) normal processing leading to electromigration site and (b) with planarization of oxide layer leading to unwanted contact.

Lastly, one must note that there are multiple definitions of IC faults which frequently overlap. For example, in memory designs, a fault could be defined in terms of electrical nets such as a bit line to ground short or a word line to bit line short. For the same memory design, one could also consider more abstract fault definitions such as

row or column failures. The two examples above are useful from testing and redundancy calculations points of view, respectively. From a yield loss point of view it is often more convenient to define faults as shorts or opens in a single layer, shorts between layers, missing contacts, etc. The implication of this is that one needs to understand the relationship shown in Figure 2.10 with alternate definitions of faults.

2.3.3 Die Yield Models

Yield loss due to particles is defined as the average fraction of defective die per wafer. This is equivalent to defining yield loss as the probability of occurrence of at least a single fault in a die. The probability of occurrence of a fault depends not only on the types, rates and sizes of the particles deposited on the wafer surface but also on the attributes of each of the transformation processes from particles to defects and ultimately to the faults. Figure 2.12 summarizes the different transformation processes that one must consider to predict the probability of occurrence of a fault on a IC.

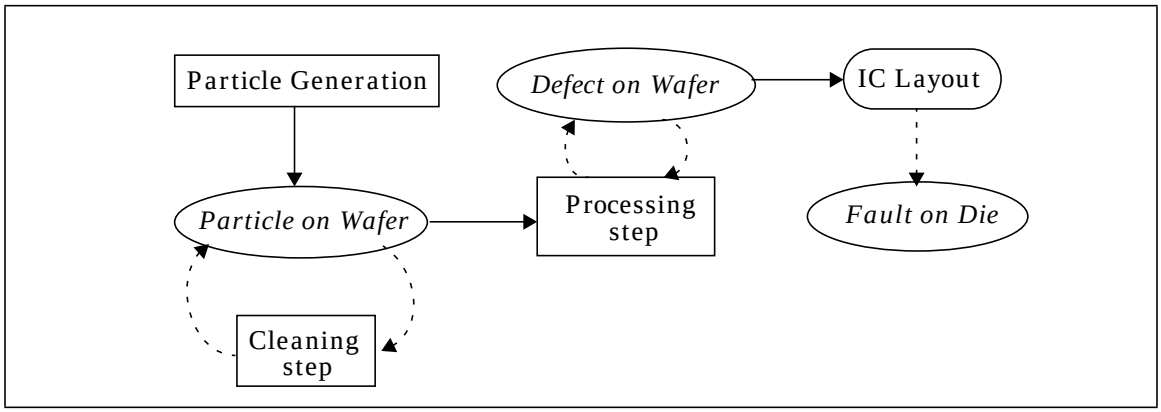

Figure 2.12 Particle and defect transformation processes in wafer fabrication.

Most of the research to date has been concentrated on the transformation of defects to faults and, thus, any quantitative analysis of yield loss is based on such formulations [14, 15, 23, 38, 39, 40]. This is mainly because of the fact that the defects are physically observable entities whereas the occurrence of particles on the wafer surface

is more difficult to observe. Recent advances in metrology equipment to monitor some of the particles on the surface of wafers is changing this situation [4]. Nevertheless, analyses based on defects provides an adequate starting point for predicting yield loss.

In the early 1960's, defects were assumed to be dimensionless points and any such defect within the area of an IC was assumed to cause a failure [41, 42]. Based on the point defect model, the Poisson model [41, 43] of yield was developed. A number of variations of this model were also developed and a complete description of these models can be found in [15, 30, 33, 38, 44]. Soon these models were found to be inadequate and a number of modifications were proposed subsequently. Modifications were made to assumptions underlying three aspects: number of defects on the wafer, the size of the defects and the dependence of fault occurrence on the IC layout.

The number of defects per wafer is characterized by a distribution function estimated from observed data. Some examples of distribution functions for defect density (number of defects per unit area) are shown in Figure 2.13 [41, 45, 46, 47, 48, 49]. This is a common way of modeling inter-wafer defect density variation commonly observed in industry [38, 50, 51, 52, 53, 54, 55, 56, 57]. Defects on the surface of the wafer are usually observed to be uniformly distributed but some researchers have observed spatial variations in the defect density [58, 59, 60]. One of the common observations is that defect density varies radially on the wafer. The important point to note is that one must consider the nature of defect density observed for a particular fabrication line in analyzing yield loss.

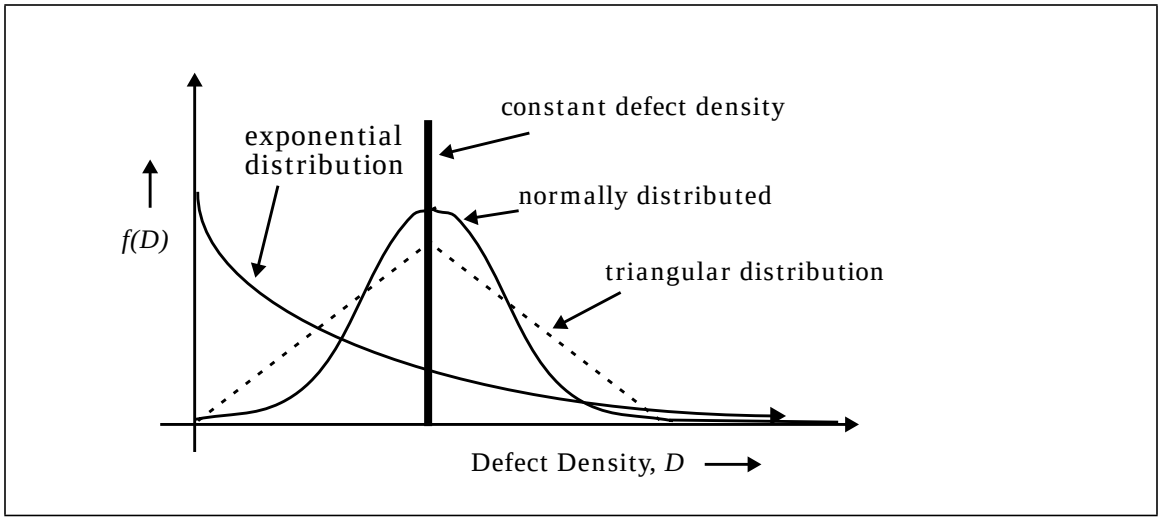

Figure 2.13 Defect density distribution functions.

It has also been observed that defects occur in various sizes that can characterized by a size distribution function. An example of a commonly used size distribution function is shown in Figure 2.14. The form of this function can be written as [24, 61]:

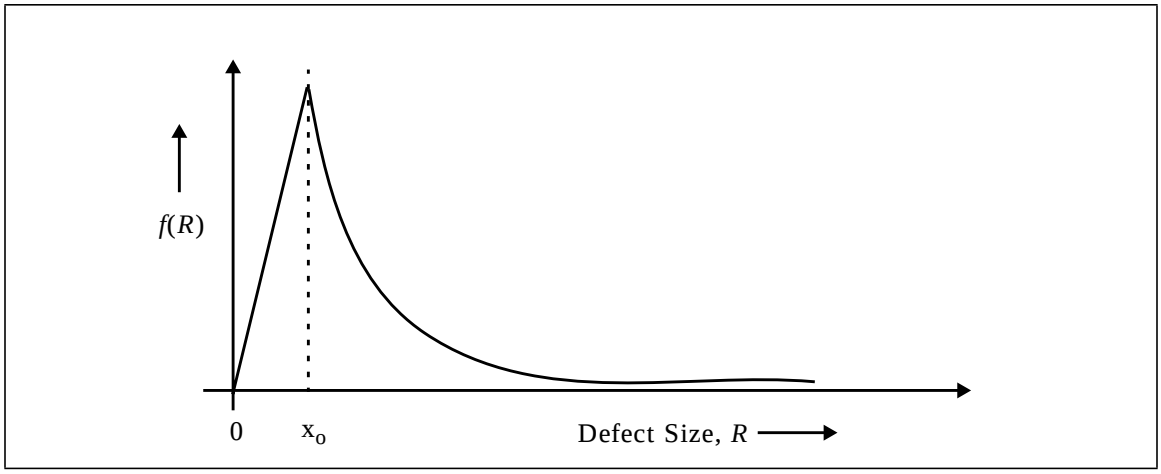

Figure 2.14 Example size distribution function.

$$f(R) = \begin{cases} \dfrac{2(p-1)R}{(p+1)x_o^2} & 0 \le R \le x_o \\ \dfrac{2(p-1)x_o^{(p-1)}}{(p+1)R^p} & x_o \le R \le \infty \end{cases} \quad (2.1)$$

where, R is the defect radius, x_0 and p are the parameters of the function $f(R)$. Other variations of this distributions, such as in [62, 63, 64, 65] have also been proposed. In these formulations, defects are viewed as circular two-dimensional deformations. From a practical standpoint, the minimum feature size is usually greater than x_0 and thus in the range of interest the probability of occurrence of a defect of a particular size, R, is inversely proportional to some power p of R.

For a given IC layout design, defects smaller than the minimum feature size obviously cannot cause any faults. Defects slightly larger than the minimum feature size can cause faults but they have to be located in such a way as to physically cause a short or an open as the case may be. Larger defects, on the other hand, are more likely to cause a fault in the IC. The dependence of layout sensitivity to the defect size is adequately captured by the *critical area* concept [29, 66, 67, 68, 69]. Critical area for a defect size, R, is defined as the area where the center of the defect must be located in order to cause a fault. For the case of a short between two metal lines, the critical area is illustrated in Figure 2.15. In the figure, the critical area is shown for three defect sizes in order to illustrate the greater sensitivity of the layout to larger defects.

Yield models based on the critical area concept were developed and extensively studied in the past [see e.g., 14, 15, 29, 38, 70]. In general, yield for a given defect type characterized by the size distribution $f(R)$ and mean defect density D_0 can be written as [29]:

$$Y = e^{-\left(D_o \int_s^{\infty} f(R) A_c(R) \, dR \right)} \quad (2.2)$$

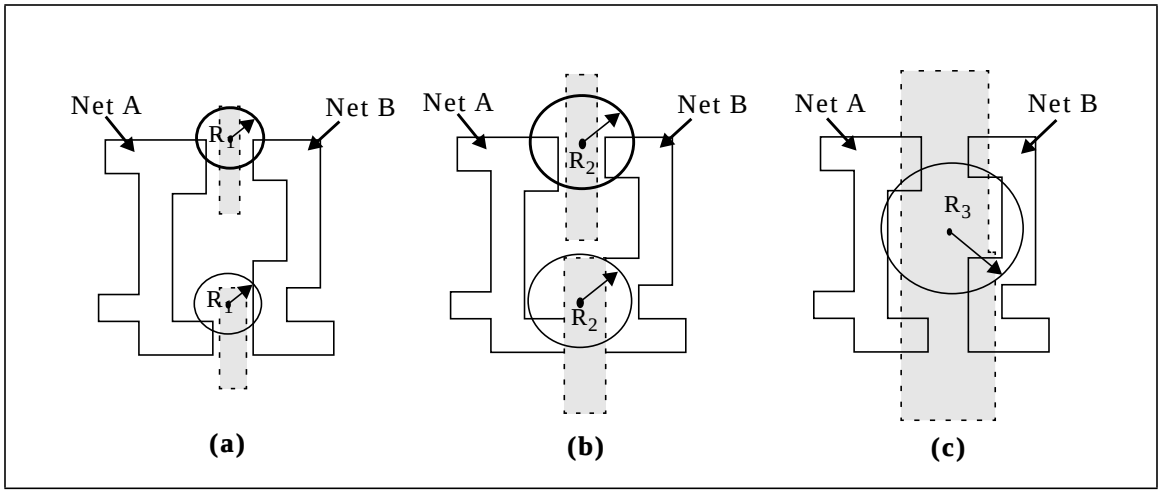

Figure 2.15 Critical area for metal shorts between two nets - (a) for defect size R_1, (b) for defect size $R_2 > R_1$. and (c) for defect size $R_3 > R_2$.

where, $A_c(R)$ is the critical area as a function of defect size and s is the minimum feature size of the layer. The above formula is a variation on the basic Poisson yield model. One can look at Equation 2.2 as partial yield, Y_i, for a given type of defect and fault type (considered in pair) i. The total yield, Y_{total}, can then be written as:

$$Y_{total} = \prod_{i=1}^{n} Y_i \tag{2.3}$$

where, *Yi* is the partial yield and is given by Equation 2.2. In the above equation the partial yields are assumed to be independent of each other. This is accomplished by selecting the defect types and fault types appropriately. However, this is only possible with *static yield estimation* where none of the particle and defect parameters change with time. As discussed later in this chapter, in a dynamically changing situation such an assumption may not be valid. One can also incorporate the defect density variation by appropriately compounding the exponent in Equation 2.2 as:

$$Y = e^{-\left(\int_0^{\infty} g(D) \left(\int_s^{\infty} f(R) A_c(R)\, dR \right) dD \right)} \tag{2.4}$$

where, $g(D)$ is the distribution function of defect density. The argument for two dimensional defects could be easily extended to three dimensional defects by defining *critical volume* instead of critical area[71]. The assumptions made above give only approximations of the yield and, hence, any method to accurately predict yield must use simulations with tools such as VLASIC [72] or DEFAM [33, 73] - two dimensional defect to fault simulators - or even more accurately with CODEF - three dimensional contamination to defect to fault simulator [24, 25].

2.4 Testing Process

During testing, every die on each wafer is tested using different types of tests applied in a predetermined sequence. These tests can broadly be classified into four types, applied in a sequence: parametric, basic functional, ac/delay and full functional tests. Parametric tests are performed on some of the specially designed test structures on each die (in the scribe area). These tests measure factors like threshold voltage, resistances, capacitances, etc. Functional tests, on the other hand, are performed to make sure that the fabricated die are operationally suitable for performing the ac or delay tests. The basic functional test ensures some of the functionality of the device under test. Delay testing of digital circuits is used to ensure that the timing of the signals meet the specifications. Full functional test involves subjecting the IC to a longer set of stimulus to ensure as much of the functionality of the IC as possible. This requires a set of predetermined stimuli to be applied to the IC and observing the response.

For large circuits, the set of stimuli required to completely test the circuit is also very large. Thus test sets are prepared in a way so as to cover as many failures as possible. This involves first evaluating the different ways failures can occur in the circuit and obtain a *fault list*. To accomplish this, one can tabulate the targeted faults manually or automatically using simulation. Creating such a fault list depends on the *fault*

model assumed. A number of fault models [see e.g., 74, 75, 76, 77] are described in the literature such as stuck-at fault model, the bridging fault model for shorts [78, 79], etc. The important point to note here is that the process of generating the set of input stimulus - or *test generation* process - also depends on the fault models assumed. For large digital circuits tests are sometimes generated automatically using automatic test pattern generators (ATPG) [81]. More detailed descriptions of testing and test generation can be found in [79, 80, 81, 82, 83, 84, 85, 86, 87]. There are three characteristics of a test set that must be considered: effectiveness of the test set, the time required to apply the tests and usefulness of the results in further analysis.

Effectiveness of a test set is usually measured in terms of an estimated metric called *fault coverage* [80, 82, 88, 89, 90, 91]. Fault coverage is a measure of the fraction of the total number of faults from the fault list that a test set can detect. Fault coverage is rarely 100% because of two reasons. First, some faults may be undetectable either because test generation is impossible or because the fault model is inadequate, and second, most ATPG tools have a prescribed time limit to find a possible test which when exceeded causes the computation to abort. Less than a 100% fault coverage naturally implies that some of the faults will be undetected [92, 93, 94]. The *escape rate* is a function of the detectability of a fault as well as the probability of occurrence of a fault. Probability of occurrence of a fault can be estimated [82, 83, 85] in a manner similar to yield computation by Equation 2.2, for example. There is another aspect of imperfect testing and that is identifying a perfectly functional die to be faulty. This is referred to as *false reject* (overkill) and is usually difficult to estimate.

One can look at the testing step in the same way as a fabrication equipment [3, 95, 96] from an operational viewpoint. Then the time interval can be divided in the same manner as in Figure 2.8. In the case of testing, equipment setup involves loading the right test set. Loading involves mounting the wafer (or packaged die as the case may be) on the tester. The difference arises in the actual time required to test a die. In the

case of wafer testing, each chip is probed with the help of a *probe-card*. The test sequence is applied and the test aborted when a fault is detected. Thus, the entire test set may not need to be applied and the time required to perform this step is dependent on the occurrence of faults. After a die is tested, the probe card is moved to the next die to be tested on the wafer till all the dies are tested. The defective die are later rejected when the wafer is diced into individual ICs. In the case of testing of packaged ICs, there are a few differences. First, instead of a probe card a *test card* is used. Second, the test sets used may be more extensive. Specifically, tests may be conducted with different supply voltages, and after subjecting ICs to elevated temperatures (Burn-in test). But such extended tests result in increased testing time and thus are only performed for a small sample of the die which pass the first sequence of tests. When yield is high these tests may be performed for most of dies and where performance is of importance (for DRAM, microprocessor, etc.) such extended tests may be performed on a larger sample. In the extreme case such as for military applications, these tests are performed for all the ICs.

In the stable production phase, the response of the chip to the applied test is only used to accept or reject the chip. But during the yield learning phase, it is important to identify the probable cause of the failures detected [97, 98, 99, 100]. This means that one must be able to analyze the test results in such a manner as to provide some clue about the nature of the failure. One mechanism is to classify (or bin) the test result according to the test set in which the fault is detected among all test sets. Test sets are usually designed to test particular partitions of a design for a large circuit. This helps to narrow the region of the location of the fault. In most cases, this is the extent to which a test result can be analyzed. One could, however, design the test set in such a way as to provide more diagnostic information [100, 101]. One could also analyze the test results in more detail off-line but this requires that the test response must be available for later analysis. Making test results available is a time consuming task

and is not a preferred option in a production setting. This aspect of usability of the test results has an important bearing on the efficiency of the defect diagnosis process and it will be discussed in the next section.

2.5 Failure Analysis Phase

Failure analysis can be broadly divided into two classes of activities in a manufacturing line: *in-line monitoring* of partially fabricated wafers and off-line *defect diagnosis* of completely fabricated wafers or packaged ICs. The nature of analyses are different for the two classes and together they are invaluable in controlling the contamination related problems in a manufacturing line. In this section, both the operational aspects and the effectiveness of these activities will be discussed.

2.5.1 In-Line Particle Monitoring

Particle monitors are employed at a number of intermediate steps defined in the process recipe. Operationally, there are treated as any metrology equipment as shown in Figure 2.7. Particle monitors scan the surface of the wafer under observation with laser beams. The light scattered and reflected from the surface of the wafer is received by one or more sensors as shown in Figure 2.16. The sensed light energy is analyzed using specialized tools (software and hardware) providing a variety of information on the particle or defect characteristics of the wafer under investigation. The type of information generated depends on the particular type of equipment which in turn governs the manner in which they are used [4, 102, 103, 104].

The simplest particle monitor is one which is able to scan only un-patterned wafers (wafers with no IC features on it). This type of equipment cannot distinguish between surface deformations (or defects) and IC features. They are usually very fast and take only a few minutes to scan a few square centimeters. To use this equipment bare wafers are introduced at some intermediate point in the process recipe. After each step

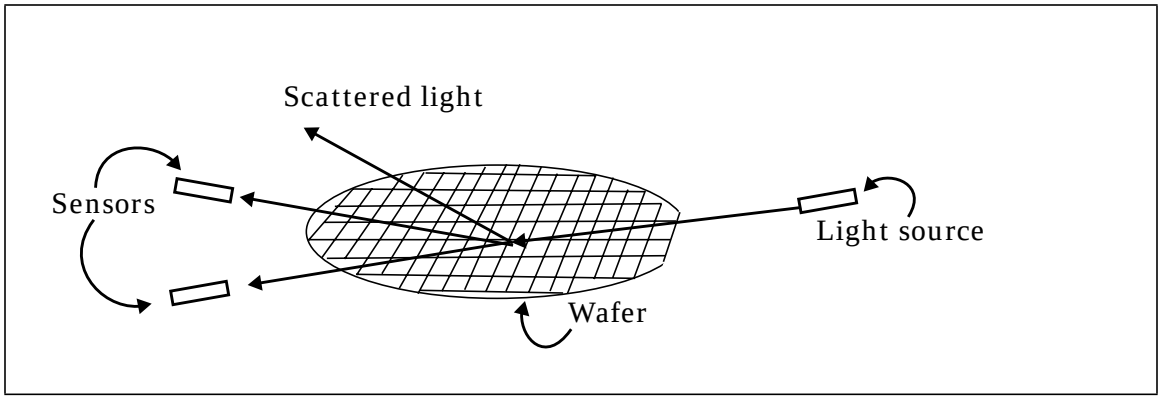

Figure 2.16 Operating principles of particle monitors

the surface is monitored for particles added at that step. For example, one can monitor particles added to surface after oxidation, metal deposition and resist deposition step in that order [102, 105]. This kind of characterization is useful for evaluating the relative contribution of each piece of equipment in terms of particles. This kind of characterization is referred to as *short loop* monitoring [105, 106]. The number of steps in a short loop is very small compared to the entire process recipe and is useful for quick feedback on the particle rates in the line. Figure 2.17 illustrates short loop monitoring in a manufacturing line.

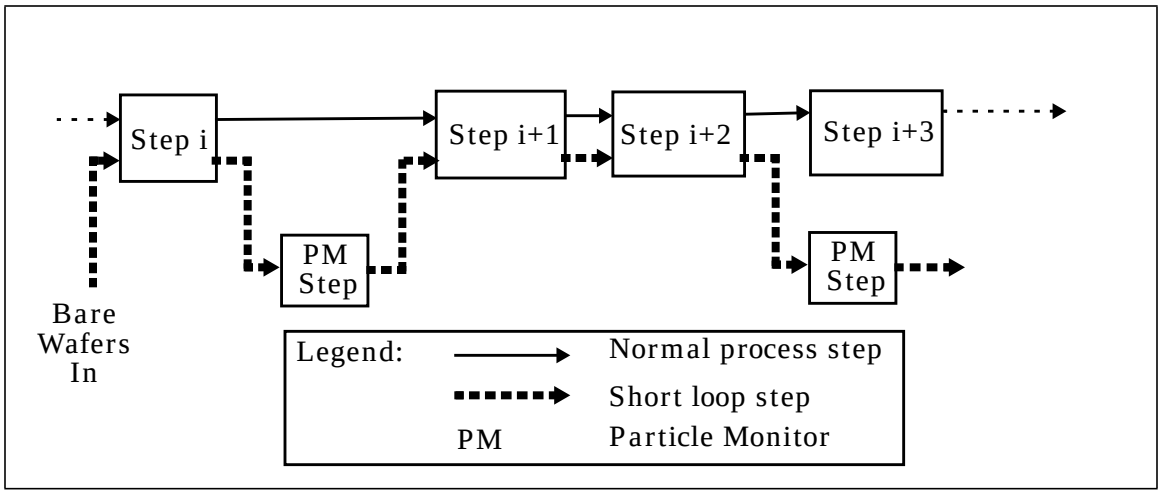

Figure 2.17 Short loop monitoring.

Monitors which can scan patterned wafers employ more sophisticated techniques for image processing and are thus more expensive and slower [102, 107, 108]. These types of equipment use one of two methods to distinguish between IC feature and particles (or defects). One method is to compare the same region of two or more adjacent IC to filter out the common features. The idea is that it is highly unlikely for two different ICs to have a particle in exactly the same location. The second, more complex, method is to compare the image obtained with the IC layout stored in the database. Such monitors have the advantage that the actual production wafers can be sampled and are thus likely to provide more useful data.

Not all particles on the wafer can be detected by particle monitors [102]. The detectability of particles depends on a number of interacting factors. The first is the size of the particle. Smaller particles, usually less than a micron, are very hard to detect. Secondly, the reflectivity of the surface below the scanned surface plays an important role. Surfaces with higher reflectivity provide more dependable results. Due to this particle monitoring is usually employed in the back end of the line where interconnects are defined. In the case of patterned wafer monitoring, the orientation of the IC with respect to the light source is an important determining factor. Certain orientations provide better resolution than others. There are other techniques to detect particles as well and these can be found in [109, 110, 111, 112].

Particle monitors can be used with different degrees of accuracy and efficiency. In the fastest mode, only the location can be obtained and that too with some error which can be quite large in some cases. In the most accurate slowest mode both the location and size can be obtained with a certain degree of accuracy. Further, one can perform computations to estimate the likelihood that the observed defect results in a fault. This could be achieved by estimating the probability of a failure given the location, size and the process step of occurrence of the observed particle. Due to inaccuracies in determination of these parameters such an estimation may not be possible. However,

one can trade off the accuracy against speed depending on how the data obtained is used. In the slowest mode, it can easily take 20-30 minutes just to scan a single die.

The throughput rate achievable depends on the sampling rate employed in the line. Sampling rules have three components. The first is the rate at which lots are sampled, in the extreme case, all lots may be used. The second is the number of wafers out of each lot selected for scanning which is usually about 3-4 wafers per lot. The third is the fraction of the area on a wafer actually scanned which is usually about 4-8 dies on a wafer. Depending on the particular sampling rules used and accuracy desired, the time required for one lot could be anywhere from 30 min to several hours. This is an important consideration since particle monitors are costly equipment and one may require many pieces of equipment to meet the effectiveness and throughput requirement [108, 113].

The manner in which the data is obtained from particle monitoring essentially decides its real effectiveness. From the point of view of controlling the manufacturing line, there are two aspects that need to be considered: first, to filter out those wafers which have excessively high probability of having low yield and second, to correct the situation by finding the source of the problem.

Filtering or screening of wafers can be accomplished by setting a threshold function. Let us assume that only the number of particles/defects is available from the monitoring equipment. For a given product one can set a limit on the number of particles which when exceeded, the wafer is rejected. If wafers are rejected early enough in the process, then there can be some saving in the costs arising out of subsequent processing steps. But one has to quantify the relationship between the number of particles observed and the expected yield. A large number of particles does not necessarily mean a low yield since the sizes of these particles also play an important role. Of course, if the threshold is set high enough then the risk can be low. For products with very low

profit margin, such a strategy can result in some cost savings and may even be desirable.

Corrective actions can be initiated on pieces of equipment suspected of introducing the particles at a higher rate than the set limits [115, 116]. Locating the piece of equipment responsible is a relatively simple task since, in all likelihood, only the equipment used in the previous step is the source. Care must be taken, however, to not react too quickly in correcting the situation since, the observed phenomenon could be a stray incident. When excessive number of particles are observed consistently over a few lots only then should corrective actions be attempted. The nature of corrective actions and their effects will be dealt with later in this chapter.

The use of data generated out of particle monitors may not be limited to controlling the line alone. One can possibly correlate all the data gathered on a particular wafer with the results of testing and generate a list of likely candidates for particles/defects for each defective die on the wafer [39, 105, 115]. This can be useful for compensating the lack of information available out of the testing process and thus aid defect diagnosis. One can create trend charts for particles for each piece of equipment and use these to guide the defect diagnosis process also. Both of these aspects of using particle monitoring for aiding defect diagnosis are discussed in the next section.

2.5.2 Defect Diagnosis after Fabrication

In this section, the focus will be on defect diagnosis of fabricated wafers. Diagnosis of defects in packaged ICs is essentially similar to that of wafers and the important differences will be pointed out where appropriate. A single cycle of defect diagnosis begins with a sample to be analyzed and ends with identification of the equipment (or any other source) responsible for the observed defect.

As mentioned earlier, a small fraction of the wafers may be selected after probe testing for failure analysis using some sampling rules. These sampling rules could simply

be to select a wafer with the highest number of defective ICs. More complex sampling rules can be used when more information can be extracted from the probe testing results. For example, when the dies fail to pass a particular test suite more frequently than others, the corresponding wafers can be selected for failure analysis.

Once a wafer is chosen for failure analysis, an attempt is made to diagnose the dominant cause of the faults found in some or all of the defective dies in the wafer. This is usually a three stage process: defect localization and identification, analysis of the particle causing the defect, and identifying the possible set of equipment as responsible for the particles [101]. Of these, the first step - defect localization and identification - is the most time consuming and uncertain, but it is a vital first step in successful analysis [117, 118]. Particle analysis is also time consuming, requiring very expensive equipment, but it may not be necessary where the shape, size, and location of the defect is enough to identify the particle source. In general, all three steps are essential for defect diagnosis to be efficient.

Defect localization is mainly achieved through direct observation methods using three types of microscopy: optical, scanning electron (SEM), and transmission electron microscopy (TEM) [119, 120, 121]. An optical microscope has a resolution of about 1μ and maximum achievable magnification about 1000X. The drawbacks of using an optical microscope are: the portion of the die to be searched has to be fairly small when dealing with small (< 2-3μ) defects which are unlikely to be located otherwise. Second, defects in lower levels of the IC (polysilicon for example) may be masked by the upper levels (metal for example), making it necessary to resort to higher magnification using scanning electron microscopy (SEM - with resolution better than 100 angstroms [120]). *Defect observability* could be greatly improved by selectively stripping away the layers [122, 123] and cross-section analysis using TEM. However, one could also strip away the defect in the process, rendering such analysis ineffective.

There are several other methods of defect localization using liquid crystal analysis, voltage contrast microscopy, electron-beam probing [124, 125, 126, 127], etc., requiring extensive sample preparation and full electrical connectivity. Hence these methods are more suitable for analyzing defects in packaged ICs where electrical analysis is simpler to perform. Also these methods tend to analyze only the surface or near surface defects, leaving the ones in the lower levels masked. Removal of certain layers may be necessary to improve "observability" [122] but it automatically excludes the possibility of an further electrical probing through normal contact pads on IC.

There are methods to electrically probe some of the internal nodes of the IC under test which include selectively etching away a small area of the IC to expose the metal lines for micro-probing [128, 129]. Sometimes, a small area is etched and filled with metal using focused ion beam (FIB) [130] techniques to reach a lower metal layer. These methods are time consuming and are applied when micro-probing is necessary.

One could also use analytical techniques on the test results to obtain tighter bounds on the defect neighborhood and reduce initial uncertainty [101, 131, 132, 133]. Such techniques, although very promising (and non-destructive), are not yet widely used in practice. One exception is the memory circuit where one can exploit the regularity of the circuit and layout to localize defects [134, 135, 136]. Here, the point in the sequence of test vectors at which the circuit fails indicates the locality of the defect. This is used to create *bit maps* which greatly speeds up the defect diagnosis process.

In summary, the efficiency and accuracy of the defect diagnosis process primarily depends on the defect size and the amount of uncertainty in the area and layer to be searched. The entire process of defect localization can be looked upon as a process of reducing this uncertainty which is primarily a function of the design of the product and the type of testing strategy used.

One outcome of this uncertainty is the rate at which different defect types are successfully identified. Defects in the top layer are, in general, easier to locate and iden-

tify than ones in lower layers. One can expect any easily diagnosable faults to be identified quickly and, as a consequence, the corresponding defects will "seem" to be the *dominant failure mechanism*. However, as the causes of the diagnosable faults are removed, the yield loss will be increasingly dominated by other less diagnosable faults. The dominant failure mechanism will keep shifting from one fault type to another. The situation could become more complex when defects in upper layers are caused by some other defects or particles in lower layers. This inter-dependence will be further elaborated in the next section.

Once a defect is localized one can possibly identify the step which introduced the particle causing the defect. Where this is not possible, more elaborate techniques can be used to identify the chemical composition of the particle using techniques such as energy dispersion spectroscopy (EDX) and wave dispersion spectroscopy (WDX) [118]. These methods require the preparation of thin cross sections of the sample which itself can cause enough damage to the IC to render any further analysis useless. Sometimes the energy of the scanning beam can cause the particle to disintegrate.

Once the origin of the dominant (or most frequently occurring) defect is narrowed down to a few steps in the fabrication process one has to pinpoint the source. It could be one of the materials used during processing or the equipment itself because of a leaky valve, contaminated chambers, etc. Usually a bank of several equivalent equipment are used in parallel to accomplish a given processing task (like resist spin-on) to meet the throughput requirement causing the source of the particle to remain ambiguous. To resolve this ambiguity one can resort to secondary sources of information like test structures [137, 138, 139], tracking records of the wafers analyzed, history of each piece of equipment, etc. Data from in-line particle monitors can be very useful in pinpointing the piece of equipment.

Let us define the throughput rate of defect diagnosis to be the number of successful diagnoses performed per unit time (one day, week, etc.). Treating the operation of

defect diagnosis equipment as any other fabrication or testing equipment the time components can then be divided as in Figure 2.8. However, the time spent in equipment for analysis is dependent on the defect characteristics of each die. Thus some defects can take very long to diagnose and some may not be diagnosable at all. In such extreme cases, the analysis may be aborted. Hence, observed throughput time is dependent on the relative proportion of diagnosable to undiagnosable faults in the dies analyzed.

Research in the area of particle monitoring and especially defect diagnosis has been limited to finding better mechanisms to perform the analyses. Currently no models or methods have been investigated to judge the efficiency and accuracy of these processes in the context of yield learning. The rate of yield learning depends to a great extent on the rate of correct feedback to the wafer fabrication process so that problems can be corrected. It also depends on the change in particle/defect characteristics as a result of corrective actions which is dealt with in the next section.

2.6 Corrective Actions in Manufacturing

Corrective actions are performed on the identified defect source when sufficient confirmatory evidence is collected for the source [140]. The frequency with which a particular resource is held responsible can also be used as a measure of the sufficiency of evidence. The corrective action could simply be setting the equipment correctly or changing certain simple parts. In this case, correction can be performed without disrupting the processing sequence. On the other hand, if the piece of equipment needs more complex repair or cleaning, it may need to be taken off-line.

Taking a piece of equipment off-line may not be preferred since this disrupts the flow of wafers by reducing the capacity of the line temporarily. In such a case, corrective actions may be performed when the equipment is idle (unlikely to be of sufficient duration) or, during the next scheduled maintenance cycle [3]. In addition, some policy

must be employed to ensure avoidance of bottlenecks due to many pieces of equipment being taken off-line. The duration of time required to correct the situation depends on the availability of maintenance resources and the type of problem being corrected. A major overhaul of the equipment can easily take several days, and cleaning of equipment, can take several hours.

After the corrective actions are applied and the equipment is put into normal operation, the observed defect characteristics like density, size distributions, etc., are expected to change so that the number of faulty chips per wafer is reduced. Depending on the particular particle source and type, the change in characteristics may be anywhere from little or no change to complete removal of the generating mechanism [140, 141, 142, 143]. For example, repairing a leaky valve can remove a particle source completely. On the other hand, cleaning a piece of equipment can only reduce the rate of introduction of particles causing faults in the IC.

The process of repairing or cleaning may affect the rate of introduction of particles of different sizes differently, for example, the rate of introduction of larger particles may be affected more than the smaller ones. This means that both the particle size and density distributions could be affected. One can describe the process of repairing/cleaning as a probability function which describes the probability of removal of a particle of a certain size. With this assumption one can estimate the new distribution of size and density by appropriately convoluting the distribution functions for particles with the probability function for cleaning [143]:

$$f_{new}(N, R) = f_{old}(N, R) \otimes g(N, R) \tag{2.5}$$

where, $f(N,R)$ is the joint probability distribution of number, N, and size, R, of occurrence of particles, and $g(N,R)$ is the joint probability distribution of cleaning the particles. Note that the number of particles, N, is simply related to the density, D, of the particles. In practice, however, joint probability distributions of particles are not available. Only observed characteristics for defects are available and in addition, the

density and size are assumed to be independent parameters - which is implicit in Equations 2.2, 2.3 and 2.4.

Let us for a moment concentrate on the particle-defect-fault relationship shown in Figure 2.10. A single type of particle can affect more than one layer, cause more than one defect and result in multiple faults. Any change in particle characteristics can result in more than one defect rate to change and thus affect the rate of observed faults of multiple types. The models presented for yield earlier are thus not directly applicable.

This inter-dependence also affects the relative rate of change in defect rates (as a result of failure analysis activities). For argument's sake let us assume that polysilicon defects also cause most of the defects in the metal layer. Since metal defects are easier to identify and locate, their source will be "discovered" quickly. In this case, if the cause is corrected then polysilicon defects will be reduced simultaneously. Such inter-dependence in the rate of change in defect attributes is not easy to capture using the yield models presented earlier.

2.7 Yield Forecasting - Discussion

Previous techniques to forecast yield were based on a macro view of the manufacturing line where the focus was on analytically describing yield as a function of time. Essentially, yield is expressed as a time series formula where at each time step, yield is expressed as a function of yield at the previous time step. Thus, yield Y_n at time step n can be written as [144, 145, 146, 147]:

$$Y_n = b\,Y_{n-1} \tag{2.6}$$

where, b is a constant (learning rate) greater than 1.0. In another method, a refinement was made to this by replacing yield with defect density in Equation 2.6 and then using an yield model to map defect density to yield [147].

The value of the learning rate, *b*, is assumed to be a parameter that can be extracted for a particular product and a given manufacturing line. But the learning rate of one product does not provide any knowledge of the learning rate for another product. Extrapolation of learning rates from historical learning curves can be difficult for a variety of reasons [148]. For example, failures can be more difficult to diagnose because of smaller feature size, larger die size, more interconnect levels, etc. Feedback cycles can be much longer because of increased cycle times, more types and sources of contamination, etc. Presence of shorter feedback cycles due to short loop monitoring can alter the learning rate. These are but a few of the reasons that prevent one from using such simple models in an effective way.

There is currently no methodology to model the yield learning process discussed in this chapter. In the next two chapters such a methodology and its corresponding models will be presented that effectively capture the important attributes of yield learning.

References

[1] P. K. Nag and W. Maly, "Yield Learning Simulation", *Proc. of SRC TECHCON 93*, pp. 280-282, Oct. 1993.

[2] *ManSim X*, User Manual, Tyecin Systems Inc., San Jose, CA, 1995.

[3] L. F. Atherton and R. W. Atherton, *Wafer Fabrication: Factory Performance and Analysis*, Kluwer Academic Publishers, 1995.

[4] L. Peters, "20 Good Reasons to Use In Situ Particle Monitors", *Semiconductor International*, pp. 52-57, Nov. 1992.

[5] S. C. H. Lu, D. Ramaswamy and P. R. Kumar, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semconductor Manufacturing Plants", *Trans. on Semiconductor Manufacturing*, vol. 7. no. 3, pp. 374-388, Aug. 1994.

[6] K. R. Baker, *Introduction to Sequencing and Scheduling*, Wiley, New York, 1974.

[7] L. M. Wein, "Scheduling Semiconductor Wafer Fabrication", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 3, pp. 115-130, Aug. 1988.

[8] C. Lozinski and C. R. Glassey, "Bottleneck Starvation Indicators for Shop Floor Control", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 4, pp. 147-153, Nov. 1988.

[9] C. R. Glassey and W. W. Weng, "Dynamic Batching Heuristics for Simultaneous Processing", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 2, pp. 77-82, May 1991.

[10] K. C. Saraswat, S. C. Wood, J. D. Plummer and P. Losleben, "Programmable Factory for Adaptable IC Manufacturing", *1993 Symposium on VLSI Technology, Digest of Technical Papers*, pp. 131-132, May 1993.

[11] C. Roger Glassey and M. G. C. Resende, " Closed-Loop Job Release Control for VLSI Circuit Manufacturing", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 1, pp. 36-46, Feb. 1988.

[12] H. Gurnani, R. Anupindi, and R. Akella, "Control of Batch Processing Systems in Semiconductor Wafer Fabrication Facilities", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 4, pp. 319-328, Dec. 1992.

[13] B. Ehteshami, R. G. Petrakian, and P. M. Shabe, "Trade-Offs in Cycle Time Management: Hot Lots", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 2, pp. 101-106, May 1992.

[14] W. Maly, "Computer-Aided Design for VLSI Circuit Manufacturability", *Proceedings of the IEEE*, vol. 78, no. 2, pp. , Feb. 1990.

[15] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated Circuit Yield Statistics", *Proceedings of the IEEE*, vol. 71, no. 4, pp. 453-470, April 1983.

[16] Peter Singer, "DI Water Filters: The Last Defense Against Microcontamination", *Semiconductor International*, pp. 77-80, Dec. 1994.

[17] M. Mishima, T. Yasui, T. Mizuniwa, M. Abe, and T. Ohmi, "Partcle-Free Wafer Cleaning and Drying Technology", *Trans. on Semiconductor Manufacturing*, vol. 2 no. 3, pp. 69-75, Aug. 1989.

[18] P. Borden, " The Nature of Particle Generation in Vaccum Process Tools", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 4, pp. 189-194, Nov. 1990.

[19] G. S. Selwyn, "The Uncoventional Nature of Particles", *Semiconductor International*, pp. 72-78, March 1993.

[20] Venu Menon, “SEMATECH’s Venu Menon Addresses Future Contamination-free Manufacturing Challenges”, *Semiconductor International*, pp.38, Jan, 1995.

[21] M. Watanabe, I. Kanno, and T. Ohmori, “Influence of particles / impurity metals in RCA Cleaning Solutions on Surface Contamination”, *Proc. of 1994 International Symposium on Semiconductor Manufacturing*, pp. 99-102, 1994.

[22] Peter Singer, “1995: Looking Down the Road to Quarter-Micron Production”, *Semiconductor International*, pp. 46-52, Jan, 1995.

[23] C. H. Stapper and R. J. Rosner, “Integrated Circuit Yield Management and Yield Analysis: Development and Implementation”, *Trans. on Semiconductor Manufacturing*, vol. 8, no.2, pp. 95-102, May 1995.

[24] J. B. Khare, *Contamination-Defect-Fault Relationship - Modeling and Simulation*, Ph.D. Dissertation, Carnegie Mellon University, Nov 1995.

[25] J. Khare and W. Maly, “Inductive Contamination Analysis (ICA) with SRAM Application”, *Proc. of Int. Test Conference*, pp. 552-560, 1995.

[26] L. Hecht, “A New Method to Determine Contamination Limited Yield”, *Trans. on Components Hybrids Manufacturing Technology*, vol. 14, no. 4, pp. 905-905, Dec. 1991.

[27] C. E. Novak, *Development of an Efficient 2D Lithography Simulator and its Application to Defect Analysis*, M.S. Thesis, Electrical and Computer Engineering Dept., Carnegie Mellon University, Pittsburgh, PA, September 1994.

[28] C. E. Novak, K. D. Lucas, Zhi-Min Ling, Andrzej J. Strojwas, “Lithography Simulation of Contamination-Caused Defects,” *Proceedings of SPIE*, vol. 2439, 1995.

[29] W. Maly, “Modeling of Lithography Related Yield Losses for CAD of VLSI Circuits”, *Trans. on Computer-Aided Design*, vol. 5, no. 3, pp/ 166-177, March 1985.

[30] D. M. H. Walker, *Yield Simulation for Integrated Circuits*, Kluwer Academic Press, 1987.

[31] C. H. Stapper, “Modeling defects in integrated circuit photolithographic patterns”, *IBM Journal of Research and Development*, vol. 28, no. 4, pp. 461-475, July 1985.

[32] C. H. Stapper, "Evolution and Accomplishments of VLSI Yield Management at IBM", *IBM Journal of Research and Development*, vol. 26, no. 5, pp. 532-555, Sept. 1982.

[33] D. Gaitonde, *Design and Application of a Hierarchical Defect to Fault Mapper*, Ph. D. Thesis, Carnegie Mellon University, February 1995.

[34] Victor Camello, "Planarization Using RIE and Chemical Mechanical Polish", *Semiconductor International*, pp. 28, March 1990.

[35] G. Nanz and L. E. Camilletti, "Modeling of Chemical-Mechanical Polishing: A Review", *Trans. on Semiconductor Manufacturing*, vol. 8, no. 5, pp. 382-389, Nov. 1995.

[36] P. E. Riley and E. D. Castel, "Planarization of Dielectric Layers for Multilevel Metallization", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 4, pp. 1154-1156, Nov. 1988.

[37] Takayuki Yanagawa, "Yield Degradation of Integrated Circuits Due to Spot Defects", *IEEE Trans. on Electron Devices*, vol. ED-19, no. 2, pp. 190-197, Feb. 1992.

[38] T. L. Michalka, R. C. Varshney, and J. D. Meindl, "A Discussion of Yield Modeling with Defect Clustering, Circuit Repair, and Circuit Redundancy", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 3, pp. 116-127, Aug. 1990.

[39] D. Betel, O. Bar-Ilan, and A. Burger, "Correlating Observable Defects and Yield", *Semiconductor International*, pp. 128-130, Oct. 1991.

[40] A. V. Ferris-Prabhu, "On the Assumptions Contained in Semiconductor Yield Models", *Trans. on Computer-Aided Design*, vol. 11, no. 8, pp. 966-975, Aug. 1992.

[41] B. T. Murphy, "Cost-Size Optima of Monolithic Integrated Circuits", Proceedings of the IEEE, pp. 1537-1545, Dec. 1964.

[42] T. Lawson, "A Prediction of the Photoresist Influence on Integrated Circuit Yield," *IEEE Journal of Solid State Technology*, vol. 9, no. 7, pp. 22-25, July 1966.

[43] R. M. Warner, Jr., "Applying a Composite Model to the IC Yield Problem", *IEEE Journal of Solid-State Circuits*, vol. SC-9, no. 3, pp. 86-95, June 1974.

[44] C. H. Stapper, "Fact and Fiction in Yield Modeling," *Microelectronics Journal*, vol. 20, no. 1-2, pp. 129-151, 1989

[45] R. B. Seeds, "Yield, Economic, and Logistic Models for Complex Digital Arrays," *IEEE International Convention Record*, pp. 60-61. March 1967.

[46] R. B. Seeds, "Yield and Cost Analysis of Bipolar LSI," *IEEE International Electron Devices Meeting*, October 1967.

[47] T. Okabe, M. Nagata, and S. Shimada, "Analysis on Yield of Integrated Circuits and New Expression for the Yield", pp. 135-141, Dec. 1972.

[48] C. H. Stapper, "On a Composite Model to the IC Yield Problem", *IEEE Journal of Solid-State Circuits*, pp. 537-539, Dec. 1975.

[49] C. H. Stapper, "Defect Density Distribution for LSI Yield Calculations", *IEEE Trans. on Electron Devices*, pp. 655-657, July 1973.

[50] A. V. Ferris-Prabhu, "A Cluster-Modified Poisson Model for Estimating Defect Density and Yield", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 2, pp. 54-59, May 1990.

[51] P. R. Pukite and C. L. Berman, "Defect Cluster Analysis for Wafer-Scale Integration", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 3, pp. 128-135, Aug. 1990.

[52] R. S. Collica, "The Effect of the Number of Defect Mechanisms on Fault Clustering and its Detection Using Yield Model Parameters", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 3, pp. 189-195, Aug. 1992.

[53] A. Tyagi and M. A. Bayoumi, "Defect Clustering Viewed Through Generalized Poisson Distribution", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 3, pp. 196-206, Aug. 1992.

[54] C. H. Stapper, "On Murphy's Yield Integral", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 294-297, Nov. 1991.

[55] C. H. Stapper, "Correlation analysis of particle clusters on integrated cicruit wafers", *IBM Journal of Research and Development*, vol. 31, no. 6, pp. 641-650, Nov. 1987.

[56] C. H. Stapper, "Yield Model for fault clusters within integrated circuits", *IBM Journal of Research and Development*, vol. 28, no. 5, pp. 636-639, Sept. 1984.

[57] C. H. Stapper, "The Effects of Wafer to Wafer Defect Density Variations on Integrated Circuit Defect and Fault Distributions", *IBM Journal of Research and Development*, vol. 29, no. 1, pp. 87-97, Jan. 1985.

[58] C. H. Stapper, "Statistics Associated with Spatial Fault Simulation Used for Evaluating Integrated Integrated Circuit Yield Enhancement", *Trans. on Computer-Aided Design*, vol. 10, no. 3, pp. 399-406, March 1991.

[59] A. Gupta and J. Lathrop, "Yield Analysis of Large Integrated-Circuit Chips," *IEEE Journal of Solid-State Circuits*, vol. SC-7, no. 5, pp. 389-395, October 1972.

[60] C. H. Stapper, "LSI Yield Modeling and Process Monitoring," *IBM Journal of Research and Development*, vol. 20, no. 3, pp. 228-234, May 1976.

[61] J. B. Khare, W. Maly and M. E. Thomas, "Extraction of Defect Size Distributions in an IC layer Using Test Structure Data", *Trans. on Semiconductor Manufacturing*, vol. 7, no. 3, pp. 354-368, Aug. 1994.

[62] A. V. Ferris-Prabhu, "Yield Implications and Scaling Laws for Submicrometer Devices", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 2, pp. 49-61, May 1988.

[63] R. Glang, "Defect Size Distribution in VLSI Chips", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 265-269, Nov. 1991.

[64] A. V. Ferris-Prabhu, "Role of Defect Size Distributions in Yield Modeling", *IEEE Trans. on Electron Devices*, vol. ED-32, no. 9, pp. 1727-1736, Sept. 1985.

[65] C. H. Stapper, "Modeling of Integrated Circuit Defect Sensitivities", *IBM Journal of Research and Development*, vol. 27, no. 6, pp. 549-557, Nov. 1983.

[66] W. Maly and J. Deszczka, "Yield Estimation Model for VLSI Artwork Evaluation", *Electron Letters*, vol. 19, no. 6, pp. 226-227, March 1983.

[67] P. K. Nag and W. Maly, "Yield Estimation of VLSI Circuits", *Proc. of SRC TECHCON' 90*, pp. 267-270, Oct. 1990.

[68] A. V. Ferris-Prabhu, "Modeling the Critical Area in Yield Forecasts", *IEEE Journal of Solid-State Circuits*, vol. SC-20, no. 4, pp. 874-880, Aug. 1985.

[69] J. Pineda de Gyvez and C. Di, "IC Defect Sensitivity for Footprint-Type Spot Defects", *Trans. on Computer-Aided Design*, vol. 11, no. 5, pp. 638-658, May 1992.

[70] W. Maly, H. T. Heineken, and F. Agricola, "A Simple New Yield Model", *Semiconductor International*, pp. 148-154, July 1994.

[71] J. P. de Gyvez, “Mound Defect Modeling in Yield Forecasts”, *Trans. on Semiconductor Manufacturing*, vol. 7 no. 4, pp. 430-439, Nov. 1994.

[72] D. M. H. Walker and S. W. Director, “VLASIC: A Catastrophic Fault Yield Simulator for Integrated Circuits,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 5 no. 4, pp. 541-556, October 1986.

[73] D. D. Gaitonde and D. M. H Walker, “Hierarchical Mapping of Spot Defects to Catastrophic Faults - Design and Applications”, *IEEE Trans. on Semiconductor Manufacturing*, vol. 8. no. 2, pp. 160-166, May 1995.

[74] J. A. Abraham and W. K. Fuchs, “Fault and Error Models for VLSI”, *Proceedings of the IEEE*, vol. 75, no. 5. pp. 639-654, May 1986.

[75] M. Syrzcyki, “Modeling of Gate Oxide Shorts in MOS Transistors”, *Trans. on Computer-Aided Design*, vol. 8, no. 3, pp. 193-202, March 1989.

[76] W. Maly, P. K. Nag, and P. Nigh, 'Testing Oriented Analysis of CMOS ICs With Opens”, *In Proceedings of Int. Conf. on Computer-Aided Design*, pp. 344-347, Nov. 1988.

[77] J. P. Hayes, “Fault Modeling”, *IEEE Design and Test of Computers*, pp. 88-95, April 1985.

[78] J. M. Acken, “Testing for Bridging Faults (Shorts) in CMOS VLSI Circuits”, *Proceedings of the Design Automation Conference*, pp. 717-718, 1983.

[79] P. J. Nigh, *Built-in Current Testing*, Ph.D. Dissertation, Carnegie Mellon University, 1991.

[80] M. Abramovici, M. A. Breuer, and A. D. Friedman, *Digital Systems Testing and Testable Desgn*, IEEE Press, New York, 1995.

[81] V. D. Agrawal and S. C. Seth, *Test Generation for VLSI Systems*, Computer Society Press, 1988.

[82] T. M. Storey, *Quality of Integrated Circuit Testing*, Ph.D. Dissertation, Carnegie Mellon University,1991.

[83] J. P. Shen, W. Maly and F. J. Ferguson, “Inductive Fault Analysis of MOS Integrated Circuits,” *IEEE Design and Test Magazine*, vol. 2 no. 6, pp. 13-26, December 1985.

[84] F. J. Ferguson and J. P. Shen, "A CMOS Fault Extractor for Inductive Fault Analysis", *Trans. on Computer-Aided Design*, vol. 7, no. 11, pp. 1181-1194, Nov. 1988.

[85] P. Nigh and W. Maly, "Layout Driven Test Generation", *IEEE Int. Conference on Computer-Aided Design (ICCAD)*, pp. 154-157, Nov. 1989.

[86] M. Jacomet and W. Guggenbuhl, " Layout-Dependent Fault Analysis and Test Synthesis for CMOS Circuits", *Trans. on Computer-Aided Design*, vol. 12, no. 6, pp. 888-899, June 1993.

[87] S. W. Bollinger and S. F. Midkiff, "Test Generation for IDDQ Testing of Bridging Faults in CMOS Circuits", *Trans. on Computer-Aided Design*, vol. 13, no. 11, pp. 1413-1418, Nov. 1994.

[88] F. N. Najm and I. N. Hajj, "The Complexity of Fault Detection in MOS VLSI Circuits", *Trans. on Computer-Aided Design*, vol. 9, no. 9, pp. 995-1001, Sept. 1990.

[89] A. D. Singh and C. M. Krishna, "On Optimizing VLSI Testing for Product Quality Using Die-Yield Prediction", *Trans. on Computer-Aided Design*, vol. 12, no. 5, pp. 695-709, May 1993.

[90] H. A. Farhat and S. G. From, "A Beta Model for Estimating the Testability and Coverage Distributions of a VLSI Cicuits", *Trans. on Computer-Aided Design*, vol. 12, no. 4, pp. 550-554, April 1994.

[91] R. L. Wadsack, "VLSI: How Much Fault Coverage is Enough?", *Digest of Papers, 1981 Int. Test Conference*, pp. 547-554, Oct. 1981.

[92] D. Gaitonde, J. Khare, D. M. H. Walker, and W. Maly, "Estimation of Reject Ratio in Combinatorial Circuits," *Proceedings of the 1993 VLSI Test Symposium*, pp. 319-325, April 1993.

[93] D. Feltham, J. Khare and W. Maly, "Design for Testability View on Placement and Routing," *1992 European Design Automation Conference*, pp. 382-387, Hamburg, Germany, September 1992.

[94] J. Khare, S. Mitra, P. K. Nag, W. Maly and R. Rutenbar, "Testability-Oriented Channel Routing," *Proceedings of the 8th Annual VLSI Design Symposium*, New Delhi, India, January 1995.

[95] *TestSim X*, User Manual, Tyecin Systems Inc, San Jose, CA, 1995.

[96] R. Uzsoy, L. A. Martin-Vega, C. -Y. Lee, and P. A. Leonard, "Production Scheduling Algorithms For A Semconductor Test Facility", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 270-280, Nov. 1991.

[97] W. Maly, B. Trifilo, R. A. Hughes, and A. Miller, "Yield Diagnosis through Interpretation of Tester Data", *Proc. of 1987 International Test Conference*, pp. 10-20, Sept. 1987.

[98] J. B. Khare, W. Maly, S. Griep, and D. Schmitt-Lansiedel, "Yield-Oriented Computer-Aided Defect Diagnosis", *Trans. on Semiconductor Manufacturing*, vol. 8, no. 2, pp. 195-206, May 1995.

[99] Y-J. Kwon and D.M.H. Walker, "Yield Learning via Functional Test Data", *1995 Proc. of International Test Conf*, pp. 626-635, 1995.

[100] W. Maly and S. B. Naik, "Process Monitoring Oriented Testing", *Proc. of Int. Test Conference*, pp. 527-532, 1989.

[101] S. B. Naik, *Computer-Aided Process Monitoring*, Ph.D. Dissertation, Carnegie Mellon University, Oct. 1994.

[102] *Close Up - Wafer Inspection*, Tencor Instruments, vol.1, no. 2, Spring 1995.

[103] Pieter Burggraaf, "Pursuing Advanced Metrology Solutions", *Semiconductor International*, pp. 62-64, April 1994.

[104] Pieter Burggraaf, "Defect Inspection: Wafers In, Process Control Out", *Semiconductor International*, pp. 58-62, Nov. 1991.

[105] D. Dance and R. Jarvis, "Using Yield Models to Accelerate Learning Curve Progress", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 1, pp. 41-46, Feb. 1992.

[106] J. A. Cunningham, "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 2, pp. 60-71, May 1990.

[107] Peter Burggraaf, "Patterned Wafer Inspection: Now Required!", *Semiconductor International*, pp. 57-60, Dec. 1994.

[108] R. K. Nurani, R. Akella, A. J. Strojwas, and R. Wallace, "Role of In-Line Defect Sampling Methodology in Yield Management", *1995 International Symposium on Semiconductor Manufacturing*, pp. 243-245, Sept. 1995.

[109] V. Murali, A. T. Wu, A. K. Chatterjee, and D. B. Fraser, "A Novel Technique for In-Line Monitoring of Micro-Contamination and Process Induced Dam-

age", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 3, pp. 214-222, Aug. 1992.

[110] N. Tokunaga, S. Okamura, S. Sasakin S. Nakamura, and F. Mieno, "Particle Count and Analysis by using a Cyclone Particle Sampler", *1995 International Symposium on Semiconductor Manufacturing*, pp. 178-181, Sept. 1995.

[111] P. G. Borden and L. A. Larson, " Benefits of Real-Time, In Situ Particle Monitoring in Production Medium Current Implantation", *Trans. on Semiconductor Manufacturing*, vol. 2 no. 4, pp. 141-145, Nov. 1989.

[112] Y. Ichikawa and J. I. Toriwaki, "ULSI Inspection System with Digital Scanning Confocal Microscopy", *1995 International Symposium on Semiconductor Manufacturing*, pp. 112-115, Sept. 1995.

[113] L. L. Pesotchinsky and Z. Fichtenholz, "Comparison of Two Wafer Inspection Methods for Particle Monitoring in Semiconductor Manufacturing", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 1, pp. 16-22, Feb. 1988.

[114] K. Mori, N. Nguyen, and J. Kantapit, "Process Equipment Particle Control with Inline Inspection", *1995 International Symposium on Semiconductor Manufacturing*, pp. 80-84, Sept. 1995.

[115] Y. Uraoka, I. Miyanaga, K. Tsuji, and S. Akiyama, "Failure Analysis of ULSI Circuits Using Photon Emission", *Trans. on Semiconductor Manufacturing*, vol. 6 no. 4, pp. 324-331, Dec. 1993.

[116] Pieter Burggraaf, "Failure Analysis: From 'Postmortem' to 'Preventive'", *Semiconductor International*, pp. 56-61, Sept. 1992.

[117] D. Corum, S. Y. Khim, and K. S. Wills, "Failure Mechanisms in Integrated Circuits", *Microelectronic Failure Analysis, Desk Reference, 3rd Edition*, pp. 277-300, 1995.

[118] T. W. Lee and Dr. S. V. Pabbisetty (Editors), *Microelectronic Failure Analysis, Desk Reference, 3rd Edition*, ASM International, The Materials Information Society, 1995.

[119] I. Banerjee, B. Tracy, P. Davies, and R. McDonald, "Use of Advanced Analytical Techniques for VLSI Failure Analysis", *Proc. of Int Reliability Physics Symp.*, pp. 61-68, 1990.

[120] E. I. Cole et. al., "Advanced Scanning Electron Microscopy Methods and Applications to Integrated Circuit Failure Analysis", *Scanning Microscopy*, vol. 2, no. 1, pp. 135-150, Jan. 1988.

[121] H. Seiler, "Secondary Electron Emission in the Scanning Electron Microscope", *J. of Appl. Physics*, vol. 54, no. 11, pp. R1-R18, Nov. 1983.

[122] G. Matuciewicz, "RIE for Failure Analysis", *Proc. of Int. Symp. of Test and Failure Analysis*, pp. 21-25, 1989.

[123] D. D'Agosta, "Non-destructive Passivation Deprocessing Using the RIE", *Proc. of Int. Symp. on Test and Failure Analysis*, pp. 257-260, 1989.

[124] C. A. Smith et. al., "Resistive Contrast Imaging: A New SEM Mode for Failure Analysis", *Trans. on ELectron Devices*, vol. 33, no. 2, pp. 282-285, Feb. 1986.

[125] E. I. Cole et. al., "Resistive Contrast Imaging Applied to Multilevel Interconnection Failure Analysis", *Proc. of IEEE VLSI Multilevel Interconnection Conf.*, pp. 176-182, 1989.

[126] W. Reiners et. al., "Electron Beam Testing of Passivated Devices via Capacitative Coupling Voltage Contrast", *Scanning Microscopy*, vol. 2, no. 1, pp. 161-175, Jan. 1988.

[127] E. I. Cole et. al., "A Novel Method for Depth Profiling and Imaging fo Semconductor Devices Using Capacitative Coupling Voltage Contrast", *J. of Applied Physics*, vol. 62, no. 2, pp. 4909-4915, Feb. 1987.

[128] D. S. Kiefer, "Window Etch Procedure for Multi-layer VLSI", *Microelectronic Failure Analysis - Desk Reference, 3rd Ed.*, ASM International, 1995.

[129] K. Hirose et. al., "Ion-Implanted Photoresist and Damage-Free Stripping", *J. Electrochemical Society*, vol. 141, no. 1, pp. 192-205, Jan. 1994.

[130] -, "Failure Analysis of Micron Technology Using Focussed Ion Beams", *Proc. of Int. Symposium on Testing and Failure Analysis*, pp. 249-254, 1989.

[131] R. C. Aitken, "Finding Defects with Fault Models", *1995 Int. Test Conference*, pp. 498-505, 1995.

[132] B. Chess, D. B. Lavo, F. J. Ferguson, and T. Larrabee, "Diagnosis of Realistic Bridging Faults with Stuck-at Information", *Int. Conf. on Computer-Aided Design*, pp. 185-192, 1995.

[133] J. A. Waicukauski and E. Lindbloom, "Failure Diagnosis of Structured VLSI", *IEEE Design and Test of Computers*, pp. 49-60, Aug. 1989.

[134] J. Khare and W. Maly, "Rapid Failure Analysis Using Contamination-Defect-Fault (CDF) Simulation", *1995 International Symposium on Semiconductor Manufacturing*, pp. 136-141, Sept. 1995.

[135] S. Griep, J. Khare, R. Lemme, U. Papenberg, D. Schmitt-Lansiedel,W. Maly, D. M. H. Walker, J. Winnerl and T. Zettler, "Speedup of Failure Analysis Using Defect Simulation", *Proc. of the 4th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF 93)*, Bordeaux, pp. , Oct 1993.

[136] S. Naik, F. Agricola and W. Maly, "Failure Analysis of High Density CMOS SRAMs Using Realistic Defect Modeling and Iddq Testing," *Special Issue on Memory Testing of the Design and Test of Computers Magazine,* March 1993.

[137] H. M. Chou, C. C. Liu, C. J. Kuo, J. H. Hwang, N. W. Wu, and M. C. Chang, "Defect Reduction Using Yield Management Test Structures", *Proc. of 1994 International Symposium on Semiconductor Manufacturing*, pp. 147-150, 1994.

[138] A. R. Comeau and J. Laneuville, "An Automated Electrical Defect Identification and Location Method for CMOS Processes Using a Specially Designed Test Chip", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 3, pp. 207-213, Aug. 1992.

[139] A. McCarthy, W. Lukaszek, C. -C. Fu, D. H. Dameron, and J. D. Meindl "A Novel Technique for Detecting Lithographic Defects", *Trans. on Semiconductor Manufacturing*, vol. 1 no. 1, pp. 10-14, Feb. 1988.

[140] A. Yamanaka, M. Mizuno, R. Ariyoshi, and O. Nozawa, "Low Cost and Flexible Data Analysis System to Find Appropriate Corrective Action for Yield Deterioration in LSI Manufacturing", *1995 International Symposium on Semiconductor Manufacturing*, pp. 103-106, Sept. 1995.

[141] R. A. Governal, A. Bonner, and F. Shadman, "Effect of Component Interactions on the Removal of Organic Impurities in Ultrapure Water Systems", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 298-303, Nov. 1991.

[142] M. Itano, F. W. Kern, Jr., R. W. Rosenberg, M. Miyashita, I. Kawanabe, and T. Ohmi, "Particle Deposition and Removal in Wet Cleaning Processes for ULSI Manufacturing", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 2, pp. 114-120, May 1992.

[143] M. Itano, F. W. Kern, Jr., M. Miyashita, and T. Ohmi, "Particle Removal from Silicon Wafer Surface in Wet Cleaning Process", *Trans. on Semiconductor Manufacturing*, vol. 6 no. 3, pp. 258-267, Aug. 1993.

[144] V. Ramakrishna and Jeanne Harrigan, "Defect Learning Requirements", *Solid State Technology*, pp. 103-105, Jan. 1993.

[145] J. A. Cunningham, "Using the learning curve as a management tool", *IEEE Spectrum*, pp. 45-48, June 1980.

[146] C. J. Teplitz, *The Learning curve deskbook: a reference guide to theory, calaculations*, and applications, Quorum Books, 1991.

[147] D. R. LaTourette, "A Yield Learning Model for Integrated Circuit Manufacturing", *Semiconductor International*, pp. 163-170, July, 1995.

[148] Robert Oliver and Kneale Marshall, *Decision Making and Forecasting*, McGraw Hill, 1995.

Chapter 3

Methodology to Predict Yield Learning Curves

In this chapter, a methodology for prediction of yield learning curves for a semiconductor manufacturing line is presented. This methodology attempts to mimic the yield learning process described in the previous chapter.

3.1 Yield Forecasting - An Overview

The yield learning process was described in the previous chapter as a sequence of events starting with the introduction of particles, followed by detection of defects and identification of their source, and concluding with eliminating the source of particles. The rate of yield learning depends on:

1. The complexity of relationship between particles, defects and faults;
2. Ease of defect localization which depends on:
   a. size, layer and type of defect,
   b. level of “diagnosability” of the IC design and,
   c. probability of occurrence of catastrophic defects;
3. Effectiveness of the corrective actions performed;
4. The timing of each of the events mentioned above;
5. Rate of wafer movement through the process.

The methodology to forecast yield as a function of time must, thus, be able to capture all of the above factors. The technique underlying this methodology is to mimic the manufacturing process using a simulator. In the next section, the basic properties of yield learning process which needs to be simulated is presented.

3.2 Characteristics of Yield Learning

In order to characterize the yield as a function of time, let us first concentrate on a single product manufacturing line. Let us also assume that only one of the pieces of equipment produces a single type of particle resulting in one type of defect. Understanding of this simplest possible case suffices to capture the essence of the yield learning process. The yield versus time curve for this scenario resembles the staircase function shown in Figure 3.1. Here, T_f is the time required for analysis and detection of the failure mechanism leading to process intervention. T_e is the time needed for process correction which decreases contamination levels and the time required for the new process parameters to be effective. T_r is the time interval after which a change in yield of the fabricated wafers is observed following process corrections. The total time required for yield change to occur is given by $T_c = (T_f + T_e + T_r)$ and the net change in yield is Y_c. Value of Y_c is determined by the new level of contamination.

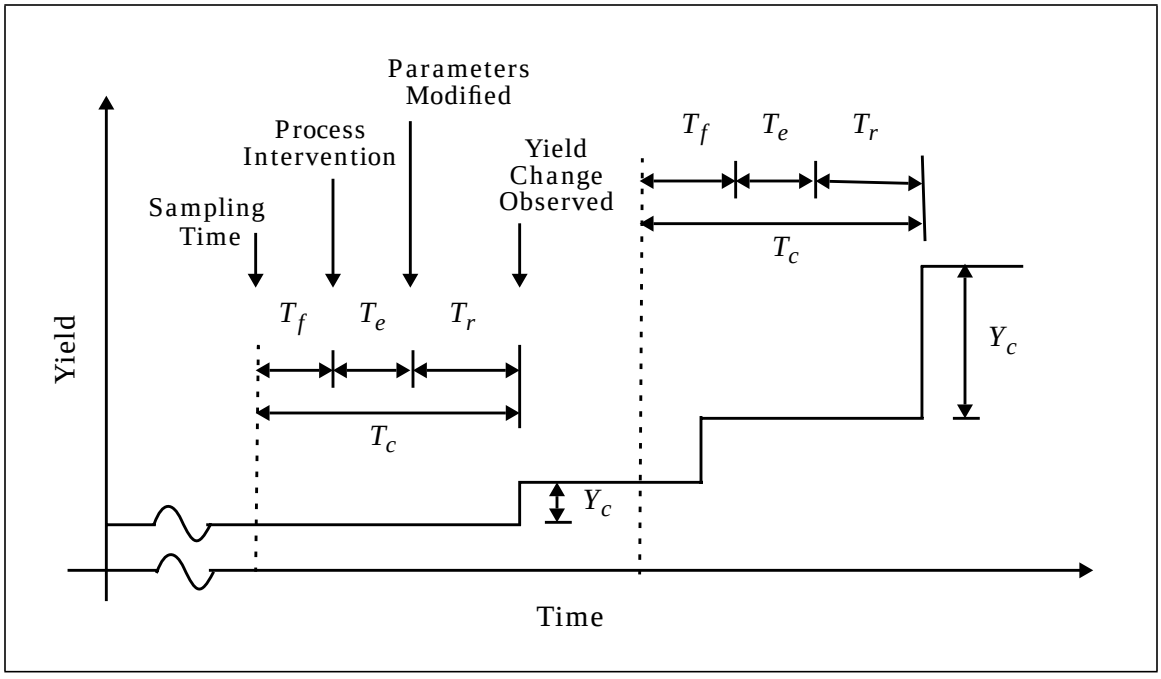

Figure 3.1 Key events in yield learning process.

Estimating T_r is equivalent to estimating the cycle time for a factory albeit partially starting from an intermediate process step until the last step. Thus, it is the sum of the raw processing time (RPT) and the queuing time for wafers waiting between process steps. One of the major contributors to the queuing time is the downtime of the equipment. Note that the factor T_e may contribute to the equipment downtime depending on the outcome of failure analysis. T_f, the time needed to detect and localize the defect depends on a number of factors as presented in the previous chapter. The change in yield, Y_c, on the other hand depends on the correctness of the diagnosis and the efficiency with which the contamination rate can be reduced as a result of the corrective actions.

Thus, even for a simple factory the inter-relationship between various attributes leading to yield improvement is quite complex. Moreover a realistic situation involves a multi-product facility with more than one source of contamination, many defect types, and several sampling and failure analysis strategies. In this case, the time to diagnose and correct different defect types will be interdependent because, for example, defects in lower layers will be "overlooked" in favor of defects in upper layers as described in chapter 2. The variability in the correctness of diagnosis will be affected since multiple sources of one type of contamination leads to ambiguity.

Note that Figure 3.1 depicts yield improvement cycles for only one type of defect originating from one source. In reality there will be a number of such cycles overlapping in time with each other. The yield learning curve for a product is, thus, a combination of all such individual overlapping learning curves.

3.3 Key Simulation Requirements

The simplified characterization of yield learning presented in the previous section can be easily modeled as the feedback cycle shown in Figure 3.2. The first step is to set the initial parameters of the particles for each source. Particles are introduced in a

wafer at the start of the cycle. After a certain delay (T_r) defects are identified and the wafer is sampled for further analyses. After another delay (T_f) the source is identified and corrective actions are initiated. Lastly after some delay (T_e) corrective actions become effective as a change in particle parameters.

Figure 3.2 represents the general structure of the algorithm for predicting yield learning curves. In order to correctly mimic a manufacturing line a number of such cycles, and their inter-dependencies have to be considered. Yield learning, as the cyclic sequence of steps described above is well suited to simulation. The production of

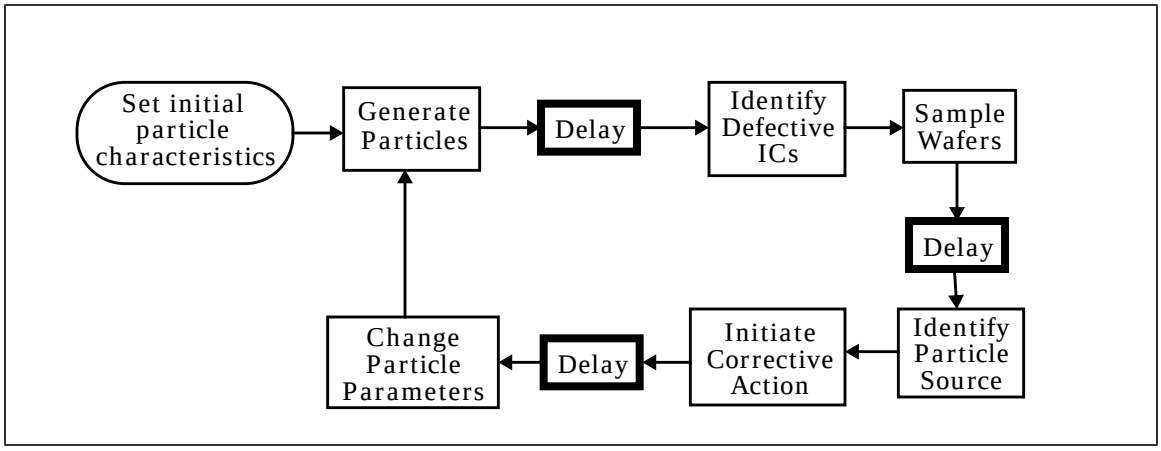

Figure 3.2 Model of yield learning.

wafers in a manufacturing line can in fact be thought of as a sequence of many such steps or events where each event affects some or all of factors T_e, T_f, T_r and Y_c. To accomplish this goal, an adaptation of a well researched discrete event model [1, 2, 3] has been used. In its simplest form an event is described by the following three characteristics:

1. *Time instance* at which the event is activated.
2. *Source* and *destination* of the event.
3. *Function* to be performed after an event is activated.

The time instance indicates when the function defined by the event is performed, or, in other, words when the state of the factory is changed. Sources and destinations are factory entities which are responsible for, and affected by, the event, respectively. After an event is activated two things can happen. First, some characteristics of the manufacturing line are altered e.g., wafers are loaded into a piece of equipment for processing. Second, a new set of events must be generated, in the above case the initiation of processing of wafers, after a known interval of time from the first event. The evolutionary nature of events for processing equipment in a manufacturing line is illustrated in Figure 3.3.

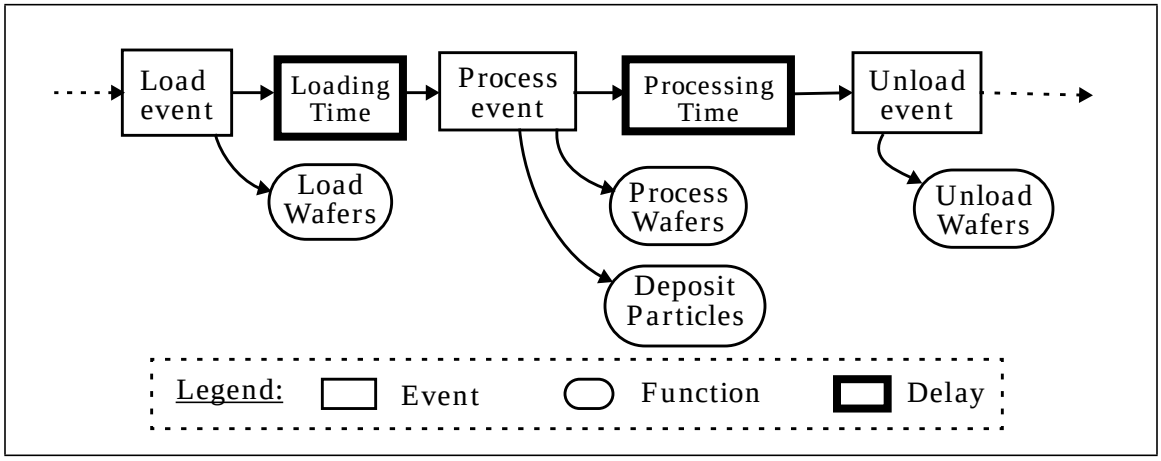

Figure 3.3 Event evolution.

One can say that there are two sets of models needed: one which models the timing of the factory, and another which alters other measurable characteristics of a factory besides time. This makes it easier to view the operational aspects of a factory independently of factors such as yield related characteristics. In some cases, however, these two may be related. For example, timing of failure analysis must be dependent on the simulated defect characteristics and the probability of their occurrence. Similarly, the number of particles deposited on a wafer may also depend on the amount of time a wafer spends in processing equipment or even waiting between steps.

Note that the time interval between two related events can be zero or indefinite. Thus, two special types of events: waiting-events and zero-delay events, are also required to facilitate simulation. Zero delay events are useful for the collection of statistics or to perform other calculations. Figure 3.4 shows a chain of event evolution to illustrate this concept. Here there are two types of events: primary and secondary events. Primary events describe the operation of the manufacturing process and secondary events are used to extract relevant data before and after an operation. Waiting events, as the name implies, are activated only by the occurrence of another event. For example, when cleaning is necessary the equipment must wait until it finishes processing the loaded wafers.

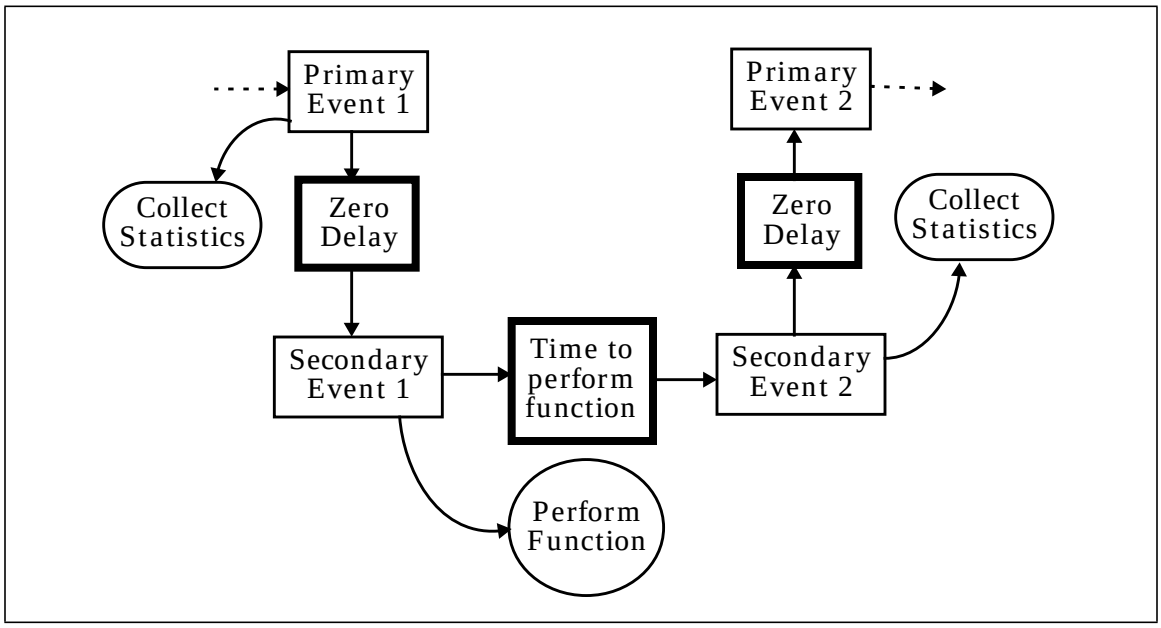

Figure 3.4 Application of Zero delay events.

Based on the nature of change in state of the manufacturing line, the primary events can be classified as being related to:

1. Movement of wafers through fabrication steps;
2. Introduction of particles and formation of defects;

3. Detection of defective ICs (testing);
4. Failure analysis activities as a result of:
   a. Particle monitoring and,
   b. defect diagnosis of fabricated wafers.
5. Corrective actions.

The primary requirement of the simulator is the ability to simulate wafer movement in a fabrication line taking into consideration such aspects as the product, process recipes, equipment, personnel and operating rules. To model yield loss due to particles, the simulator must be able to introduce particles, and transform particles to defects and ultimately to faults for each wafer. The simulation must be able to take into account the random nature of the particle introduction and the transformation processes. Hence, some variation of Monte Carlo [4, 5] simulation capability must be achieved to estimate manufacturing yield loss. Wafer testing simulation must factor in both the time required for test execution and the effect of imperfect testing on the observed yield. Testing time should be made dependent on the rate of defective ICs.

Particle scanners should be simulated taking into consideration the sampling rules, variability in the accuracy of detection and the efficiency of equipment. Fabrication line control policies such as wafer rejection and initiating corrective actions must be taken into account in modeling particle scanners. Defect diagnosis of fabricated wafers must be simulated as a sequence of four steps: sampling of wafers, defect localization, particle identification and identification of source. Wafer sampling should be guided by certain rules which depends upon the outcome of testing. Efficiency of defect localization should be a function of defect characteristics, product design and initial uncertainty due to lack of information from testing activity. Uncertainty in particle and source identification should be modeled as well.

Next, a model to simulate the effect of corrective actions is required which is consistent with the model for yield simulation. This model must take into account continu-

ous changes in size and density of particles as well as an abrupt removal of the particle source. Ideally, corrective actions should also take into account changes in particle to defect and defect to defect transformation processes.

To achieve the capability of performing cost revenue trade-off studies, the simulator must also be able to take into account the capital and operating cost of fabrication, testing and failure analysis equipment. One should be able to estimate the cost of wafers, cost of good die and the number of good die fabricated (or productivity). The cost calculations must be dependent on the usage of various manufacturing resources (e.g, equipment, materials, etc.) in such a way that one can perform "what-if" analysis. In the next chapter, detailed formulation of specific modeling requirements for each of the required simulation capabilities is presented.

References

[1] J. Banks and J. S. Carson II, *Discrete Event System Simulation*, Prentice-Hall, Engelwood Cliffs, New Jersey, 1984.

[2] D. J. Miller, "Simulation of a Semiconductor Manufacturing Line", *Communications of the ACM*, vol. 33, no. 10, pp. 99-108, Oct. 1990.

[3] C. D.Pegden, R. P. Sadowski, and R. E. Shannon, *Introduction to Simulation Using SIMAN*, 2nd ed, McGraw Hill, 1995.

[4] R. Y. Rubinstein, *Simulation and The Monte Carlo Method*, John Wiley and Sons, 1981.

[5] H. Walker and S. W. Director, "VLASIC: A Catastrophic Fault Yield Simulator for Integrated Circuits", *Trans. on Computer-Aided Design*, vol. 5, no. 4, pp. 541-556, 1986.

Chapter 4

Simulation Models

In this chapter, the models needed for yield learning simulation are presented in detail. The models are of seven distinctively different classes:

1. Wafer movement simulation models;
2. Yield simulation models;
3. Models for simulating the testing process;
4. Particle monitoring simulation models;
5. Models for the defect diagnosis process;
6. Models for simulating the effect of corrective actions;
7. Cost models.

In this chapter, modeling assumptions and representative equational forms are presented.

4.1 Wafer Movement Simulation

The wafer fabrication phase is modeled as being comprised of five entities: equipment, products, process recipes, operators and factory rules. The models of these entities are as follows.

Manufacturing equipment is organized as described in Chapter 2. The factory is divided into work areas consisting of one or more workstations which in turn are made up of one or more pieces of equipment generally capable of performing the same processing steps. Each workstation is associated with a queue (storage area) where incoming wafers are temporarily stored. Each piece of equipment is characterized by

its *capacity* expressed as the number of wafers that can be processed in a *single run*. The minimum allowable number of wafers in a *single load* is a lot. Batch equipment can have a capacity greater than the lot size. The piece of equipment is also associated with a set of operating rules defined later. Setup, load and unload times are defined for a piece of equipment. Equipment processing time is modeled as a sum of required process time (defined by recipe step) and equipment timing error defined as a distribution function with a given mean and variance.

Each product is identified by a name and a unique design associated with it. Depending on a manufacturing line operating policies, sometimes many different products can be the same design. The main parameters of a product are its lot size, wafer size, size of die and the number of dies per wafer. Each product is also associated with a process recipe to be used in its manufacturing. Wafers are released into the factory according to the *wafer generation mechanism*. Wafer generation is conceptually performed by a virtual piece of equipment. The wafer release rate can be defined in two ways, either by a distribution function with a given mean and variance of wafer starts per week (WSPW), or by a time dependent function of WSPW.

Process recipes consist of a number of steps to be performed in a particular sequence. Each step of the recipe is identified by a name, the workstation to be used and processing time. In some cases, like a metrology step, processing time can vary and in this case the process step is associated with a *process time estimation model*. The real processing time for a single run is estimated by adding equipment timing error.

Operators in the line are assigned to each workarea and it is assumed that any assigned operator can operate any piece of equipment in the workarea. The function of the operator is to set up the equipment, load and unload wafers and move the wafers as and when required depending on their availability.

Factory rules include lot release and dispatch rules, and equipment load and setup rules. Lot release rules defines the manner in which lots are to be introduced into the factory for each product which could be just one lot at a time or several lots in a group. The lots are placed in the queue of the first workstation defined by the process recipe for that product. Lot dispatch rules decides the order of lots in the queue; it can be based on the arrival time (FIFO or first-in-first-out rule) of the lots, their predefined priority (hot lots), or by bottleneck indicators. A bottleneck indicator could be that the waiting time of lots in a particular queue has exceeded a predefined threshold value. Equipment related operating rules are the setup and load rules discussed in Chapter 2 and considered to be input parameters. The general algorithm for moving the wafers at an intermediate step, beginning with the arrival of a new lot, is shown in Figure 4.1.

Operating rules are defined in a hierarchical fashion starting from factory wide rules, down to workareas, workstations and finally to the equipment level. In the absence of lower level rules, the immediate higher level rules take precedence and are applied if appropriate. The wafer movement simulation models are adaptations of more extensive and complex models described in [1].

4.2 Yield Simulation

The primary aim in yield modeling is to classify each die on a wafer as fault-free or faulty so that yield can easily be estimated by evaluating the ratio of good die to the total number of die on a wafer. One can further associate with each die a list of faults, the corresponding defects, the steps at which the defects were formed, the particles causing these defects and the equipment responsible for the particles. In this way a complete trace of the fault to the source can be ascertained. As will become apparent later, such information can be used to model the efficiency of defect diagnosis.

Ideally, one can perform yield simulation using established Monte Carlo techniques to mimic the introduction of particles on to the wafer and then simulating the interac-

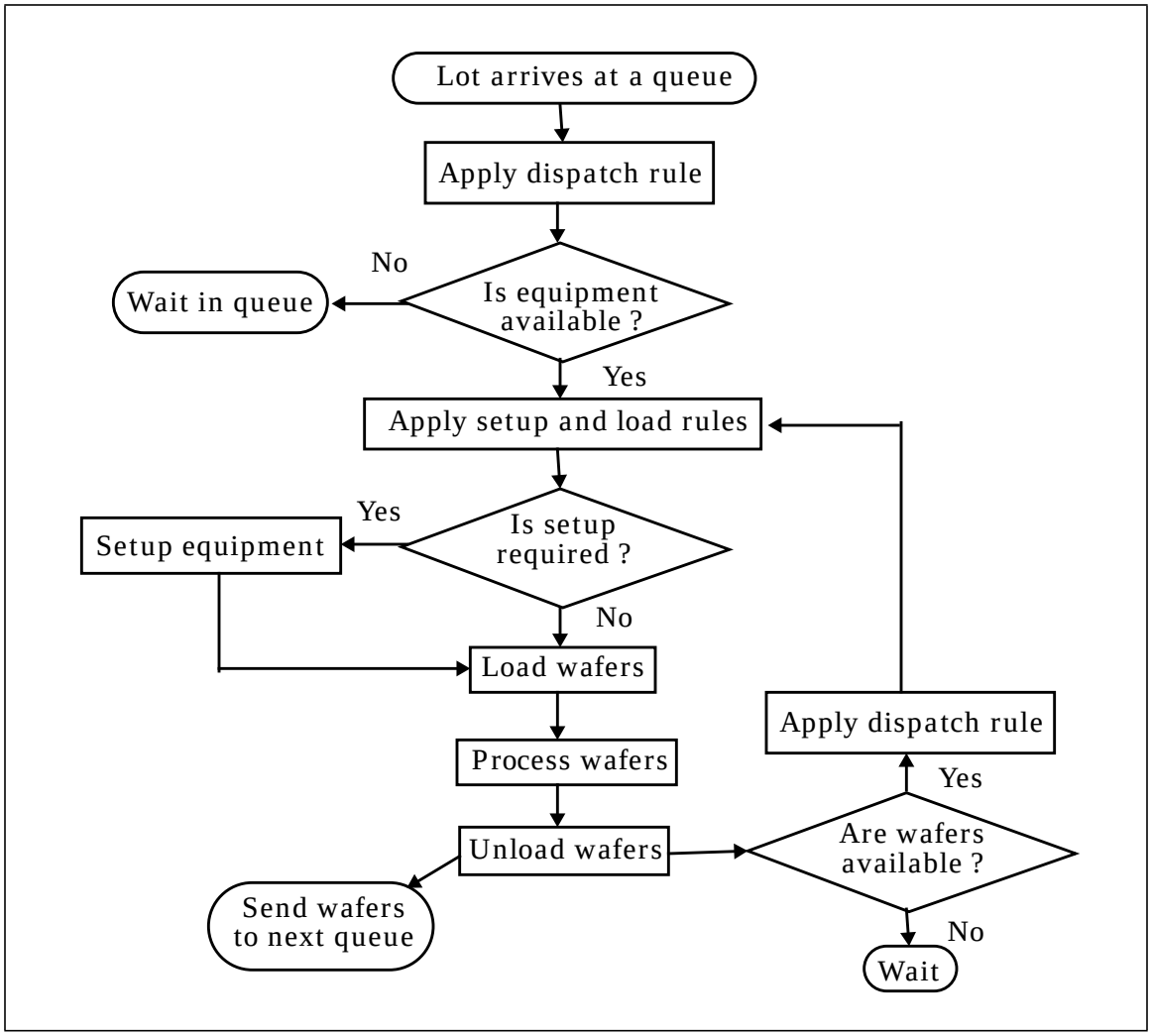

Figure 4.1 Algorithm for sequencing wafers at a single step.

tion of the particle with processing steps as in CODEF [2]. The defective circuit should then be simulated or compared with circuits stored in a database to determine the presence of a fault. Such a simulation would be, however, excessively time consuming since millions of particles would need to be simulated. Hence some simplifications, with perhaps some reduction in accuracy and limitations, are needed to simulate yield loss efficiently.

4.2.1 Simplified Method of Yield Estimation

First, let us assume that particle types are uniquely identified by, for example their chemical and physical characteristics, and associated with their source i.e. the generating equipment. However, each source may generate more than one type of particle and several sources can generate the same type of particle. A source and particle type pair will be referred to as a disturbance type.

Each disturbance type is assumed to generate two or three dimensional particles of a certain size, R_c. If the number of particles on a wafer is given by N_c, then the occurrence of particles on a wafer can be described by the joint probability distribution $f(N_c,R_c)$. Figure 4.2 illustrates the hierarchy of relationships between disturbance types and particle characteristics. Generally, the distributions of N_c and R_c are

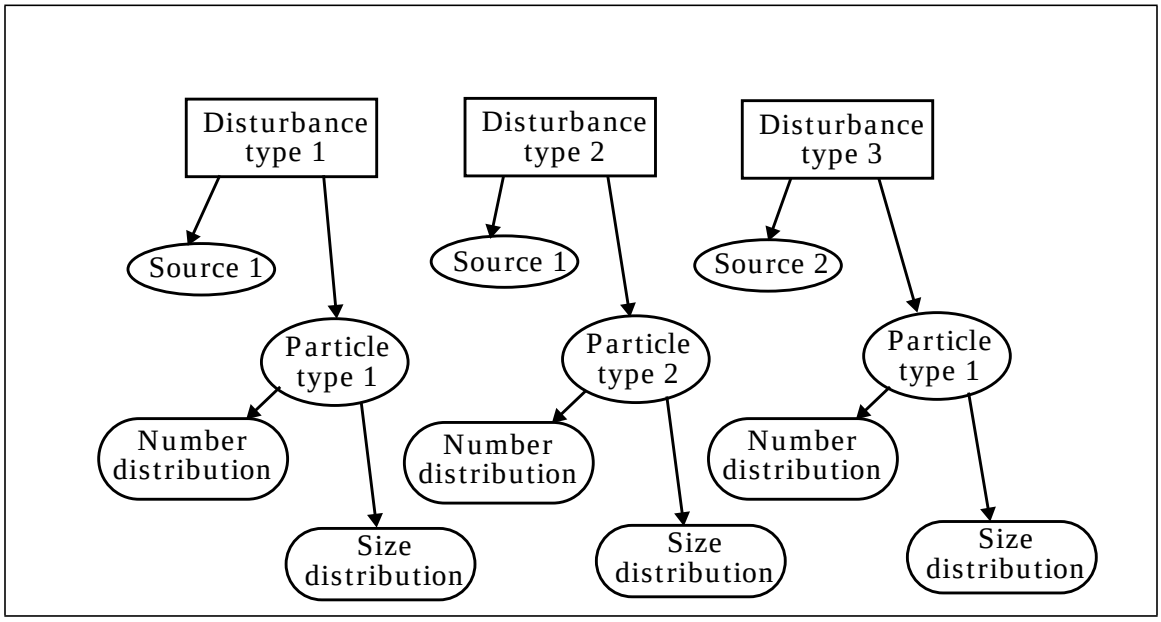

Figure 4.2 Disturbance type characteristics.

assumed to be independent of each other and modeled as gaussian and polynomial dis-

tributions respectively [3, 4, 5, 6]. Therefore, the particle distribution can be described by:

$$f(N_c, R_c) = f_N(N_c) f_R(R_c) \tag{4.1}$$

where, $f_N(N_c)$ is the distribution function of the number of particles per wafer and $f_R(R_c)$ is the particle size distribution function. $f_N(N_c)$ is given by the gaussian distribution:

$$f_N(N_c) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(N_c - m)^2}{2\sigma^2}} \tag{4.2}$$

where, m and σ are the mean and standard deviation and are assumed to be known. $f_R(R_c)$ is given by:

$$f_R(R_c) = \frac{K}{R_c^p} \qquad R_c \geq R_{min} \tag{4.3}$$

where, K is a constant and p is a known exponent extracted experimentally [6]. R_{min} is the minimum size of particles, greater than zero, that can cause a defect in the IC. The parameters of the above equations may vary with time making the occurrence of particles on a wafer a non-stationary stochastic process. One should also assume a spatial variation in the distribution of particles across a wafer surface, $g(x_c,y_c)$, where (x_c,y_c) is a point on the surface [7]. But such an assumption makes yield modelling much more complex and, instead, particles are assumed to be distributed uniformly.

The next step in yield modeling is to characterize the occurrence of a fault on a die as a result of particles introduced on the wafer surface of certain sizes. This can be achieved in two steps where the transformation of contamination to defect is modeled first and then the defect to fault transformation. These modeling steps are presented in the next two sections.

4.2.2 Mapping Contamination to Defect

There are three types of transformations of contaminants to defects. These are: formation of defects from contamination at a certain process step, removal of contamination at a cleaning step and formation of defects at a process step from defects formed in earlier steps. Note that the key to modeling these components is identifying the process step and the layer in which the transformation occurs. The three possible scenarios of the fate of a particle are depicted in Figure 4.3.

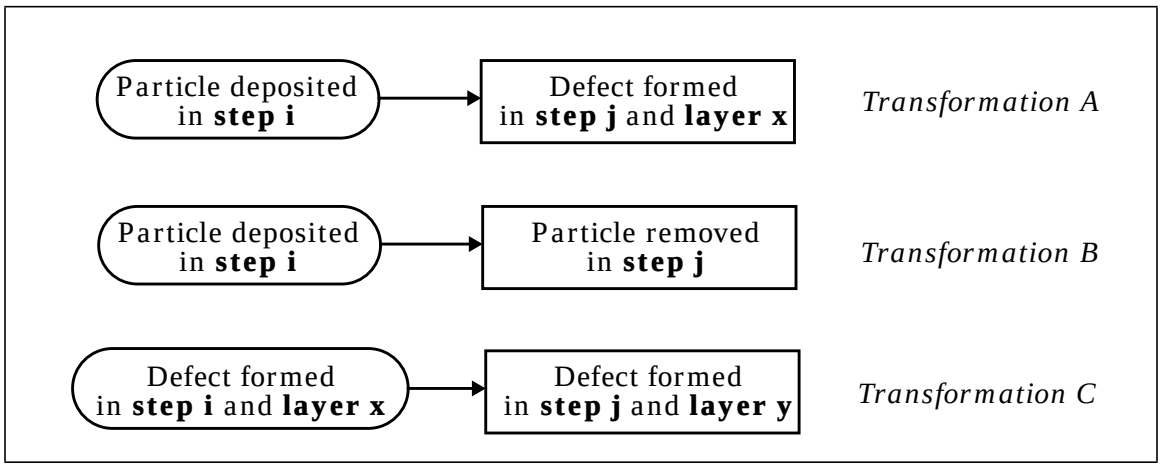

Figure 4.3 Possible transformations of particle.

In the model for transformation of particles to defects one has to consider change in defect size also. Such a transformation can be formulated as:

$$R_d = h_c(R_c, x_c, y_c) \tag{4.4}$$

where, R_d, x_d and y_d are the size and location of defects, R_c, x_c and y_c are those of the particles, and $h_c()$ is the transforming function. In its simplest form, $h_c()$ can be assumed to depend only on the particle size and thus, R_d can be given by the function $C_c R_c$ where, C_c is a given constant.

During cleaning steps, surface particles are removed and this can easily be modeled as the probability of particle removal, $P_c(R_c)$. The probability function can be made dependent on the particle size since larger particles are observed to be easier to

remove than smaller ones. This model is the more general form of the experimental determination of wet cleaning efficiency as presented by [8, 9] and it is further assumed that the same model holds for other methods of cleaning as well.

Defect to defect transformation (three dimensional defect propagation) can be modeled in a manner similar to particle to defect transformation. Thus the formula can be written as:

$$R_{dd} = h_d(R_d, x_d, y_d) \tag{4.5}$$

where, R_{dd}, x_{dd} and y_{dd} are the new size and location, respectively, and $h_d()$ is the transforming function. Again, this transforming function is assumed to be of the form $C_d R_d$ where, C_d is another constant. Note that defects may also get removed by steps such as layer polishing. Such an effect can be modeled similarly to the ones proposed above.

4.2.3 Mapping Defect to Fault

Defect to fault mapping has been extensively studied in the past as noted in Chapter 2. There are two distinct methods, one uses Monte Carlo techniques [2, 10, 11] and the other uses models based upon the critical area concept [12, 13]. Monte Carlo techniques are excessively time consuming, whereas models based on critical area estimate only the average yield. Neither method is directly suitable to answer the question: given the size, location and layer of a defect on a wafer, what is the resulting fault, if any?

This question can easily be addressed by making certain modifications to the critical area based yield models. It is first assumed that faults are defined as shorts or opens in a particular layer, and that their corresponding defects are extra and missing material in that layer. Let us now assume that a defect of certain type and size, R, occurring

in a particular die causes a fault of type *i*. Then the probability, p^i_f, of this fault occurring given that a defect of size R has occurred is given by:

$$p_f^i = \frac{A_c^i}{A_{chip}} \tag{4.6}$$

where, A^i_c is the critical area for fault *i*, and the defect type and size (*R*) under investigation (see e.g., Figure 2.15), and A_{chip} is the total area of the die. Then if *n* types of faults can be caused by this defect, the total probability that a fault has occurred, P_f, and that no fault has occurred, Q_f, are given by:

$$P_f = \frac{\sum_{i=1}^{n} A_c^i}{A_{chip}}$$
$$Q_f = 1 - P_f = p_f^0 \tag{4.7}$$

where, p^0_f is a convenient representation for the probability that no fault occurs. Hence, to simulate the occurrence of a fault (for a given defect size, *R*) one has to select a number from 0 to *n* randomly using the values p^i_f.

Next, one has to consider the probability of occurrence of a defect of certain type and size which is the same as that of a particle, p_c (since it is assumed that location of particles and of corresponding defects are the same). Further, since the particles are assumed to be distributed uniformly one can express p_c as:

$$p_c = \frac{1}{N_{chip}} \tag{4.8}$$

where, N_{chip} is the number of dies on a single wafer. (Note that Equations 4.6 and 4.8 are valid only if the particles are uniformly distributed over the wafer surface.)

Critical area is assumed to be known before simulation. In reality it can be extracted using a number of available methods which analyze design layouts for this purpose. Available methods for such computations have been observed to be prohibitively

expensive for a large layout containing several million transistors. To improve computational efficiency, a new method was developed to extract critical area using layout hierarchy. A complete description of the method appears in [14].

4.2.4 Estimating Yield

Based on the models presented above, the method to estimate yield can now be formulated as depicted in Figure 4.4. The first step is applied to each wafer being processed in a piece of equipment at a particular step. The equations presented earlier are applied in sequence as shown in the figure. At the end, each defective die is associated with a list of faults and their corresponding defect and particle characteristics. Finally, when the wafer is completely fabricated yield can easily be estimated by counting the number of defective die on each wafer. The mean of yield values obtained from all wafers gives the average yield in the case where particles do not change with time (static case).

4.3 Test Simulation

The primary objectives in developing a model for the testing process are: estimating sort yield, estimating time requirement, and characterizing the nature of defective die used for failure analysis. In this section, only the first two factors will be dealt with. Characterization of defective die is more relevant to the development of the failure analysis models.

4.3.1 Sort Yield

Since only contamination related yield loss is considered, the sort yield appears to be higher than actual yield due to two factors: primarily due to less than 100% fault coverage, and to a lesser extent due to some of the low yielding wafers being sampled for failure analysis. Fault coverage is normally defined at the top level for all faults taken together. At the level of a fault on a die a *detection parameter* which takes a

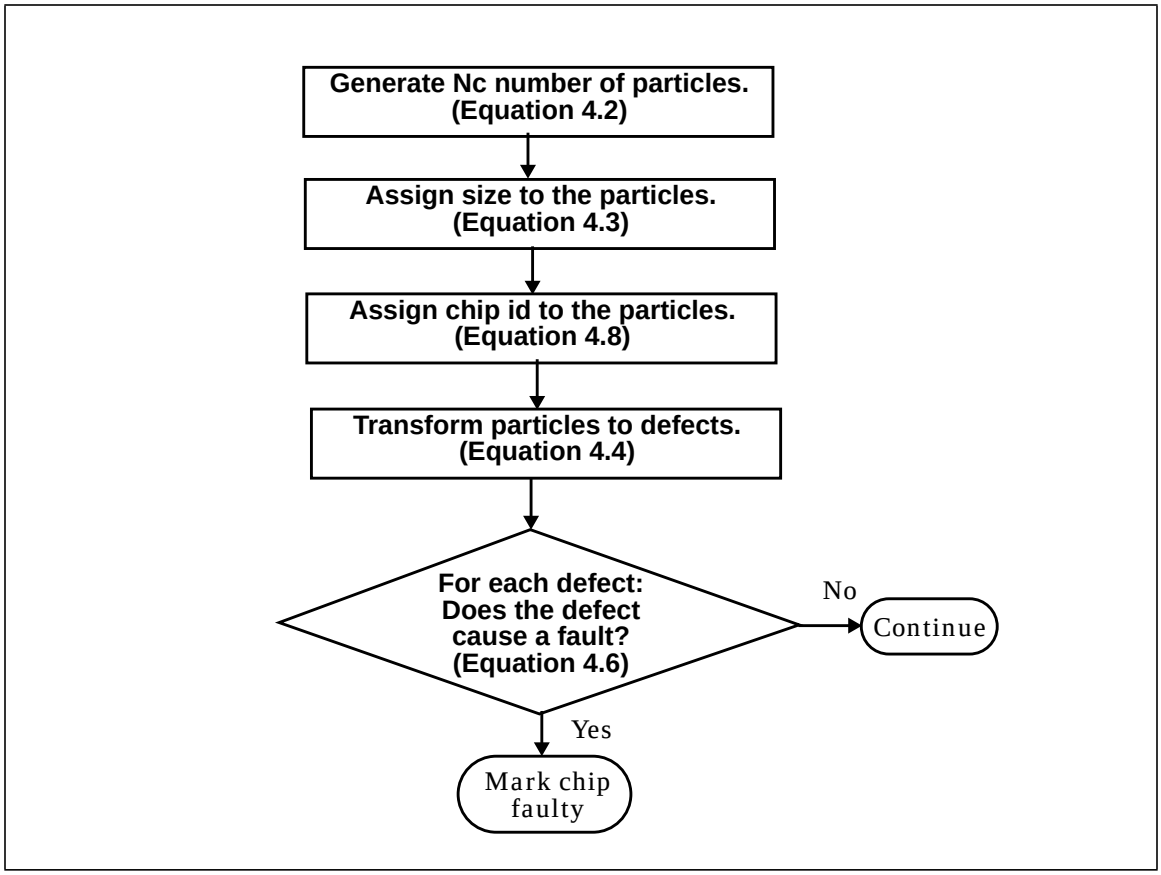

Figure 4.4 Method to estimate yield.

value of 1 or 0 depending on whether or not a fault is detected by testing process. This can achieved by treating fault coverage as a probability value and assigning detection parameter for each fault a value of 1 with this probability. One could have also defined fault coverage values for each subset of faults or at the level of individual faults modeled for testing purposes. The same method can be applied in all such cases.

To estimate the sort yield, one has to now determine whether at least one fault for each defective die is detectable. In this case, a die can be assigned a *tested faulty* value of either 1 or 0. This model of imperfect test quality does not take into account good dies being tested faulty or false reject outcomes. For this model, the sort yield will always be higher than or equal to the manufacturing yield.

4.3.2 Time Required to Test

Operationally, testers are equivalent to normal processing equipment. Their operating rules are also similar and thus scheduling of wafers through testers can be simulated in the same fashion as any other piece of equipment in the fabrication phase. Rules for lot dispatch, equipment setup and loading are assumed to be known input parameters [15]. The capacity of each tester is one wafer in a single run and the time required to completely test a lot, T_{lot}, is given by:

$$T_{lot} = N_{lot} T_{wafer} \tag{4.9}$$

where, N_{lot} is the number of wafers in a lot, and T_{wafer} is the time taken to test a wafer and is given by:

$$T_{wafer} = T^{l}_{w} + T^{u}_{w} + N_{chip} T_{chip} \tag{4.10}$$

where, T^{l}_{w} and T^{u}_{w} are the load and unload times of a wafer, respectively, and T_{chip} is the time to test a chip. T_{chip} is given by:

$$T_{chip} = T^{l}_{c} + T^{u}_{c} + T_{exec} \tag{4.11}$$

where, T^{l}_{c} and T^{u}_{c} are the load and unload times of a chip, respectively, and T_{exec} is the time to execute the tests. T_{exec} depends on whether a tested die contains a detectable fault as less time is required when a fault is detected early in the test sequence. T_{exec} is assumed to be either a constant parameter, a single distribution function for all types of faults, or separate distributions for different fault types.

4.4 Particle Monitoring Simulation

Particle monitors have two primary purposes in a manufacturing line: Collecting data on the frequency and size of particles and controlling the manufacturing line using such data. Efficiency and accuracy of data collection depends on the sampling

methods and particle detectability. This section presents the requirements for a model of a particle monitor to be able to capture yield-related characteristics.

4.4.1 Sampling rules

Sampling rule for a particle monitor can be defined by three factors for each product being monitored: lot sampling rate, wafer sampling rate and the area of the wafer to be scanned. Lot sampling rate is defined as the number of lots to be skipped before sampling one for analysis. Wafer sampling rate defines the number of wafers out of a lot to be chosen for analysis. It is assumed that any wafer can be selected with equal probability. Area of the wafer to be sampled is given as the number of adjacent dies to be scanned for each wafer.

Sampling rules should be such that the capacity of the equipment is not exceeded. Thus, the number of equipment required to perform particle scanning has to be pre-computed by looking at the wafer start rates for each product and the average time required to scan each die. Note that the time required to scan a die is also a function of the accuracy desired. Hence, prior to any simulation both accuracy and capacity requirements must be known.

4.4.2 Accuracy of Monitoring

Accuracy of a particle monitor is characterized as the probability that a particle of a given size is detected on the IC surface. Small particles are generally hard to detect than large ones. Probability of detection or *particle detectability,* $p_d(R)$ is assumed to be given and examples of such a function is shown Figure 4.5. It is tacitly assumed that detection efficiency does not depend on any other factor such as orientation, surface properties, etc. The particular characterization chosen for $p_d(R)$ is:

$$p_d(R) = K_d\left(1 - e^{-\alpha R}\right) \tag{4.12}$$

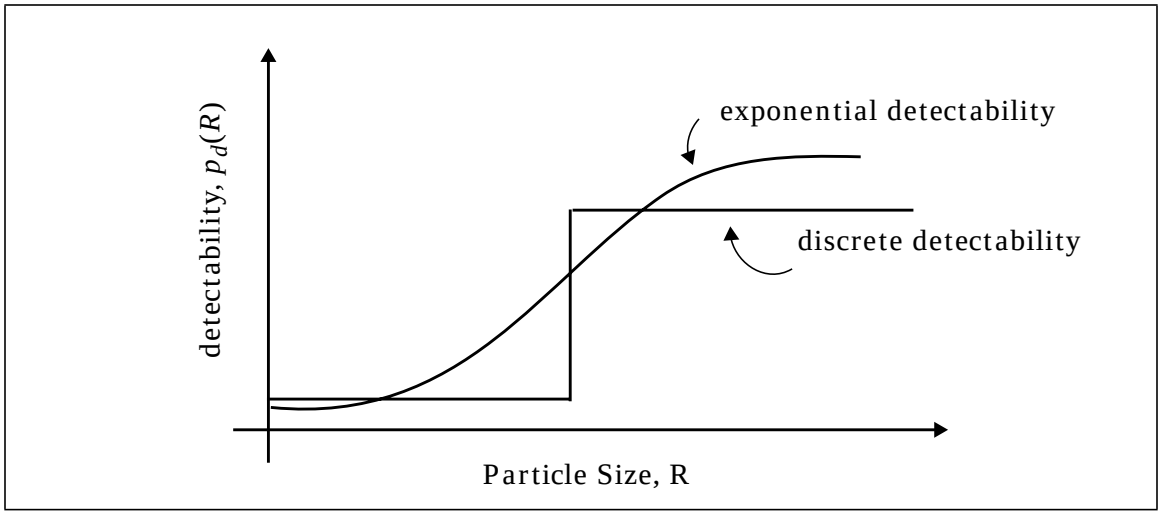

Figure 4.5 Particle detectability.

where, K_d is a constant and α is a parameter determining accuracy of the equipment. Higher the value of α, the more likely it is to detect a smaller particle. Also note that T_{chip} must be consistent with the value of α, since a larger α implies a larger T_{chip} value.

Each particle on a sampled wafer is associated with a detection value of 0 or 1 depending on whether it is detected by particle monitoring activity or not. This information can be used later to model manufacturing line control and can possibly be used to model fault to defect correlation in observed test or defect diagnosis results.

4.4.3 Controlling Manufacturing Line

The two aspects of manufacturing line control using particle monitor data considered are rejection of wafers and initiation of corrective actions. Reworking of wafers as a result of particle monitoring activity is not considered. Wafer rejection can be simulated by defining a threshold function on the number of particles detected. It is assumed that the threshold value of number of particles is a given parameter and whenever the observed number of particles exceeds this value the sampled wafer is rejected. Note that rejecting wafers has an adverse effect on loading of the line. This

can be easily corrected by using a feedback control to replace lost wafers by introducing an equal number of bare wafers. This is referred to as WIP control.

Initiating corrective actions can be based on another threshold function defined for the source of particles. To simplify modeling, it is assumed that the particle source is always diagnosed correctly. A *particle number counter* is associated with the source to indicate whether the source has been held responsible for excessive particle introduction. Application of corrective actions and the corresponding models will be presented later.

4.4.4 Discussion of Particle Monitor Modeling

A simple model was presented in this section to model particle monitoring activity which in reality is much more complex [16, 17]. Available data on the functioning of a particle monitor is limited [18] and the aspect of accuracy and efficiency of particle monitors should be investigated further. Specifically, the threshold functions defined above are over-simplified; one must also take into account the size and locations of the particles observed. The data collected by particle monitors might not be used immediately to control the manufacturing line. In fact, the data could be used to correlate the failed die observed on a wafer to its corresponding particle monitor data. Appropriate models are required for this which have not been dealt with here.

4.5 Defect Diagnosis Simulation

The primary objective in modeling the defect diagnosis process is to estimate the time required to identify a subset of equipment responsible for defective die from the point it is sampled. The analysis process itself is composed of a predefined sequence of steps, a specific type of equipment being used at each step. The time required at each step is determined using a diagnosability model, presented later, which takes into account the attributes of the product, the defect and the failure analysis equipment.

The analysis identifies a subset of fabrication equipment and estimates for each piece of equipment a measure of correctness of diagnosis, using an assignment rule. This functional model of the failure analysis process is depicted in Figure 4.6. Such a representation is well suited to discrete event simulation. A set of rules and models which are consistent with this functional representation follows.

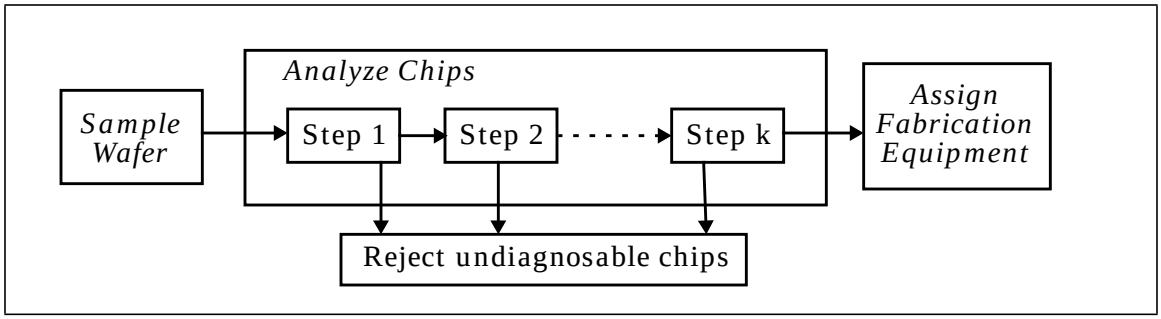

Figure 4.6 Functional representation of the defect diagnosis process.

4.5.1 Sampling Strategy

Sampling strategy of wafers is an input to the failure analysis process. A possible sampling strategy is to select those wafers that have at least a minimum number of defective die, D_{min} or to select wafers from special bins predefined by the testing process. This last rule has not been included in the test process simulation model but this should not impose any limitations on the applicability of the failure analysis model. Whatever the chosen rule, it can lead to overloading of the equipment available for failure analysis. To avoid this, a sampling rule should be combined with the requirement that wafers can be sampled only when the number of wafers in the input queue of failure analysis is less than an allowed maximum, Q_{limit}. Figure 4.7 shows the sequence of operation for successfully sampling wafers for failure analysis.

4.5.2 Timing of Analysis

A model to estimate the time required at each step of the failure analysis process is presented in this section. First, consider a diagnosability measure, m, with a value

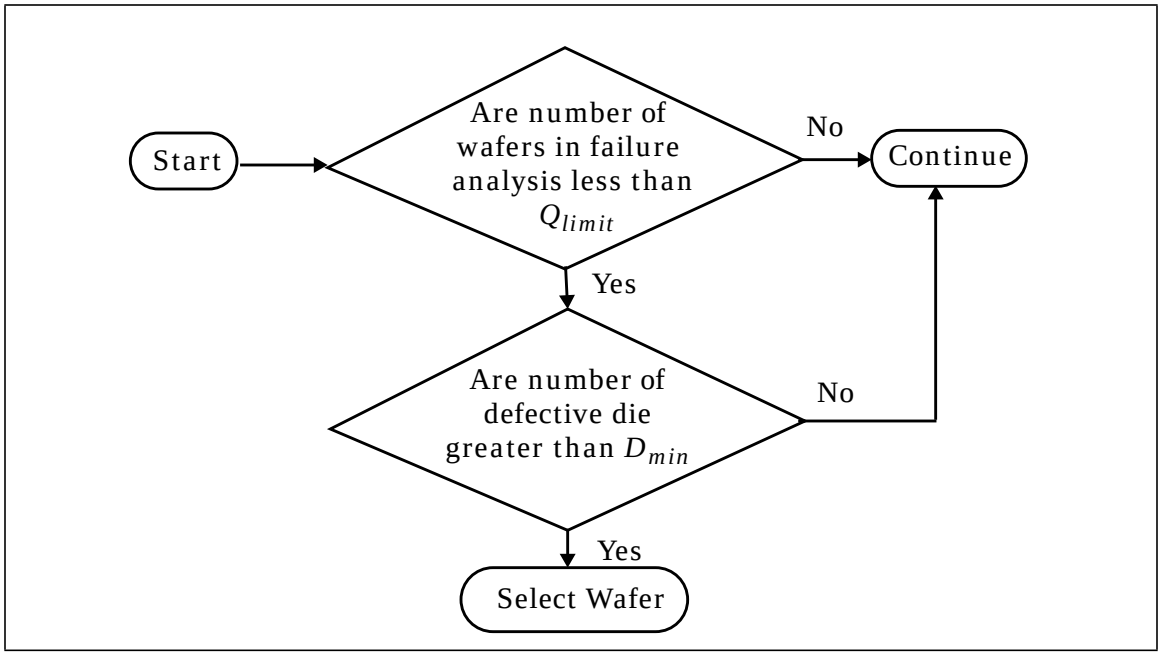

Figure 4.7 Sampling strategy for defect diagnosis.

between 0.0 and 1.0 for each fault defined for a product. A value close to 1.0 indicates that the fault is easily diagnosable and a value of 0.0 indicates the fault to be undiagnosable. Such a measure should be dependent on the type of defect causing the failure, the layer in which the defect occurs, and the size of the defect. Suppose that at each step, starting with an initial value of m_i, a final value of m_f is achieved in time t_f. One possible form of the function is given by:

$$m_f = 1.0 - (1.0 - m_i)\, e^{-e_d t_f} \tag{4.13}$$

where, e_d represents the efficiency of the diagnosis process and is a parameter which depends on the analysis equipment. Higher the value of e_d, the more efficient is the diagnostic process. The above function is graphically shown in Figure 4.8. Note that the parameter e_d can itself be a function of the type, layer, and size of the defect in addition to being dependent on the efficiency of the equipment used. The above model of the diagnostic process implies that more the time spent on analysis, the higher are the chances of detecting the cause of the fault.

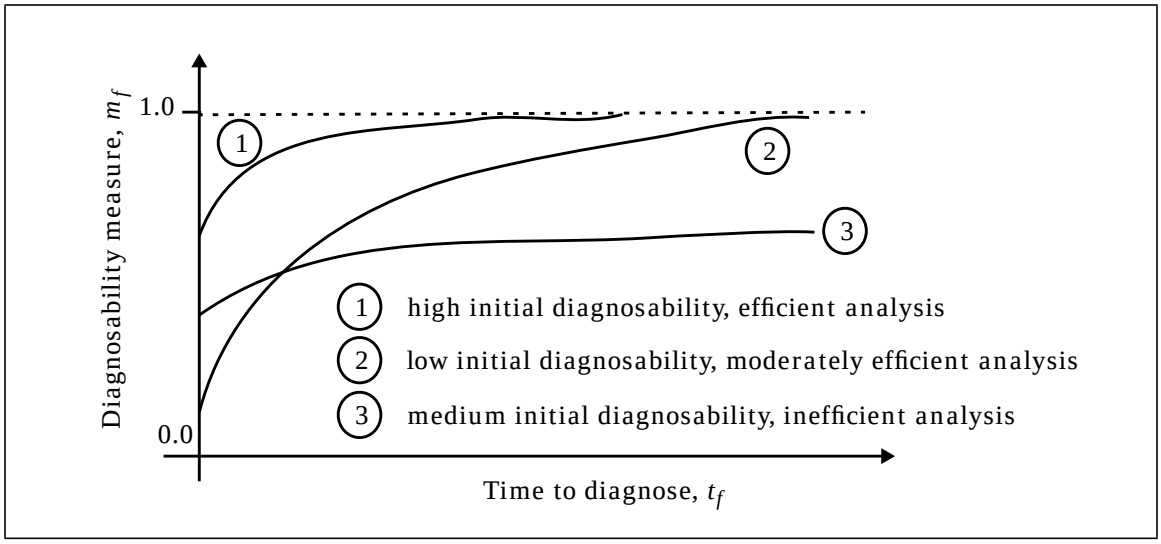

Figure 4.8 Diagnostic efficiency of failure analysis.

Since, the amount of time spent on analyzing a single die in a piece of equipment is finite, a limit T_{cmax} should be assumed for each step. Further, a minimum allowable value m_{thresh} must be defined for the initial diagnosability measure, m_i, for each step. This parameter is introduced to reflect the fact that the analysis on a chip may be discontinued after a certain step because at that point the cause of the fault is deemed to be "undiagnosable".

It remains now to define a model for estimating the initial diagnosability for the first step of the analysis. It is assumed that each fault for a product is characterized by:

1. An estimate for the area on the chip where the defect may be present, A_s. The maximum value for A_s is A_{chip} or the total area of the chip.
2. The size of the defect, R.
3. The layer n, in which the defect is manifested, $n = 0$ for the top layer.

Using these parameters one can estimate the initial diagnosability using the following equation:

$$m_i = (1 - a \cdot n)\,(1 - b \cdot A_s + c \cdot A_s \cdot R) \tag{4.14}$$

where, a, b, and c, are positive constants which captures the relative importance of each of the three attributes defined above. A high value of a indicates that defects in lower, unexposed, layers are more difficult to observe. A high value of b limits the search area for the cause of the fault. A high value of c indicates that larger defects are increasingly easier to detect in spite of large area of search for the cause of the fault. This model provides the ability to capture differences in products which are affected by the same kind of defects.

To illustrate the nature of the Equations 4.13 and 4.14 let us first assume that n = 0 which means only defects in the top layer are of interest. Figure 4.9 shows the nature of Equation 4.14 as a function of A_s and R for certain values of b and c. One can also

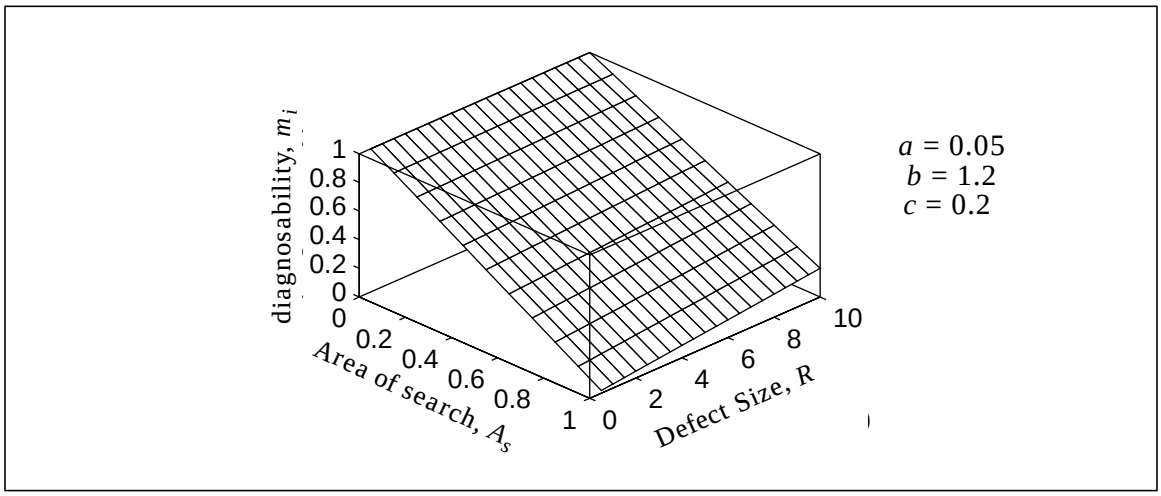

Figure 4.9 Initial diagnosability measure as a function of A_s and R.

calculate the limiting values of t_f needed to reach a value of m_f = 0.99; this is shown in Figure 4.10 as a function of A_s and R for the case shown in Figure 4.9. The time is in log scale since its range is large due to sharp changes near the extreme points.

4.5.3 Sequencing of Wafers

Sequencing wafers for defect diagnosis requires that a recipe be defined as a sequence of steps much like a process recipe in the fabrication phase. However, this

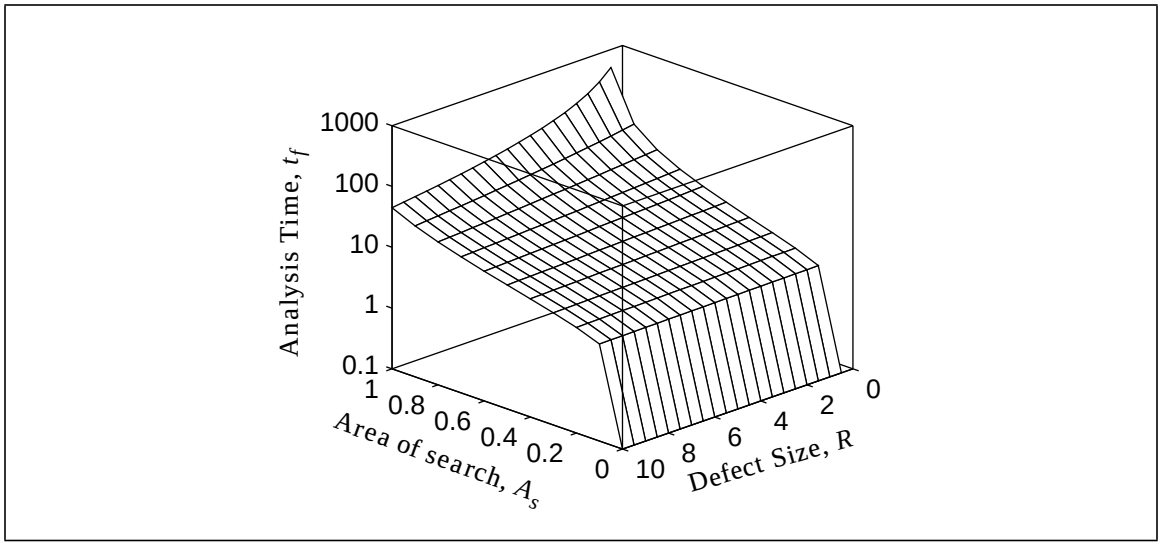

Figure 4.10 Analysis time t_f as a function of A_S and R.

recipe cannot be a simple linear arrangement of steps since, in general, the choice of the next step depends upon the outcome of the current step. As a specific example, assume that most of the defects occur in the polysilicon layer in a single metal process. There can be two possible outcomes after initial analysis:

1. A defect in polysilicon is observed through the top layer or;
2. Most of the defects are missed.

In such a case, layer stripping will become necessary and the first few steps of simple optical microscopy may need to be performed again. Hence, one may need to model the fact that at an intermediate step during defect localization any one of three things are possible:

1. defects are localized and analysis continues to particle identification,
2. strip layer and continue analysis or,
3. abort analysis altogether.

In this model the actual time spent by a wafer at a single step is a combination of a number of interacting phenomena. The equipment used for defect diagnosis and more

importantly the engineers are considered to be limited resources just as in the wafer fabrication phase. This way the queuing time can also be factored into the model.

4.5.4 Assignment Rules

In order to formulate the assignment rule, the source (equipment) of the particle has to be associated with each fault in a chip (one of the assumptions of the presented yield model). In reality however, identification of the source may be uncertain because a given process step may be carried out on many similar equipment operating in parallel, or because similar defects may be produced by closely related steps before and after the suspected step. Since the contamination-defect relationship is assumed to be known, one can pre-evaluate the possible set of equipment for any combination of observed defect and particle type.

Equipment assignment is achieved by keeping a count, $E_{suspect}$, of the number of times a piece of equipment is held responsible for a defect in an analyzed die. In case of ambiguity a simple rule that can be used in the simulator is to hold all such suspected equipment equally responsible. Other types of rules that can be modeled are:

1. *Perfect diagnosis* i.e. the responsible equipment is always correctly identified,
2. Diagnosis aided by particle monitors and
3. Probability of incorrect diagnosis where a wrong piece of equipment is held responsible.

4.5.5 Issues in Simulating Defect Diagnosis Process

In practice, the efficiency and accuracy of failure analysis of defects in a semiconductor manufacturing depends on a number of attributes interacting in a complex manner. These can be broadly related to either design, testing, or particle monitoring attributes. Each parameter in the model proposed above is related to one of these classes. The ability to extract these parameters is an important issue which has not been addressed here or even in the available literature. The models presented for

defect diagnosis are believed to represent reality based on current understanding of the process [16, 17, 19, 20, 21]. Other aspects which have not been considered are the contribution of historical learning, and the role of the expertise of failure analysis personnel on accuracy and efficiency of diagnosis which, in reality is very important.

4.6 Simulation of Corrective Actions

The main objectives in modeling corrective actions are: first, to decide when a piece of equipment needs to be repaired or cleaned and when the equipment can be taken off-line (if required). Second, to estimate the new parameters of the particle model. The discussion here will be based on the particle model where equipment is the source. However, similar arguments can be extended to the models for other sources of particles.

4.6.1 Decision to Take Corrective Actions

Although, defect diagnosis or particle monitoring processes may hold a particular equipment suspect, it may not be necessary to take any action. First, taking equipment off-line too frequently can cause bottlenecks due to temporary decrease in line capacity. Second, confidence in defect diagnosis must be considered. The second requirement is achieved by keeping a count, $E_{suspect}$, of the number of times a piece of equipment is held responsible for a defect in the die fabricated as presented earlier. When this count exceeds a predefined threshold, E_{thresh}, the particular piece of equipment needs to be cleaned. Setting this threshold high means that the confidence in diagnosis is low.

The piece of equipment targeted for cleaning can be taken off-line in two ways. The first rule is to wait for the next scheduled maintenance period if the estimated waiting time, CL_{wait}, is less than a predefined interval of time. Otherwise, the second rule is applied where the piece of equipment is taken off-line as soon as it completes any on-

going processing step. More complex rules may be necessary when more than one piece of equipment operating in parallel needs to be taken off-line. This is achieved by defining a limit on the time interval, $CL_{interval}$, within which two separate pieces of equipment cannot be taken off-line. Figure 4.11 shows the flow chart for taking equipment off-line for applying corrections.

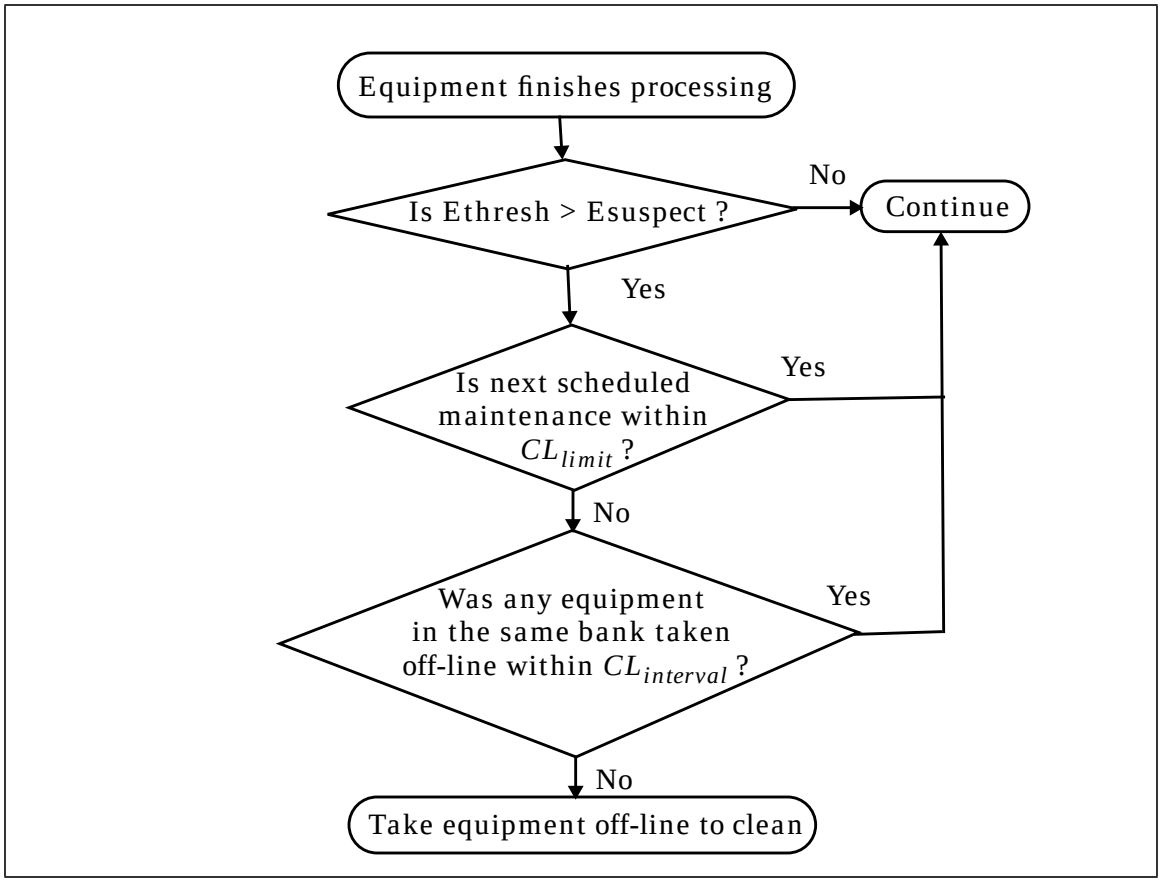

Figure 4.11 Taking equipment off-line.

4.6.2 Effect of Corrective Actions

It should be assumed that both the particle rate and the relative occurrences of different particle sizes change as a result of cleaning. It is further assumed that the new

distribution of particle number, N_c, is also a normal distribution with new mean and standard deviation given by:

$$\begin{aligned} m_{new} &= m_{old} \cdot k_m \\ \sigma_{new} &= \sigma_{old} \cdot k_\sigma \end{aligned} \tag{4.15}$$

where, k_m and k_σ are given constants between 0.0 and 1.0. A value of zero indicates that the operation of cleaning removes the source entirely. Similarly, we assume that the distribution of particle size, R_c, is still a polynomial distribution with new exponent p_{new} given by:

$$p_{new} = p_{old} + p_{diff} \tag{4.16}$$

where, p_{diff} is a positive constant. Note that the constant multiplier K (see Equation 4.3) of the new distribution must be adjusted in order for the integral of the distribution to be 1. This model implies that change in particle number is independent of change in size distribution. Though this is unlikely in reality, it is a reasonable assumption in the absence of any experimental data otherwise. The other underlying assumption in this model is that cleaning or repairing changes particle characteristics so as to reduce the rate of occurrence of defective die.

4.7 Cost Simulation

One of the accepted industry standards for estimating cost of manufacturing is the cost-of-ownership model developed by Sematech [22, 23]. In this model, the focus is on estimating the effective contribution of equipment and other resources to the wafer cost. This analytical model requires prior estimates of attributes like uptime, throughput yield, die yield, cycle times, etc. These parameters can be extracted by direct observation in a factory with varying degrees of confidence, but they cannot easily be extrapolated to different speculative scenarios. The cost model described in [24] gives a direct estimate of wafer cost arising out of equipment usage in a multi-product fab-

rication line. In the latter model, wafer cost is defined as being composed of two components: the first being direct equipment usage, and the second, the fraction of time during which a piece of equipment is not processing any wafers. The focus is on “fair” allocation of the cost incurred when equipment is idle in a multi-product facility. Here a variation of the latter cost estimation model is presented which can then be combined with yield estimates to model die cost.

4.7.1 Wafer Cost Model

Average wafer cost C_w can be expressed as the sum of three components: average cost due to equipment usage, C_{equip}, average cost due to wafers waiting between processing steps (in input queues, for example), C_{wait}, and, any fixed cost C_{fixed}. The averaging is done over a fixed interval of time. Thus, the wafer cost can be expressed as:

$$C_w = C_{equip} + C_{wait} + C_{fixed} \tag{4.17}$$

C_{equip} is calculated in a manner similar to as presented in [24] and is given by:

$$C_{equip} = C_{active} + C_{inactive} \tag{4.18}$$

where, C_{active} is the average cost contribution for active usage of the equipment and $C_{inactive}$ is the average cost contribution from equipment when no wafers are being processed (equipment may be idling or off-line). C_{active} in turn is given by:

$$C_{active} = T_{active} \cdot K_{active} \tag{4.19}$$

where, T_{active} is the average amount of time wafers of a particular product are processed in the piece of equipment under consideration and K_{active} is the cost of utilizing the equipment per unit time. Note that K_{active} must include the contribution of capital cost and the operating cost of the equipment. $C_{inactive}$ can also be estimated in a similar way by expressing it as:

$$C_{inactive} = T_{inactive} \cdot K_{inactive} \tag{4.20}$$

where, special attention must be paid to estimating $T_{inactive}$ in a fair manner. Similar to that presented in [24] $T_{inactive}$ can be estimated as a fraction of the total inactive time in such a way that the corresponding T_{active} is also the same fraction of the total active time of the equipment. $K_{inactive}$ then must take into account not only the contribution of fixed cost but also the cost of repairing, etc., if any.

C_{wait} should be calculated as:

$$C_{wait} = T_{wait} \cdot K_{wait} \tag{4.21}$$

where, T_{wait} is the total time a wafer is waiting in the fabrication line without being actively processed and K_{wait} is the sum total of cost per unit time of any variable costs when the wafers are waiting. An increase in cycle time means that more resources are necessary for handling larger inventory, delay in product delivery, and as presented in [25], yield can also suffer. Thus, K_{wait} can be seen as a penalty cost for any increase in cycle time. (T_{wait} can theoretically be minimized and most efforts to reduce cycle times have been focused on reducing T_{wait} [26, 27, 28].)

Finally, C_{fixed} is composed of the raw cost of the wafer (cost of bare silicon) and any other cost that should be associated with every wafer. Cost per minute parameters in the above model need to be extracted out of cost data like capital cost, depreciation rates, operational cost, etc. which are usually well characterized data.

4.7.2 Die Cost

Having modeled wafer cost, one can estimate the cost of good ICs by taking into consideration the average yield during a given period of time. If, for a given product, average yield is denoted by Y, and the number of chips per wafer is N_{chip}, then the cost of a good chip can be expressed as:

$$C_{chip} = C_w \cdot Y \cdot N_{chip} \tag{4.22}$$

In the above model for wafer cost and consequently, die cost, it is tacitly assumed that the cost contribution of the failure analysis equipment and resources can be cal-

culated in exactly the same manner. However, when only a subset of products are subjected to failure analysis, wafer cost for these products will be unnecessarily inflated. Such inflated estimates can be avoided if one allocates the entire cost of failure analysis to all the products. This assumption is reasonable since failure analysis can be viewed as a common resource to "debug" the entire manufacturing line.

4.7.3 Cost of Manufacturing

If N_w denotes the number of wafers manufactured during a given time period with an average wafer cost of C_w, the total cost of manufacturing, C_{total}, is given by:

$$C_{total} = C_w \cdot N_w \tag{4.23}$$

This estimate of the total cost of manufacturing serves as an important reference point for comparing different scenarios. One can, for example, compare different wafer release policies or scheduling rules. Of course, an increase in cost of manufacturing does not necessarily mean the overall productivity of the manufacturing facility is any worse. For example, an increase in failure analysis capacity will increase the cost of manufacturing, but at the same time the yield learning rate may improve producing more good ICs to sell. Examples presented later illustrate both the cases: first, where cost of manufacturing alone is sufficient to compare scenarios and second, where yield must be taken into account too.

References

[1] *ManSim X*, User Manual, Tyecin Systems Inc., San Jose, CA, 1995.

[2] J. B. Khare, *Contamination-Defect-Fault Relationship - Modeling and Simulation*, Ph.D. Dissertation, Carnegie Mellon University, Nov 1995.

[3] C. H. Stapper and R. J. Rosner, "Integrated Circuit Yield Management and Yield Analysis: Development and Implementation", *Trans. on Semiconductor Manufacturing*, vol. 8, no.2, pp. 95-102, May 1995.

[4] T. L. Michalka, R. C. Varshney, and J. D. Meindl, "A Discussion of Yield Modeling with Defect Clustering, Circuit Repair, and Circuit Redundancy", *Trans. on Semiconductor Manufacturing*, vol. 3 no. 3, pp. 116-127, Aug. 1990.

[5] R. Glang, "Defect Size Distribution in VLSI Chips", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 265-269, Nov. 1991.

[6] J. B. Khare, W. Maly and M. E. Thomas, "Extraction of Defect Size Distributions in an IC layer Using Test Structure Data", *Trans. on Semiconductor Manufacturing*, vol. 7, no. 3, pp. 354-368, Aug. 1994.

[7] C. H. Stapper, "Statistics Associated with Spatial Fault Simulation Used for Evaluating Integrated Circuit Yield Enhancement", *Trans. on Computer-Aided Design*, vol. 10, no. 3, pp. 399-406, March 1991

[8] R. A. Governal, A. Bonner, and F. Shadman, "Effect of Component Interactions on the Removal of Organic Impurities in Ultrapure Water Systems", *Trans. on Semiconductor Manufacturing*, vol. 4 no. 4, pp. 298-303, Nov. 1991.

[9] M. Itano, F. W. Kern, Jr., R. W. Rosenberg, M. Miyashita, I. Kawanabe, and T. Ohmi, "Particle Deposition and Removal in Wet Cleaning Processes for ULSI Manufacturing", *Trans. on Semiconductor Manufacturing*, vol. 5 no. 2, pp. 114-120, May 1992.

[10] D. M. H. Walker and S. W. Director, "VLASIC: A Catastrophic Fault Yield Simulator for Integrated Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 5 no. 4, pp. 541-556, October 1986.

[11] D. D. Gaitonde and D. M. H Walker, "Hierarchical Mapping of Spot Defects to Catastrophic Faults - Design and Applications", *IEEE Trans. on Semiconductor Manufacturing*, vol. 8. no. 2, pp. 160-166, May 1995.

[12] W. Maly and J. Deszczka, "Yield Estimation Model for VLSI Artwork Evaluation", *Electron Letters*, vol. 19, no. 6, pp. 226-227, March 1983.

[13] W. Maly, "Modeling of Lithography Related Yield Losses for CAD of VLSI Circuits", *Trans. on Computer-Aided Design*, vol. 5, no. 3, pp/ 166-177, March 1985.

[14] P. K. Nag and W. Maly, "Hierarchical Extraction of Critical Area for Shorts in Very Large ICs", *Proc. of Int. Workshop on Defect and fault Tolerance in VLSI Systems (DFT)*, pp. 19-27, Nov. 1995.

[15] *TestSim X*, User Manual, Tyecin Systems Inc., San Jose, CA, 1995.

[16] Dr. Splittgerber, *Personal Communications*, Siemens AG, Munich.

[17] Mr. Melzner, *Personal Communications*, Siemens AG, Munich.

[18] *Close Up - Wafer Inspection*, Tencor Instruments, vol.1, no. 2, Spring 1995.

[19] Dr. Schafer, *Personal Communications*, Siemens AG, Munich.

[20] Dr. Zeller, *Personal Communications*, Siemens AG, Munich.

[21] Mr. Hess, *Personal Communications*, Siemens AG, Munich.

[22] J. Crest and P. Burggraaf, "The Reasoning Behind Cost of Ownership", *Semiconductor International*, pp. 56-60, May, 1993.

[23] E. Neacy et. al., "Cost Analysis for Multiple Product/Multiple Process Factory: Application of SEMATECH's Future Factory Design Methodology", *1993 Advanced Semiconductor Manufacturing Conf. (ASMC) Proc.*, pp. 212-219, Oct. 1993.

[24] W. Maly, H. Jacobs, and A. Kersch, "Estimation of Wafer Cost for Technology Design", *Proc. of 1993 IEDM, pp. 35.6.1-35.6.4*, Washington, D.C., Dec. 1993.

[25] L. M. Wein, "On the Relationship Between Yield and Cycle Time in Semiconductor Wafer Fabrication", IEEE Trans. on Semiconductor Manufacturing, vol. 5, no.2, pp. 156-158, May, 1992.

[26] B. Ehteshami, R. G. Petrakian, and P. M. Shabe, "Trade-Offs in Cycle Time Management: Hot Lots", IEEE Trans. on Semiconductor Manufacturing, vol. 5, no. 2, pp.101-106, May, 1992.

[27] S. C. Wood and K. C. Saraswat, "Factors Affecting The Economic Performance of Cluster-Based Fabs", Electrochemical Society Third Int. Symp. on ULSI Science and Technology, Washington, D.C., May 1991.

[28] S. C. H. Lu, D. Ramaswamy, and P. R. Kumar, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing", IEEE Trans. on Semiconductor Manufacturing, vol. 7, no. 3, pp. 374-388, Aug. 1994.

Chapter 5

Yield Learning Simulator -Y4

The methodology and the models for yield learning described in the previous chapter has been implemented in a program -Y4 (an acronym for Yield Forecaster). Y4 can be used in two ways: as a stand-alone simulator using internal models to mimic a fabrication line or, as a library of routines with externally implemented user models for custom simulations. In this chapter, the structure of the prototype simulator Y4 is presented and some of the important features of its modules are described.

5.1 Implementation Structure

Figure 5.1 shows the overall structure of the Y4 framework which simulates cost and yield learning curves. The heart of Y4 is the event handler which communicates with six modules: The wafer movement simulator (WSIM), the yield simulator (YSIM), the failure analysis simulator (FASIM), the in-line particle monitor simulator (PSIM), the cost simulator (COSIM), and, the probe tester simulator (TSIM). The operation of the event handler and these six modules can be controlled through the simulation control unit. In addition, the user can implement different models with the help of the toolkit of functions to access and modify the common database for all the modules including the event handling routines. A basic user interface is also available to read input files for the models, write output of statistics gathered, and, customize the simulation control strategy.

The events are maintained in a balanced binary tree sorted by increasing value of time. Since more than one event can occur at the same instance, each node of the tree

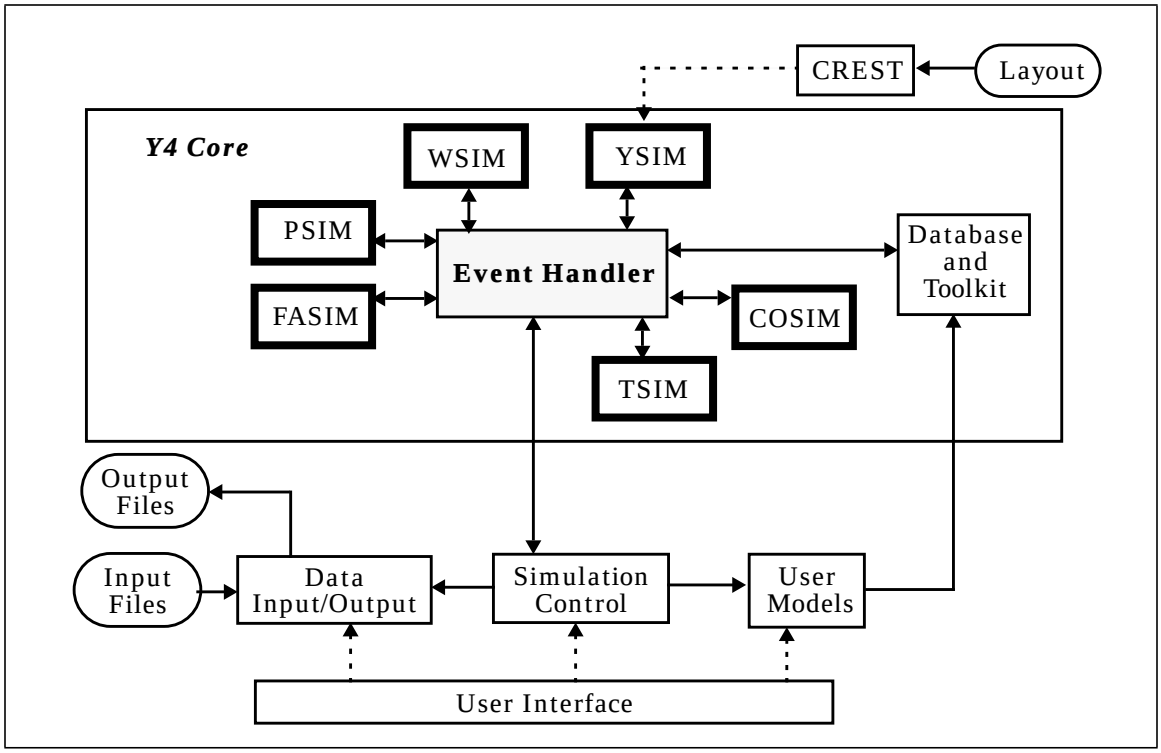

Figure 5.1 Top level structure of the Y4 framework.

contains a list of events with the same time value and sorted in the order they were generated. Zero delay events are appended to the end of the current list of events in the order they are generated and this alleviates the problem of concurrent events [1]. All communications and interdependence between the modules of Y4 are handled by sending appropriate events of specific types.

The models described in Chapter 4 are the internal models of Y4 and are implemented in its submodules (WSIM, etc.). Although these models are based on reasonable simplifying assumptions, not every situation and factory environment can be described by a single set of models. In order to be flexible enough, external models can be implemented and linked to Y4 to expand the repertoire of available models. However, the general methodology to predict cost and yield learning curves is intended to remain the same.

5.2 WSIM

The wafer movement simulator, WSIM, is composed of various factory entities and models presented in Section 4.1 and is shown in Figure 5.2. The simulator is governed by a controller which modifies the states of factory entities such as equipment, products, wafers and personnel. Scheduling of wafers is governed by a set of factory operating rules such as wafer release policy, dispatch, setup, and load rules, etc. Statistics are collected and passed on to the simulation control for updating the database and output. Y4 can be used with WSIM alone for cycle time and throughput analysis. WSIM is similar to ManSim (a commercial simulator [2] to perform semiconductor factory analyses) in its capability. In fact, a comparison experiment showed them to be in excellent agreement (within 1% of estimated cycle times for a variety of scenarios).

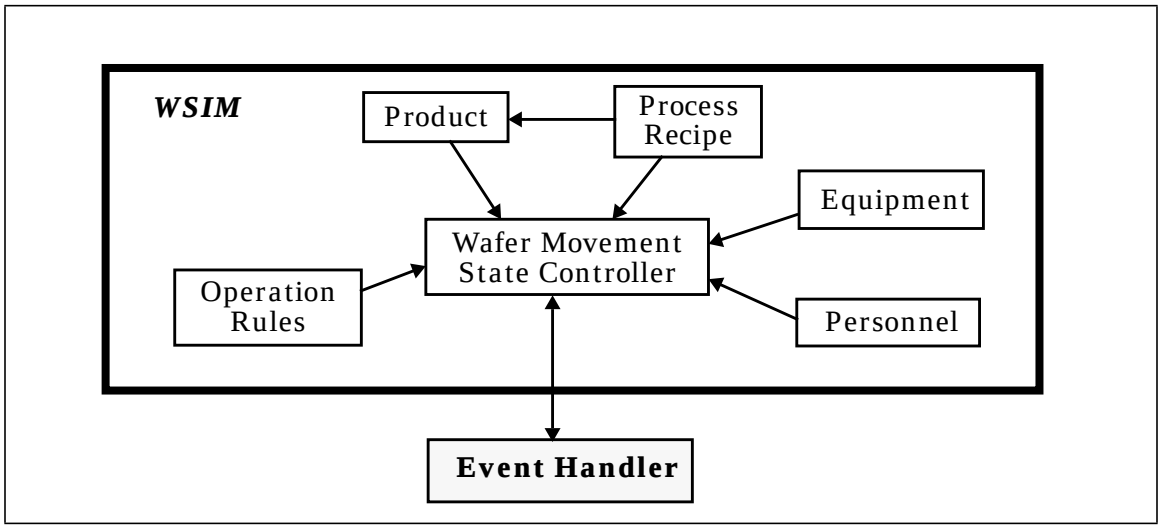

Figure 5.2 Wafer Movement Simulator - WSIM.

There are two main *time-lines* of events that occur in the simulated wafer fabrication line. The first one is initiated by the process of introduction of wafers in the fabrication line according to some release policy. Most of the events in the factory are initiated in some fashion due to the movement of wafers. The second time line is an independent sequence of events generated by the process of unscheduled (asynchro-

nous) breakdown of equipment. It is assumed that if such equipment breakdown occurs during processing, the partly processed wafers have to be rejected and further processing of new lots takes place only when the piece of equipment is back on line.

5.3 YSIM

The structure of the yield simulator, YSIM, is shown in Figure 5.3. Particle formation is mimicked using random numbers. The particle number generator produces N_c particles where N_c is distributed normally (Equation 4.2) for each wafer. Then each generated particle is assigned a random size between R_{min} and R_{max} (Equation 4.3) taken from a polynomial distribution. One can also simulate the lot to lot variation in number of particles by activating a third random number generator in one of the following ways:

1. generate random means and standard deviations using a gaussian number generator for use as an input to the particle generator for wafers.
2. generate a random multiplier, m_c, taken from a predefined distribution function. The new number of particles is simply then $m_c N_c$.

All the random number generators and statistics estimators are adapted from algorithms in [3, 4].

Depending on the step of the recipe being performed, the particle to defect mapper performs the proper translation to defects (Equations 4.4 and 4.5). Similarly, the defect to fault mapper uses the critical area of the fault (from the layout) to assign a fault, if any, to each defect (Equations 4.6, 4.7 and 4.8). The interface between the critical area extractor - CREST [5] - and Y4 is rudimentary and is essentially accomplished through input files. CREST can currently extract critical area for shorts in interconnects for fairly large designs.

YSIM works with the wafer movement simulator to introduce particles on wafers and estimate the yield of each fabricated wafer obtained from WSIM. Further, the

parameters of the particle generator can be dynamically controlled by a cleaning/ repairing model such as the one presented earlier (Equations 4.15 and 4.16). One could also use YSIM purely as a static yield simulator without changing the parameters of the particle generators. This is useful for estimating yield and its distributions for a stable manufacturing line where particle distribution parameters do not change with time.

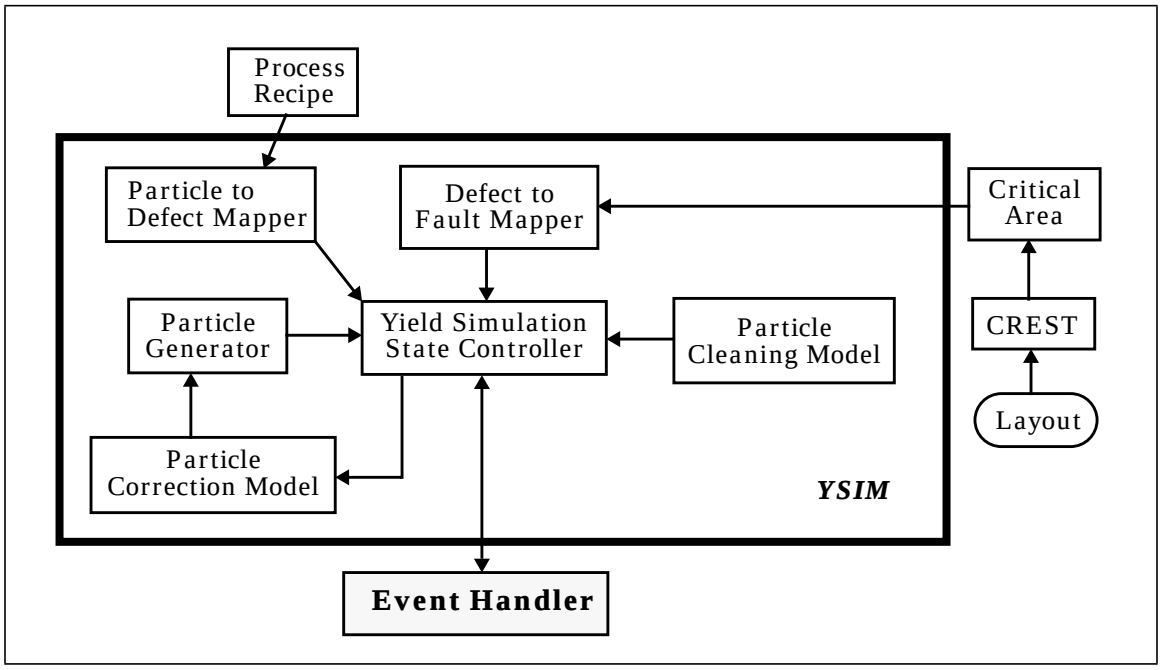

Figure 5.3 Yield Simulator - YSIM.

5.4 TSIM

The purpose of the test simulator - TSIM - shown in Figure 5.4 is to be able to predict the time required for testing and to model less than 100% fault coverage situations. The input to the simulator is a wafer lot of a product with some dies marked defective, and the list of faults that have occurred. The sequencer schedules the wafers into available testing equipment and also steps through individual die to estimate time. The time estimation is achieved by a model which requires the list of possible faults

and the distribution of testing time (Equations 4.9, 4.10 and 4.11). In scheduling testing equipment both personnel and rules are taken into account. The fault list and coverage values for each fault defined for a product is used to determine which of the defective die tested is found to be faulty. The simulator can be switched off in which case, a pseudo simulation is done with zero testing time and 100% fault coverage.

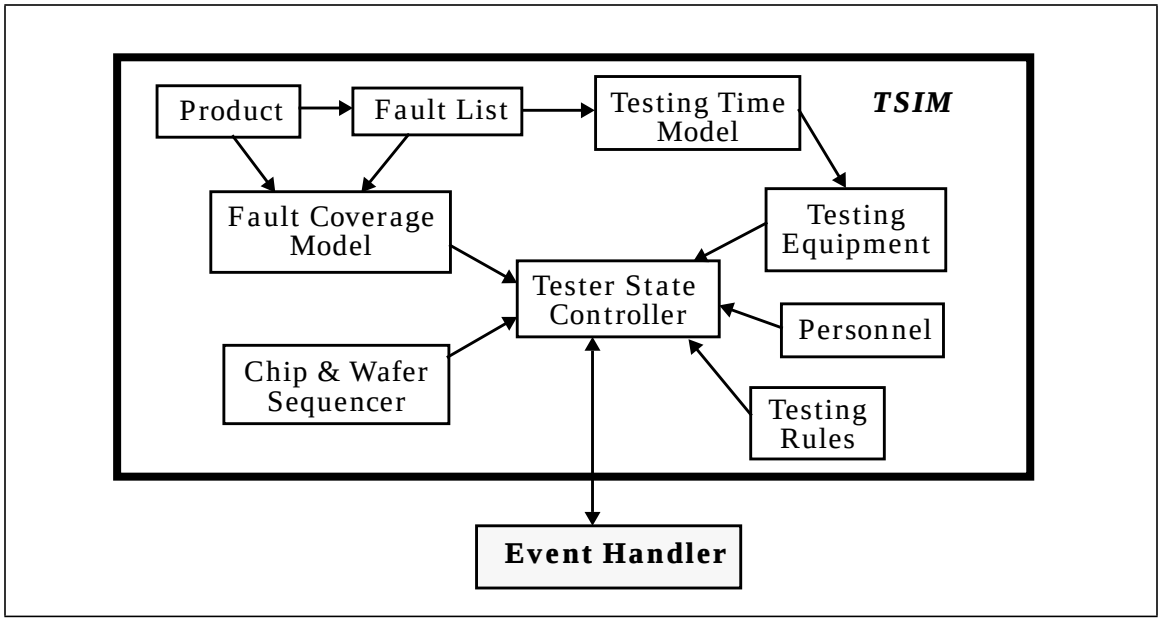

Figure 5.4 Tester Simulator - TSIM.

5.5 PSIM

The particle scanning module or PSIM mimics the efficiency and accuracy of a particle scanner; its sub-components are shown in Figure 5.5. The two main components used are the detectability model which determines whether a particle in a scanned area is detected (Equation 4.12), and the timing model which estimates the time required to scan a single wafer. The data structures associated with each scanned chip is also updated to reflect that certain particles are successfully detected. Subsequent control of the fabrication line depends upon the wafer rejection and equipment correction criteria applied. Note that this requires the functioning of PSIM to be intimately

tied to WSIM. PSIM can also be switched off, either completely or partially by deactivating the detectability model and simulating particle scanners purely as normal equipment.

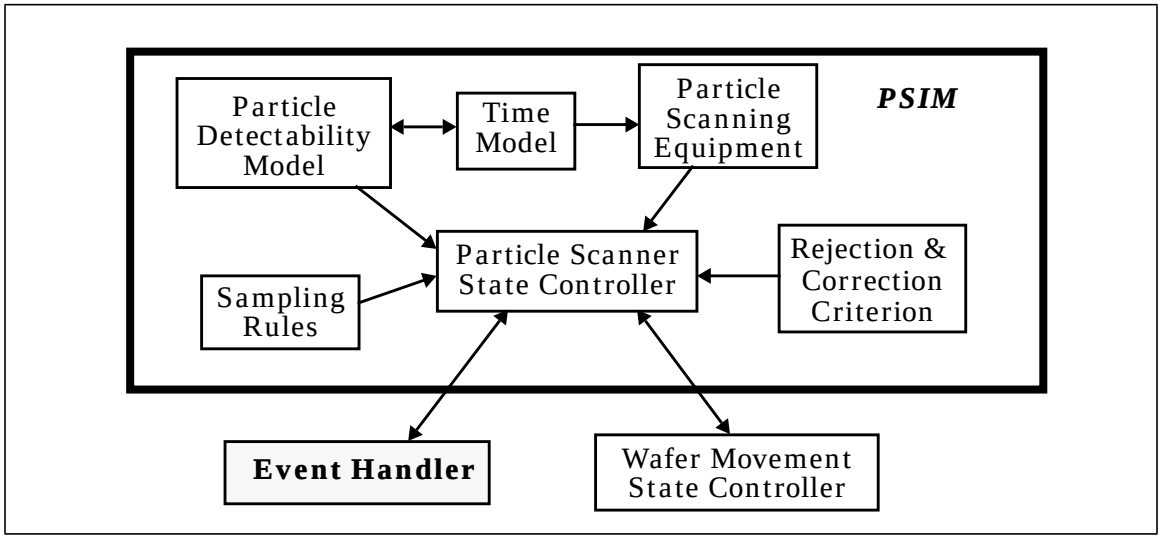

Figure 5.5 Particle Monitor Simulator - PSIM.

5.6 FASIM

The failure analysis simulator, FASIM, consists of three parts: a sampling model, a diagnosability measure estimator, and a sequencer of chips through different analysis equipment. Operational policies control both the sampling model and the chip sequencer by either discarding or accepting chips for further analysis using the diagnosability measure. Initial estimates for the diagnosability measure can be modeled as a function of the layout design or as plain inputs (Equation 4.14). The updated measures depend on both the design and the equipment characteristics as postulated earlier (Equation 4.13). The structure of FASIM is shown in Figure 5.6. Note that the operation of FASIM depends on WSIM, the wafer movement simulator, which produces wafers with defects. It also depends on YSIM which provides information on particle sources necessary for correctly assigning the source of analyzed defects to the

set of equipment responsible. Mapping of a defect to contamination and finally to its source is computed internally and used in assigning the source of observed defects. One can, however, provide this information externally to simulate incomplete information on source of defects.

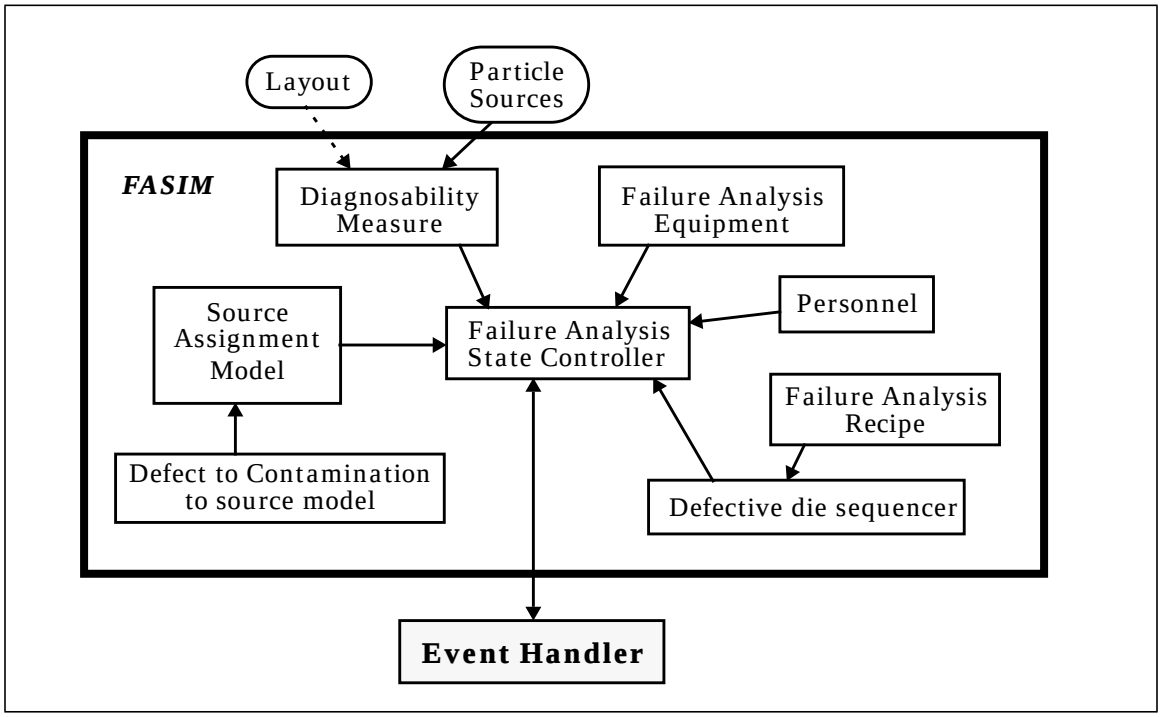

Figure 5.6 Failure Analysis Simulator - FASIM.

5.7 COSIM

Figure 5.7 shows the cost simulator COSIM, which works first with the WSIM to estimate the cost of the wafer, and then with YSIM to estimate the cost of good die. Relevant statistics such as equipment utilization, down times, idle times, etc. are extracted from WSIM (and PSIM, TSIM, and FASIM as required) and stored in the database. Values of cost per unit time are used for all these factors to assign a cost to each wafer for every product manufactured. The cost amortizer then uses the model presented earlier (Equations 4.17, 4.18, 4.19, 4.20 and 4.21) to fairly allocate cost con-

tributions to the various factors. The costs of testing and failure analysis can also be amortized in a predefined manner by using the usage statistics from TSIM, PSIM and FASIM. Yield results are used to compute die cost per wafer for a given period (week, month, etc.) of time (Equation 4.22). Cost of manufacturing is obtained by integrating all input costs (Equation 4.23).

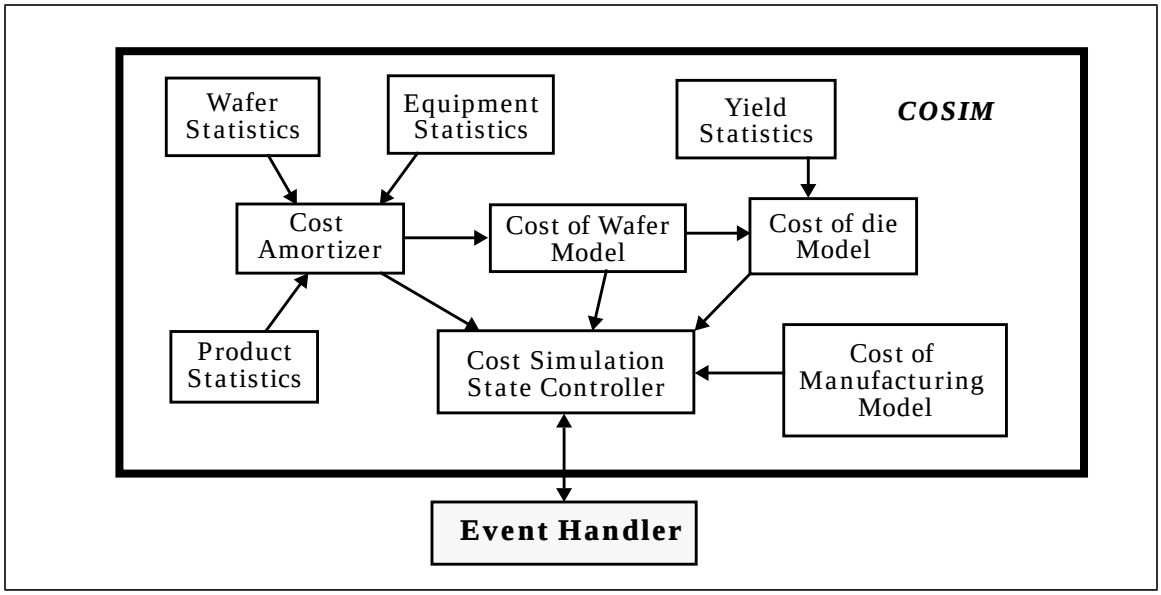

Figure 5.7 Cost Simulator - COSIM.

References

[1] J. Banks and J. S. Carson II, *Discrete Event System Simulation*, Prentice-Hall, Engelwood Cliffs, New Jersey, 1984.

[2] *ManSim X*, User Manual, Tyecin Systems Inc, San Jose, CA, 1995.

[3] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C - The Art of Scientific Computing,* Cambridge University Press, 1990.

[4] D. E. Knuth, Seminumerical Algorithms, 2nd. ed., vol.2 of The Art of Computer Programming, Addison-Wesley, Reading, MA, 1981.

[5] P. K. Nag and W. Maly, “Hierarchical Extraction of Critical Area for Shorts in Very Large ICs”, *Proc. of Int. Workshop on Defect and fault Tolerance in VLSI Systems (DFT)*, pp. 19-27, Nov. 1995.

Chapter 6

Basic Capabilities of Y4

In this chapter, results of a number of simulation experiments replicating known phenomena in a manufacturing line are presented to illustrate the capabilities of Y4. In designing the experiments, an attempt has been made to exercise each of the modules (WSIM, YSIM, TSIM, PSIM, FASIM and COSIM) in such a way as to demonstrate their features and properties individually.

First, cycle times and throughput analyses of a single and a two product manufacturing line are presented. Subsequent simulation results illustrate the general difference in wafer cost of single and two product factories. Then simulations of the impact of imperfect testing on escape rate are presented. Particle monitor simulations are illustrated along with wafer rejection in the case where there is a high yield variance. A simulation example of yield learning for a single product factory with defect diagnosis follows. This is followed by simulation of yield learning using particle monitors alone.

The process recipes, equipment and cost data used in these examples were taken from an existing manufacturing line. To design manufacturing lines with different capacities, the equipment set had to be altered to meet desired capacity requirements. Operators in the manufacturing line were not simulated, and thus, any variability in observed cycle times and cost is solely due to the equipment. The duration of simulation in each case is at least one year. The fabrication line is assumed to be empty at the beginning of each simulation because of limitations of the current implementation. Statistics are collected after the first 12 weeks of simulation (*warm-up* period) and

therefore the conclusions drawn are not biased by initial variations occurring as a result of this initially empty line assumption.

The following examples use mainly for a 0.5 micron 3 metal CMOS process recipe. The original recipe from an existing manufacturing line has been modified by merging steps that logically define a layer into a single step. An example would be a lithography step which is actually composed of resist spin, bake and expose steps. The original equipment step has also been changed in order to reflect the merging of the process steps. These modifications result in cost and cycle times that are nearly the same as the original. Modification was necessary to protect the proprietary nature of the original data. The modified recipe consists of 145 steps using 183 pieces of equipment for a 2496 wafer starts per week (WSPW in short) capacity factory (a medium sized factory).

Some examples also use a 0.5 micron, 2-metal, trench capacitor, DRAM process. After preprocessing, this recipe consists of 174 steps using 214 pieces of equipment also with a capacity of 2496 WSPW. The modified process recipes and equipment set are given in Appendices A and B, respectively.

6.1 Cycle Time and Throughput Analysis

Figure 6.1 shows the average cycle time (in minutes) and mean throughput rate (in wafers per week) versus the wafer start rate for the CMOS factory. As expected, cycle time increases as wafer starts per week (WSPW) is increased. The rate of increase is rapid when input WSPW is greater than the capacity of the line (2496 WSPW).The fabrication line is unstable since the throughput rate saturates and excess wafers in the line cause inventory to build up rapidly. Normally a factory should never be operated in such an unstable region. The opposite case is when the factory is under-utilized by having a small number of wafer starts per week compared to the line capacity.

Here, the mean cycle time is very close to the theoretical raw processing time (RPT) since human operators have not been taken into account.

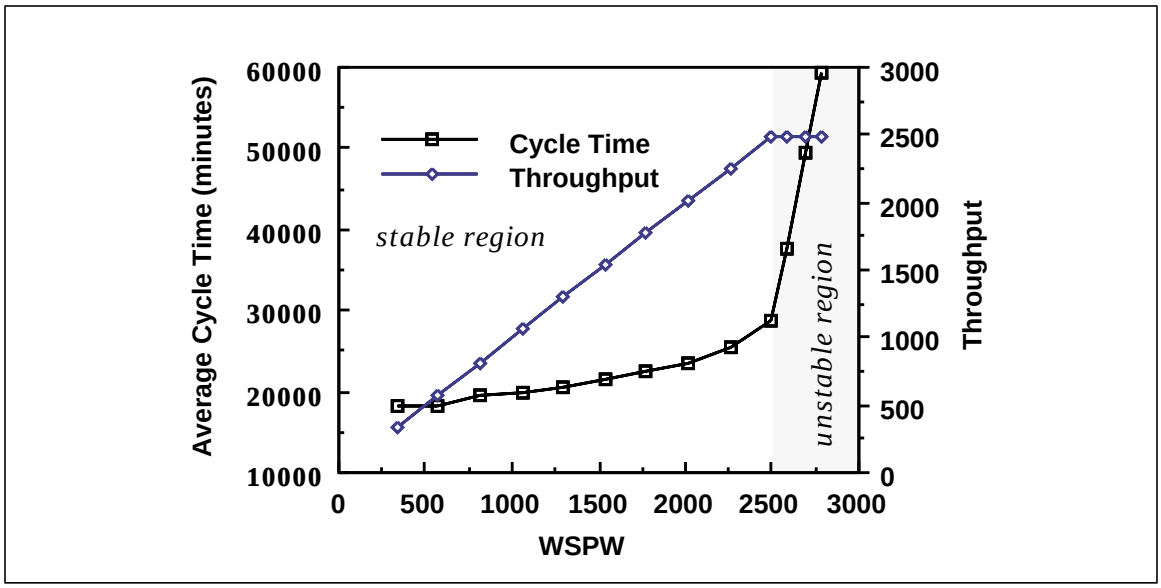

Figure 6.1 Cycle Time and Throughput of CMOS factory.

The DRAM factory shows the same trend as the CMOS factory as illustrated in Figure 6.2. The difference lies in the value of the cycle times and achievable throughput

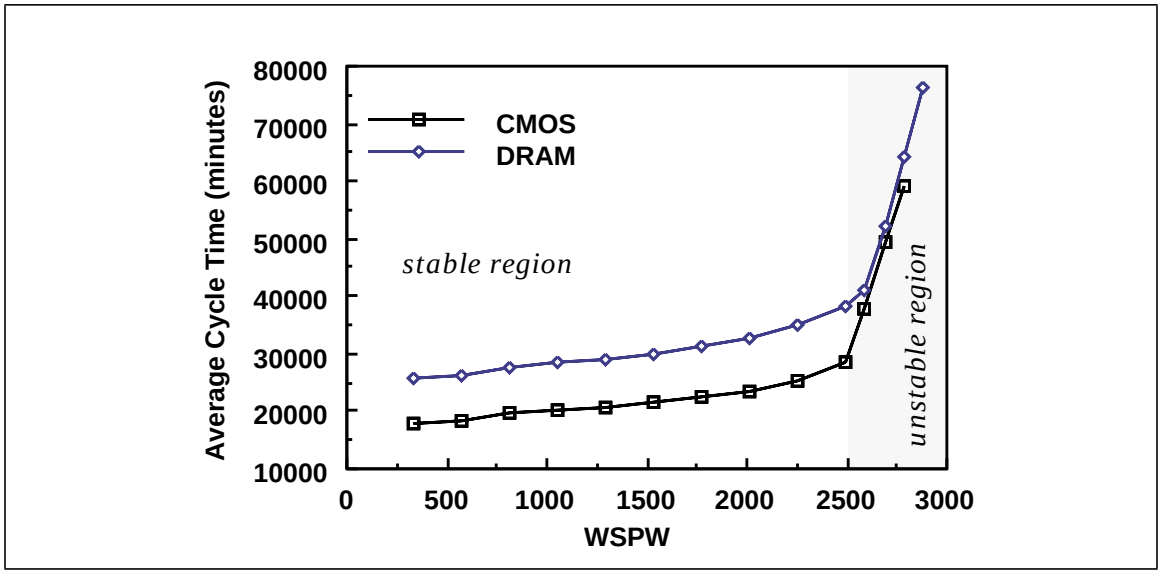

Figure 6.2 Cycle Time and Throughput comparison of DRAM vs. CMOS factories.

rate. The raw processing time is 25607 minutes for DRAM and 18106 minutes for CMOS. The variance in cycle time is also a function of the input WSPW as shown in Figure 6.3. Variance increases as the input WSPW is increased and is due to the fact that queues become larger. The increase in variance in the stable region is due to the likelihood of having a full load available for batch equipment being higher.

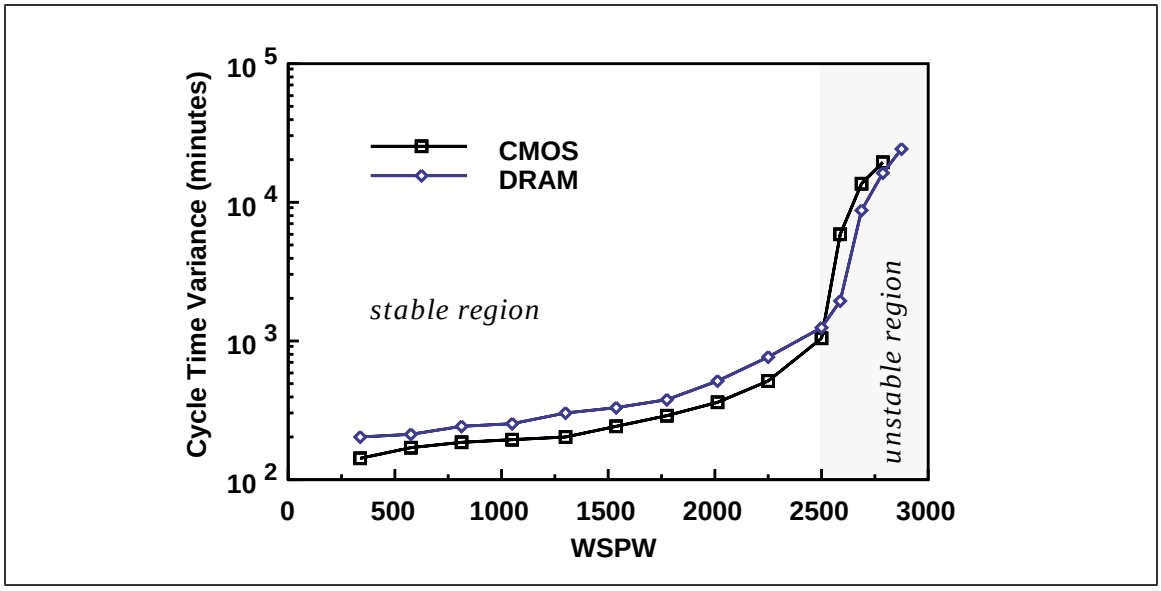

Figure 6.3 Variance in cycle time comparison for DRAM vs. CMOS factories.

In a multi-product factory, scheduling of wafers for various products are inter-dependent, hence, the cycle time of each product could be affected by other products. Amongst many factors, cycle time is affected by the proportion of each product, or product mix, being manufactured. To illustrate this dependence, a two-product factory was designed with a capacity of 832 and 1664 WSPW (for a total of 2496 WSPW) for the CMOS and DRAM products, respectively. This factory has 222 pieces of equipment and has been derived from the original DRAM factory with minor modifications. The designed proportion of wafer start rate is 33% of CMOS product, the total being 2496 WSPW. In this experiment the proportion of CMOS product is varied from 10% to 90% of total wafer starts.

Figure 6.4 shows the cycle time of each product as a function of the percentage of the CMOS product. The manufacturing line shows two distinct regions of operation: a sta-

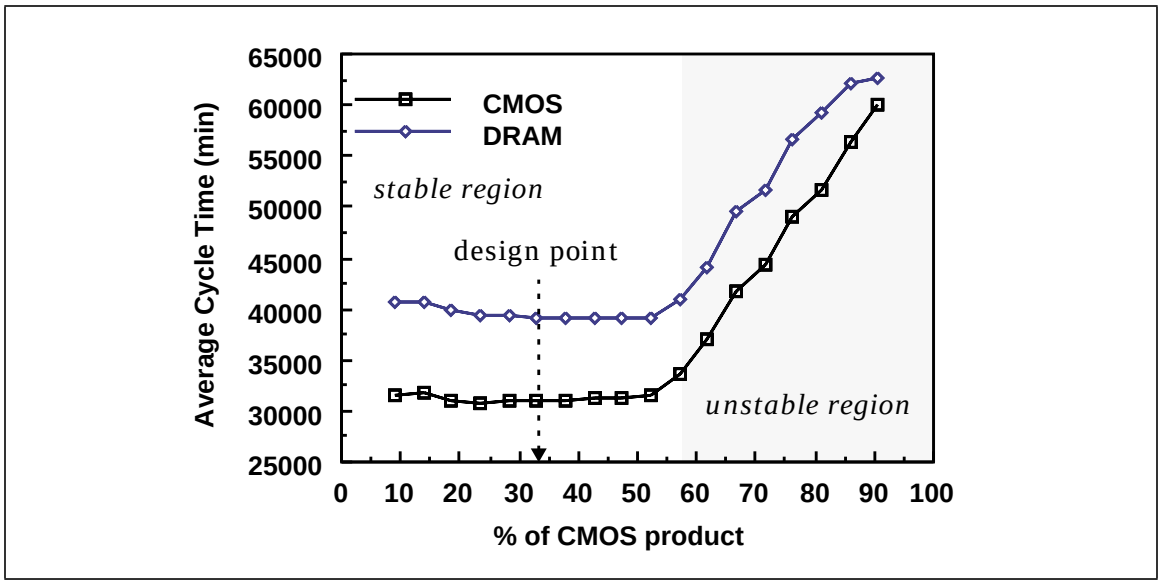

Figure 6.4 Cycle Time of two product factory (CMOS and DRAM).

ble and an unstable region. The stable region is centered around the design point of the manufacturing line. In this region, the cycle time of CMOS product increases slowly as the proportion of it is increased. The cycle time of DRAM product, on the other hand decreases slowly in this range. This factory is derived from the DRAM line, hence, the operation of the line is dominated by the DRAM process. Therefore, the cycle time of CMOS product increases as its proportion is increased. In the unstable region cycle times of both products increase rapidly since the capacity available is not enough to fabricate such a large proportion of CMOS product. In general, one can expect unstable operating conditions on both extremes around the design point depending on the particular organization and design of the factory. However, this simulation experiment illustrates the applicability of simulation tools in assessing the operating flexibility of a multi-product facility.

6.2 Analysis of Wafer Cost

In this section, wafer cost estimates are presented for the manufacturing lines illustrated in the previous section. Figure 6.5 shows the cost of wafer for the CMOS product as a function of input wafer start rate. The cost of wafer decreases nearly inversely with increasing wafer start rate as long as the factory is operated within its designed capacity. The minimum value attained is $2845. Beyond the line capacity, wafer cost increases a little corresponding to a small drop in equipment utilization. Long queues of wafers waiting at the bottleneck equipment cause other equipment to starve. Note that, in these experiments the value of K_{wait} in Equation 4.21 is zero and, hence, the increase in cost is entirely due to decrease in utilization and not due to increase in waiting time, T_{wait}. Similar results were obtained for the DRAM product except that the minimum wafer cost obtained is $3532, mainly because the DRAM process requires expensive equipment to define the trench capacitors and executes more lithography (mask) steps.

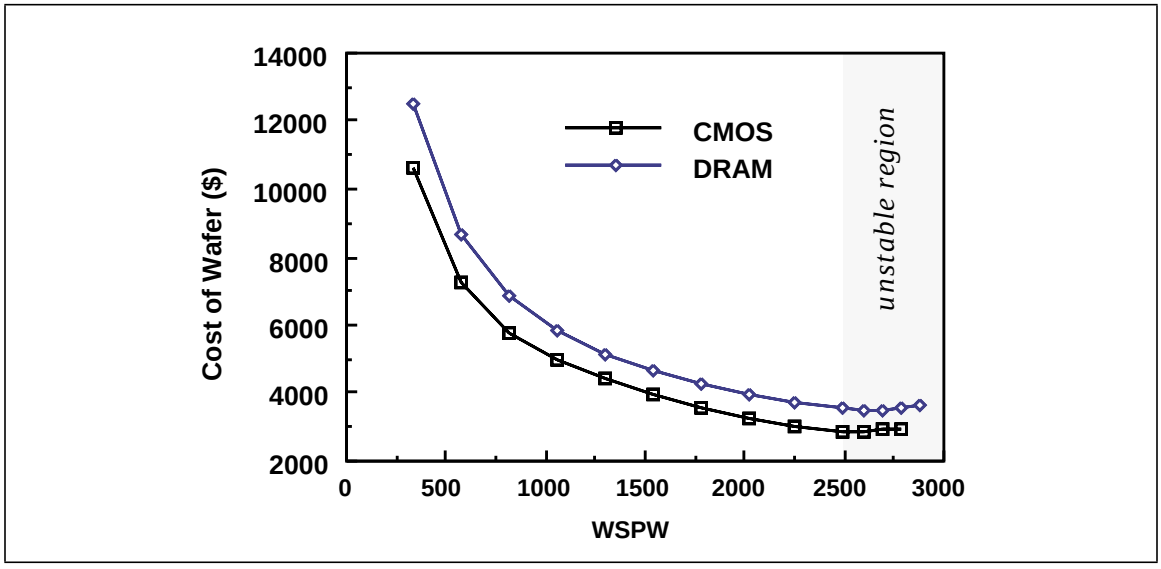

Figure 6.5 Cost of Wafer vs. volume for CMOS and DRAM factories.

Figure 6.6 shows estimates of cost of wafer as a function of the product mix for the two-product factory presented earlier. In the stable operating region, the cost of CMOS product decreases as its proportion is increased due to better utilization of the CMOS processing capacity (same as in a single product factory). Exactly the opposite is true for the DRAM product. In the "fair" allocation cost model, cost incurred due to the idle times of equipment used by only one process is wholly allocated to the corresponding product. This effect is more pronounced for the DRAM product since it requires specialized equipment (for trench capacitors and epitaxial layers) not required by the CMOS process.

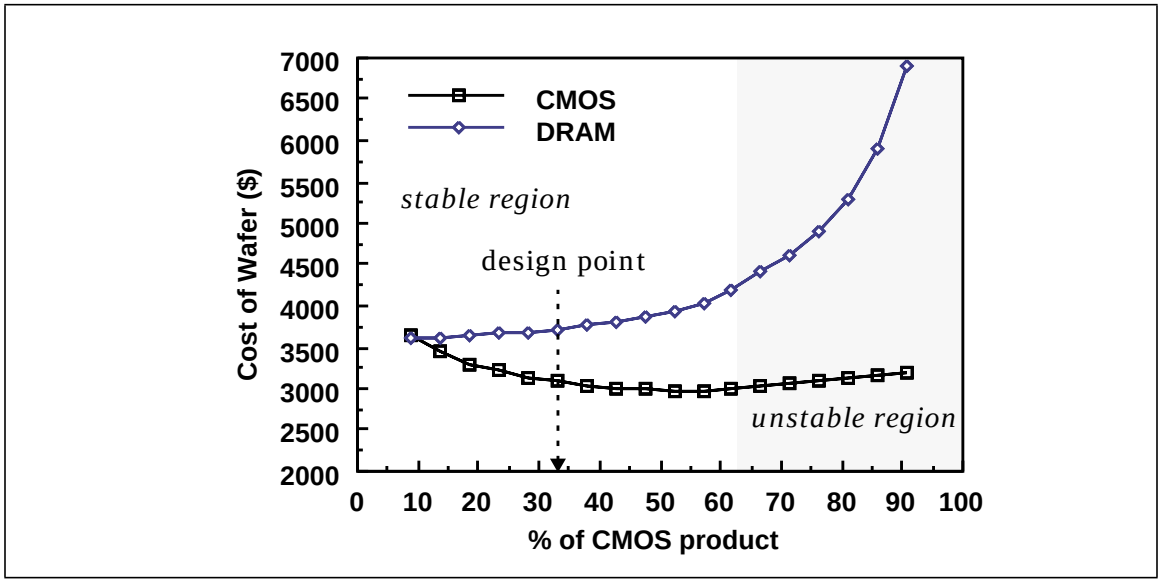

Figure 6.6 Cost of Wafer vs. product mix.

In the unstable operating region the cost of wafer for both products increases as the proportion of CMOS product is increased. The cost of CMOS product increases because of starvation of non-bottleneck equipment in spite of the fact that more wafers are being produced. In fact, the throughput of the CMOS product is no longer equal to the input wafer start rate as one may expect. Instead, it is less than the input start rate because of capacity limitations. The cost of DRAM product, on the other hand,

increases dramatically mainly because of under-utilization of the dedicated equipment. Notice that a similar effect is not apparent for the CMOS product at low start rates since the CMOS process does not have any significant (expensive) equipment dedicated to it. Almost all the process steps of CMOS are shared by the DRAM process leading to uniformly high utilization of the shared equipment.

6.3 Static Yield Estimation

The fabrication line described above for the 0.5 micron 3 metal CMOS process is also used to illustrate yield simulation with Y4. Line capacity is 2496 WSPW and wafer diameter is taken to be 150 mm. Only defects in the polysilicon and the three metal layers are considered as dominant yield detractors. Defects in the polysilicon layer are assumed to be the result of particles introduced during the poly deposition step. Defects in metal layers are assumed to be due to particles generated at the common sputtering step. It is also assumed that these defects result in shorts in the respective layers.

Critical area, as a function of defect sizes, for a product must be extracted from its layout as noted in Chapter 4. In this case, it is assumed that the critical areas for each defect type must closely resemble those of a modern microprocessor design. Since a typical layout was unavailable certain reasonable assumptions were made. Sensitivity of the polysilicon layer to shorts is assumed to be lower than the three metal layers. Using this assumption, the critical areas were derived by appropriately scaling the corresponding critical areas obtained from 24-Port Register File [1] and 32x32 Crossbar Switch [2] both designed with a 1.5 micron design rule. The results for the two designs were obtained using the software CREST and are presented in [3]. The critical area functions for each of the defect types scaled to a 0.6 micron design rule are given in Appendix C. For this assumed design, the chip size is 3 cm^2 and the number of usable die per wafer is 50.

The mean and variance of number of defects (Equation 4.2) for each of the 4 defect types are set to be 40 and 200 resulting in a mean density of 0.3 defects/cm^2. The experiment was repeated for a number of values for the exponent, *p*, of the size distribution (Equation 4.3). Figure 6.7 illustrates the plots of yield values for each layer and the total yield. Increasing values of *p* result in increasing yield since the likelihood of larger defects is lower. Also note that the metal yields are nearly equal to each other and are consistently lower than the polysilicon yield. This is because sensitivity of the polysilicon layer to shorts is lower than that of metal layers. The metal layer defect sensitivities are comparable to each other. The experiment was also conducted with various values of means and variances of defect number distributions

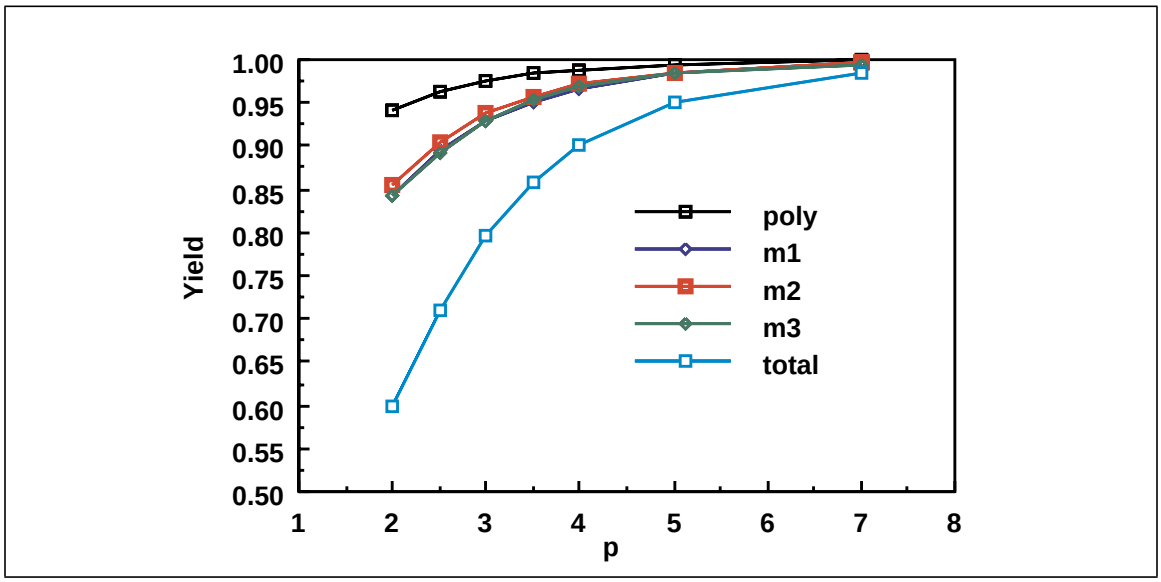

Figure 6.7 Layer and total yield vs. defect size distribution parameter, *p*.

Critical area functions for 0.5 and 0.4 micron design rules are also derived by applying scaling (shrink) transformations to the critical area assumed for the previous case. The resulting chip areas are 2.11 cm^2 and 1.4 cm^2, and the number of usable dies per wafer are 73 and 110, respectively. The same simulation experiments were repeated for these two cases of shrink. In Figure 6.8 the total yields for each of the three cases

are plotted against size distribution parameter, *p*. Note that as expected the yield values for the three cases coincide when *p* = 3 - an artifact of the yield models discussed in Chapter 2 and also presented in Chapter 4. At *p* = 3, increased sensitivity to defects due to shrinking is compensated by the increased number of dies on a wafer. Also note that the design with highest shrink (0.4 micron design rule) shows lesser sensitivity to change in *p* - again due to increased number of dies per wafer.

The cost curves for the three cases illustrated in Figure 6.8 are shown in Figure 6.9. Applying shrink to designs means that the number of dies per wafer increases which

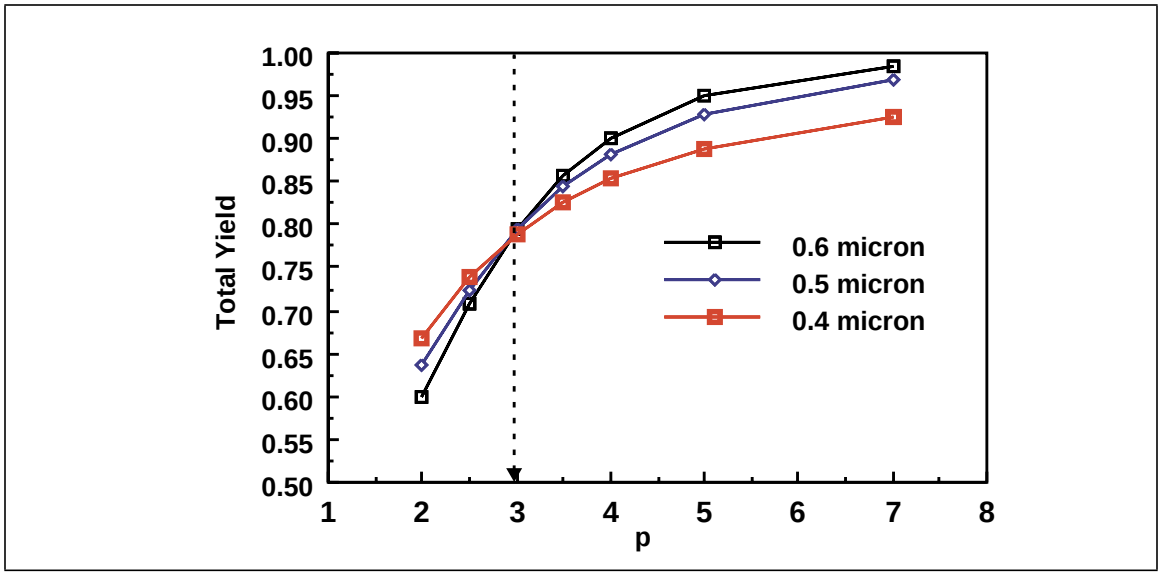

Figure 6.8 Yield vs. p comparison for three versions of a design.

lowers the cost of a good die. However, shrinking designs may lead to parametric yield loss and even an increase in particle rates. Higher particle rates can be because of *proximity effects* where closely spaced IC features are susceptible sites for formation of defects (due to limitation of lithography resolution). Such effects may affect the optimistic results shown in Figure 6.9. At higher defect density (1.2 defects/cm^2) the sensitivity of yield and cost of die to the size distribution parameter, p, increases

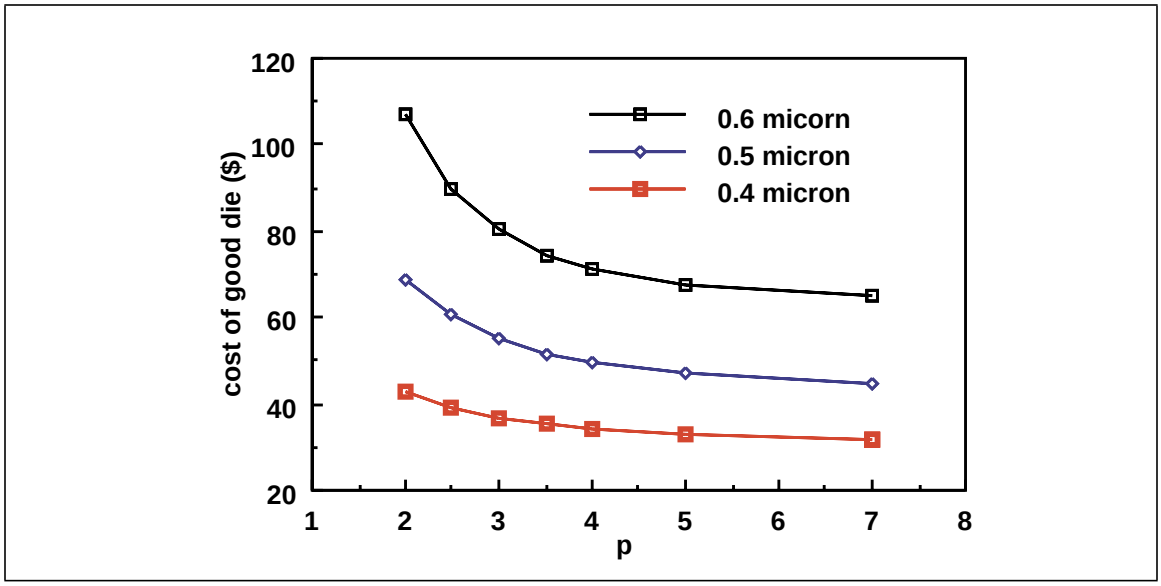

Figure 6.9 Cost vs. p comparison for three versions of design.

substantially. Note the increase in range of yield variation as shown in Figure 6.10 and the steepness of the cost of good die trends as *p* changes. However, these trends still suggest that shrinking the design reduces cost.

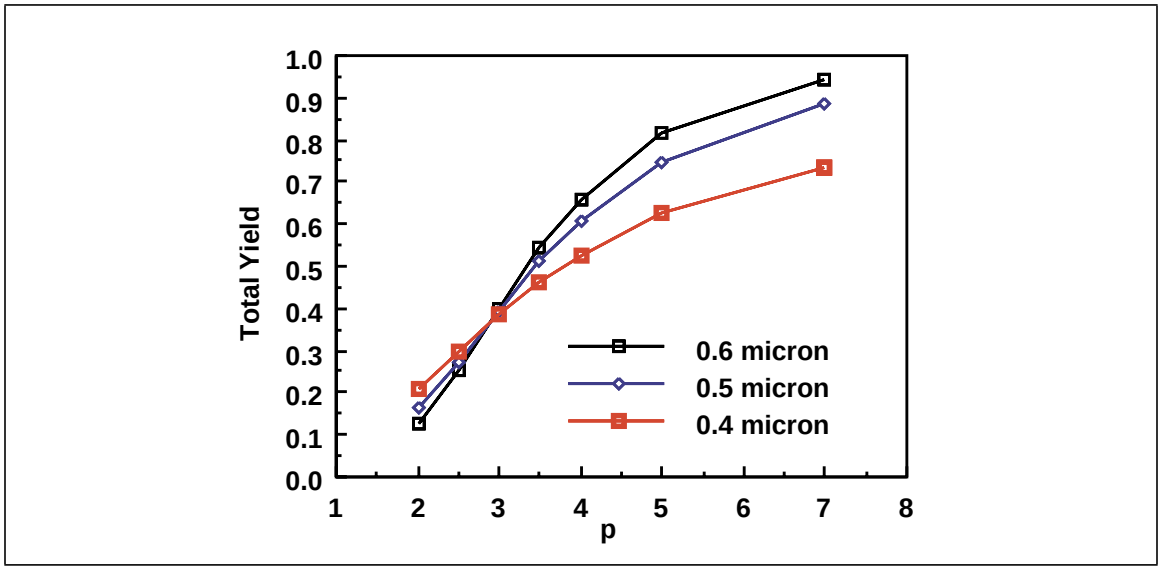

Figure 6.10 Yield vs. *p* for higher defect density (1.2 defects/cm^2).

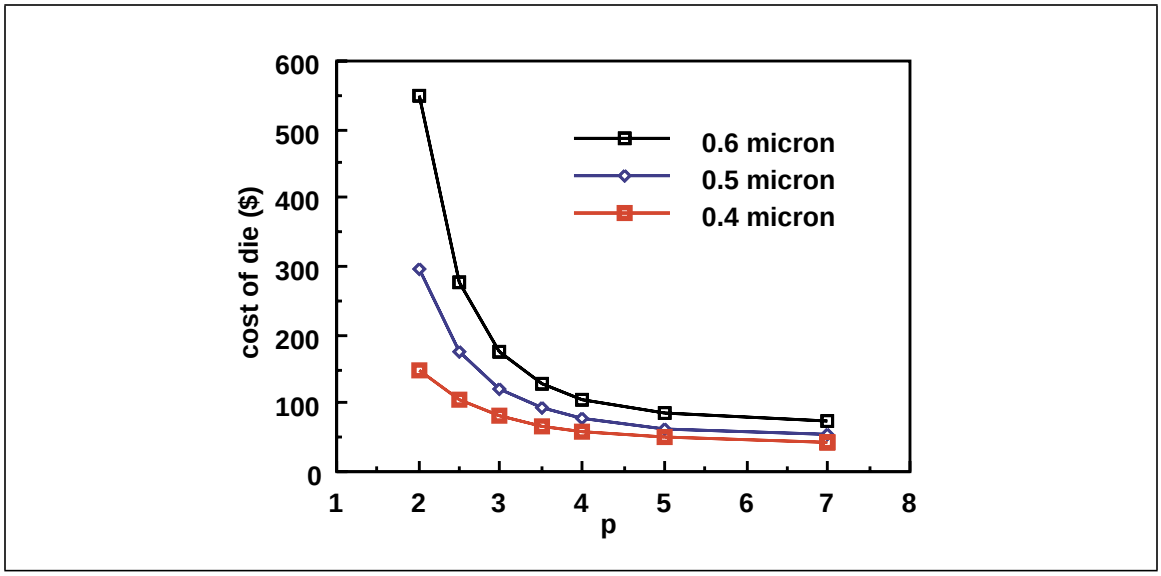

Figure 6.11 Cost of die vs. *p* for higher defect density (1.2 defects/cm^2).

6.4 Imperfect Test Simulation

In this section, results of simulation for the testing process will be presented from the point of view of less than 100% fault coverage. To achieve this, Y4 is set up for simulating defect related yield loss for the 0.4 micron 3-metal CMOS design presented earlier. In this case, the value of the size distribution parameter, p, is fixed at 3.0. A spectrum of defect densities for each defect type (extra material defect of polysilicon, metal1, metal2 and metal3) is considered.

In estimating the time to test a lot it is assumed that all load and unload times in Equations 4.10 and 4.11 are zero. The time to test a fault-free die is assumed to be 12 seconds. Time to test a defective die is assumed to be half of the nominal testing time i.e. 6 seconds. To be precise, however, a distribution of time to test a faulty die must be obtained to correctly simulate the tester properties. Note that, according to the assumptions made, time to test an entire lot will be about 8.8 hours for 100% yield and about 6.6 hours for 50% yield.

Figure 6.12 shows the apparent yield after test (sort yield) as a function of density of defects for a layer. It illustrates the dependence of sort yield on defect density for

90% and 100% fault coverage values. The sort yield at 100% fault coverage value is the actual manufacturing yield and is thus lower. For fault coverage values between 90%

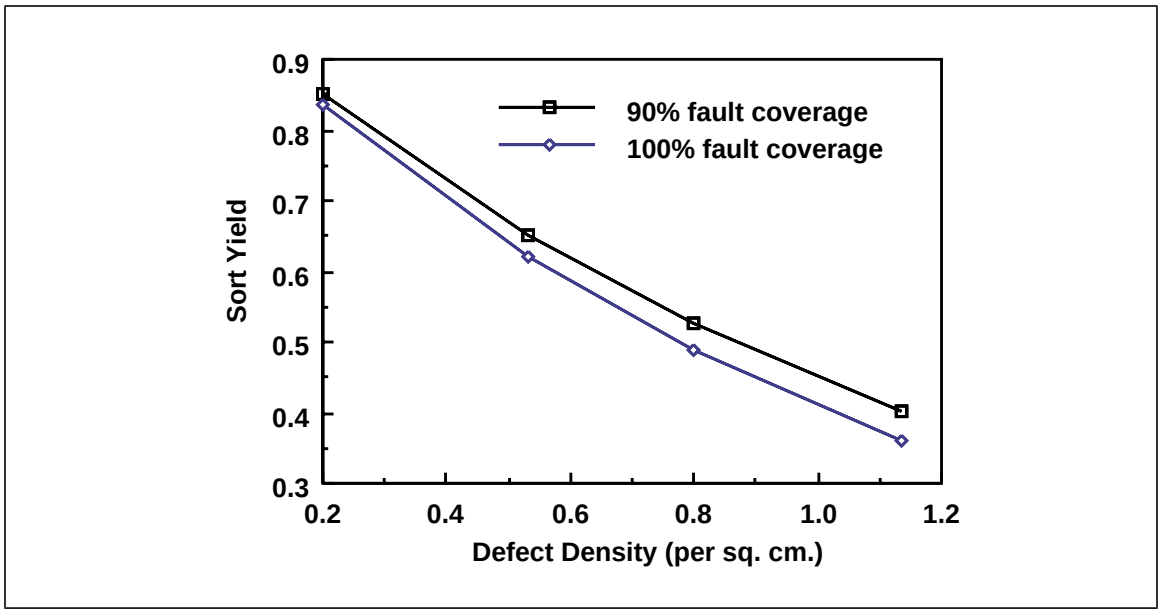

Figure 6.12 Sort yield vs. layer defect density for various fault coverage values.

and 100% the corresponding yield values are also between the curves shown. At higher defect densities and lower fault coverages the difference in sort and real yield is higher. The difference in yield is also a measure for the escape rate i.e., the fraction of tested fault-free dies which are defective. This is illustrated in Figure 6.13 and shows the higher sensitivity of escape rate with increase in defect density and decrease in fault coverage.

The average cycle time of the tester is also a function of defect levels and fault coverage. This is illustrated in Figure 6.14 showing that cycle time decreases with increasing defect density values. Note that for a given defect level, tester cycle time decreases with increasing fault coverage. In summary, tester utilization is higher when both yield and fault coverage are high.

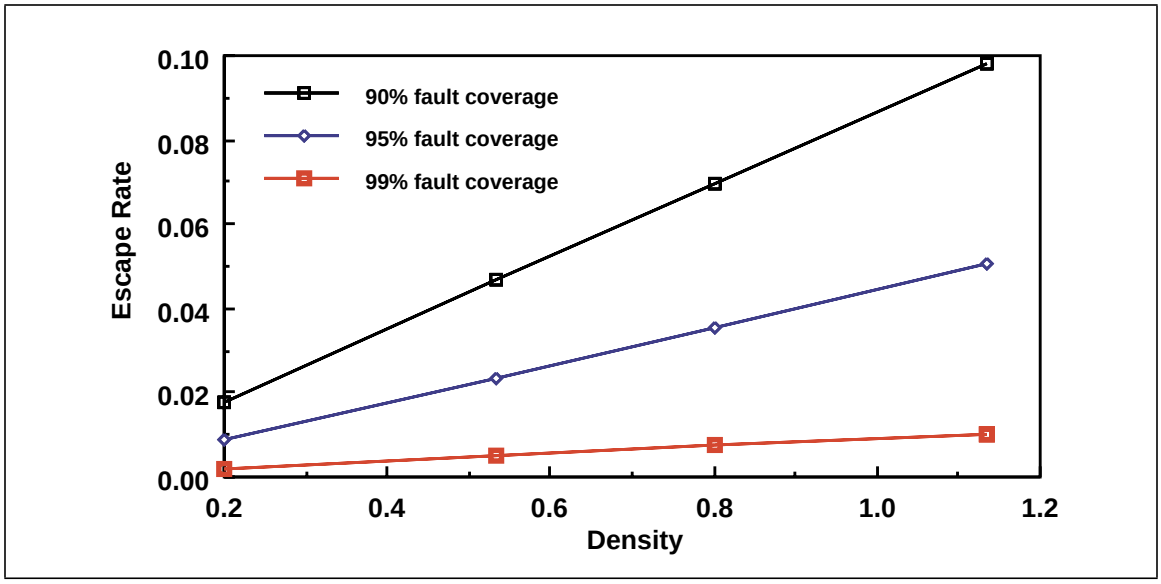

Figure 6.13 Escape rate as a function of defect density and fault coverage.

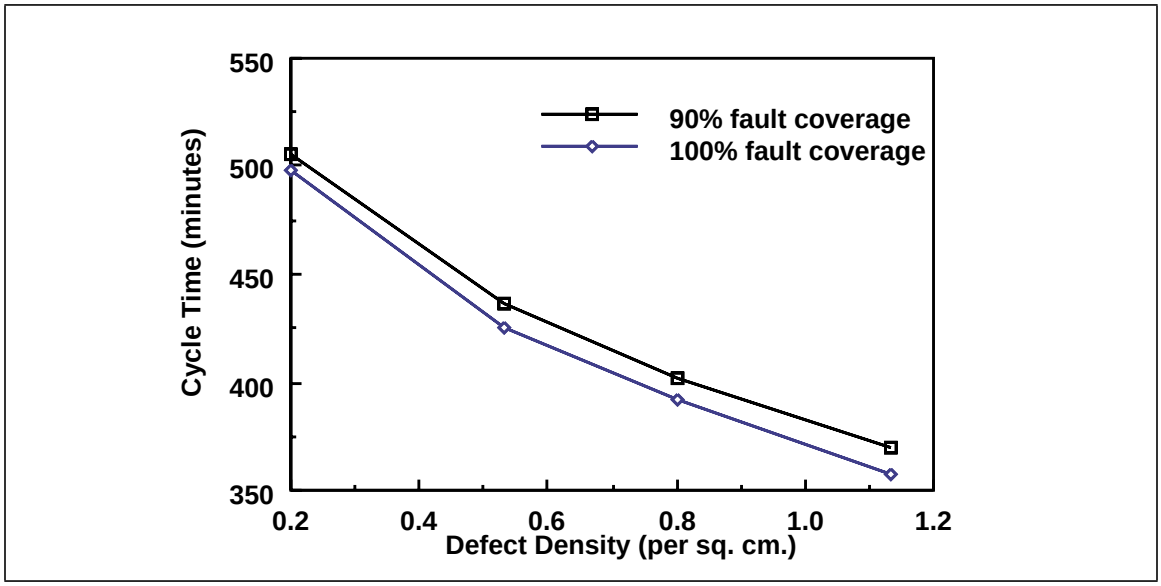

Figure 6.14 Tester cycle time vs. defect density and fault coverage values.

6.5 Simulation of Particle Monitoring

In this section, results of simulating the effect of wafer rejection with the aid of particle monitors is presented. It is assumed that particle monitors can count particles on the wafer surface deposited in a previous step. For each lot sampled for analysis, 4

wafers are randomly chosen and scanned for 30 minutes each for particles with sizes larger than a minimum of 0.5 micron. The average number of particles per wafer in a sampled lot is estimated and when this number exceeds a given threshold, the lot is rejected. Simultaneously a new lot is added at the input of the fabrication phase to compensate for the rejected lot.

It is expected that if the threshold for rejection is set correctly then very low yielding lots can taken out of the in-process inventory which, in turn, would increase the yield. This can be useful in a scenario where the yield is usually high but occasionally some low yielding lots are produced, i.e., the yield variance is large. In order to simulate this, Y4 is set up to randomly vary the lot-to-lot mean and variance of the number of particles. The resulting yield distribution is shown in Figure 6.15 for the 0.4 micron, 3-metal CMOS product with a mean yield of 0.6964. The die cost without any particle monitors is $39.57.

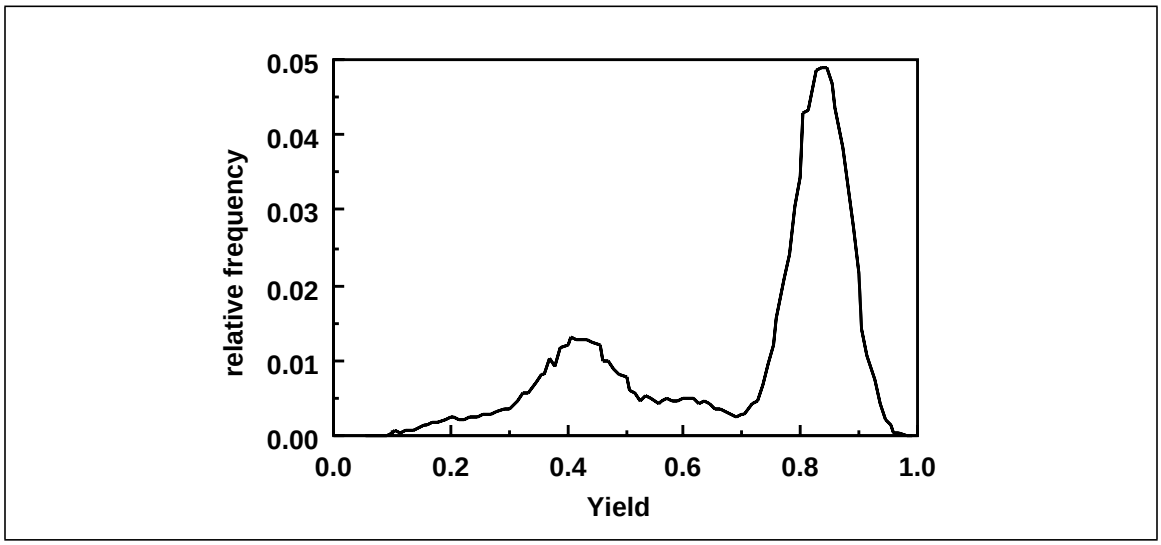

Figure 6.15 Yield distribution for particle monitor simulation.

Two parameters of interest in this simulation are the threshold number of particles per wafer for it to be rejected and the number of available particle monitors. The maximum sampling rate for lots is limited by the number of particle monitoring equip-

ment. In this case, with one monitor one in 20 lots is sampled; with 5 monitors all the lots can be sampled for analysis. The expected yield and cost of good die as a function of the number of monitors and the threshold number of particles is shown in Figure 6.16. As shown, the yield has a tendency to increase for certain threshold values but die cost also increases and in some cases it is almost the same as the nominal scenario without any monitors. Further analysis showed that any increase in yield is effectively counteracted by:

1. an increase in operating costs due to particle monitors,
2. decrease in productivity due to decrease in throughput because of lot rejections and,
3. cost of partially fabricating those wafers that are rejected.

The simulation experiments were repeated with rejection of specific low yielding wafers instead of lots and the results were similar.

It is speculated that the imbalance in the factory load caused by wafer or lot rejection in an intermediate step has adverse effects on the performance of the line. It is possible that using a better scheduling mechanism for wafers could improve the yield and cost performance. For example, several lots with less than a full compliment of wafers can be merged, at an appropriate step, to a smaller number of lots. This can improve throughput rate and may therefore decrease die cost. However, such complex rules are not implemented in Y4. It is also likely that increasing variance in yield may reveal a different outcome and this is left as a possibility for future explorations.

6.6 Yield vs. Time Simulation with Defect Diagnosis

In this section, yield learning curve simulation results are presented for the CMOS design with a minimum feature size of 0.4 microns. It is assumed that the initial exponent, p, of the particle size distribution is 2.0. The initial mean and variance of the particle number distribution is set in a such a way (about 3 defect/cm^2) that the total

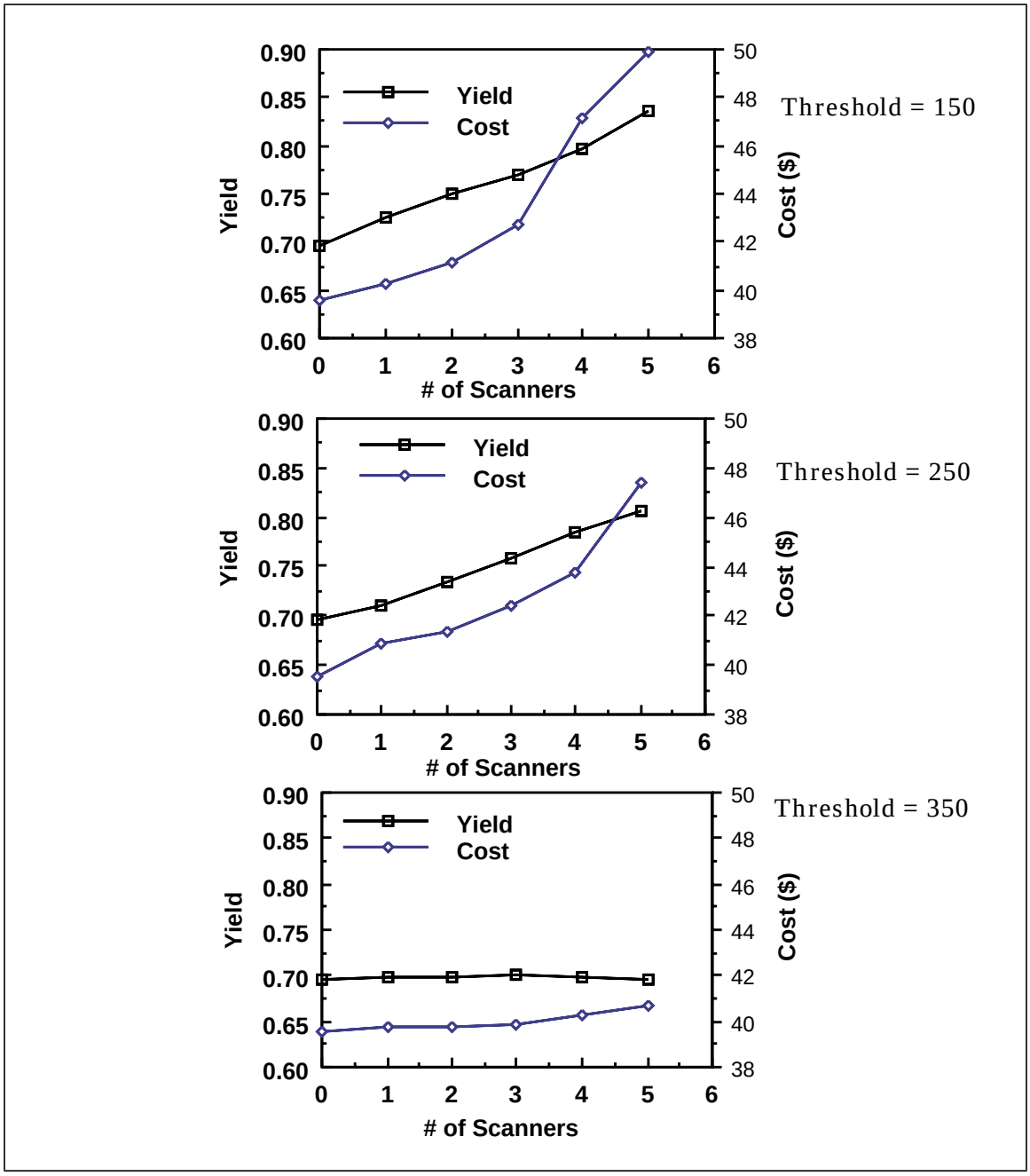

Figure 6.16 Yield and die cost as a function of number of monitors.

initial yield is less than 10%. Note that the initial yield also depends on the critical areas assumed for each of the defect types.

Wafers are sampled for failure analysis when there are more than 30 (D_{min}) defective die on a wafer and when there are less than 3 wafers (Q_{limit}) waiting to be analyzed. The failure analysis consists of five steps: observation under microscope, observation with SEM, stripping layers (if required), cross section analysis, and, spectroscopic analysis (WDX, EDX, etc.). These steps are carried out in sequence and the time required at each step is calculated using Equation 4.13. The limiting value of final diagnosability, m_f, for each step is assumed to be 0.99. The parameter, e_d, for each piece of equipment is chosen such that it reflects the expected time required for each of these steps. The maximum allowable time, T_{cmax}, is fixed at 1, 3, 1, 3, and 4 hours respectively. This means that the maximum time required to analyze 30 defects in the top metal layer will be about 2 weeks (not considering the queuing time).

To calculate the initial diagnosability value, (Equation 4.14), the layer number, n, and defect size, R, are extracted from defective die data. One also needs to also consider the extent of search area, A_s, for the product under consideration. This parameter is assumed to be defined by a normal distribution with a mean of 0.2 cm^2 (variance = 0.008). This assumption is necessary since a fault is defined to be a short in a given layer. In reality, the types of electrical faults (shorts in our case) tested at probe testers depends on the fault models used to generate the test vectors. Thus, a single layer short can mean many electrical faults each having its own range of area to be searched for the corresponding defect. The values of the parameters given here were arrived at by performing an array of simulations with reasonably acceptable values for average time for analyzing a single wafer.

Assignment of the equipment responsible is accomplished by incrementing the variable $E_{suspect}$ by 1 for the piece of equipment responsible for the defect. For the rest of the equivalent equipment the increment value is 0.5 (mimicking uncertainty in correctness of diagnosis). Corrective actions on a piece of equipment is deemed necessary when this count exceeds 20 (E_{thresh}). The equipment is taken off-line for cleaning as

soon as it has finished processing the current lot of wafers. The value of p_{diff} (Equation 4.16) is set to be 0.2 for each type of particle, and k_m and k_σ (Equation 4.15) are set to be 0.95 each.

Figure 6.17 shows an example of the trend plot of total die yield for each lot. The yield starts increasing only after about 15 weeks of simulations. This is because failure analysis is not conducted for the first 10 weeks in order to let the simulated fabrication line settle into an equilibrium. The total period of simulation is 75 weeks and the yield values shown in the figure are for nearly 7500 lots.

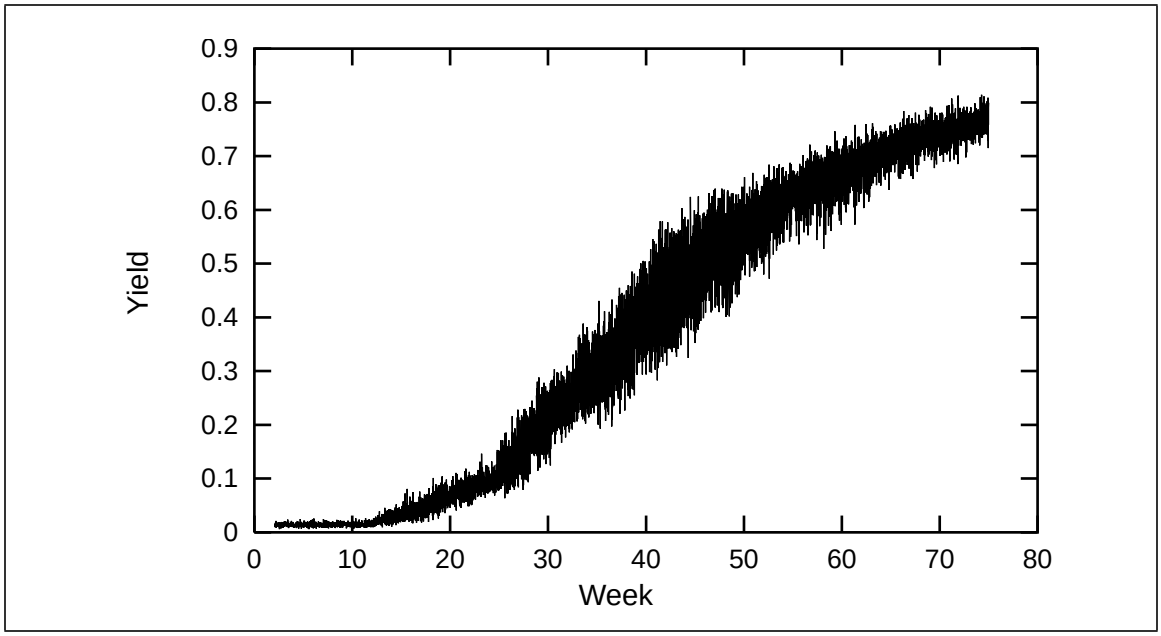

Figure 6.17 Example yield learning curve.

Observe that in Figure 6.17, the variance in yield increases as the yield ramps up and then decreases again as the mean yield increases. When the probability of a die failing is 0.50, the variance is also at its highest (same as the variance of a binomial distribution with p = 0.50). This is also due to the fact that during the yield ramp-up period, some lots are processed in relatively cleaner equipment than the others. Even-

tually, as the yield saturates, the particle characteristics of all the equipment in a workstation become more comparable.

The weekly average of the yield trend plot is shown along with the yield of polysilicon and the metal 3 layers is shown in Figure 6.18. Observe that the yield of the metal 3 layer starts to increase almost right after the failure analysis is initiated (after 10th week). Polysilicon layer yield, on the other hand, starts to increase only after another 15 weeks (around 25th week). This illustrates the fact that polysilicon defects are more difficult to detect than defects which are near the surface of the chip like metal 3 defects. Further, the yield of metal 3 is low providing enough samples to keep the failure analysis resources busy analyzing metal defects. Polysilicon defects are effectively ignored until the metal 3 yield reaches about 0.65. However the rate of yield learning for the polysilicon layer is higher than metal 3 since the availability of more samples with polysilicon defects compensates for decreased diagnosability of these defects.

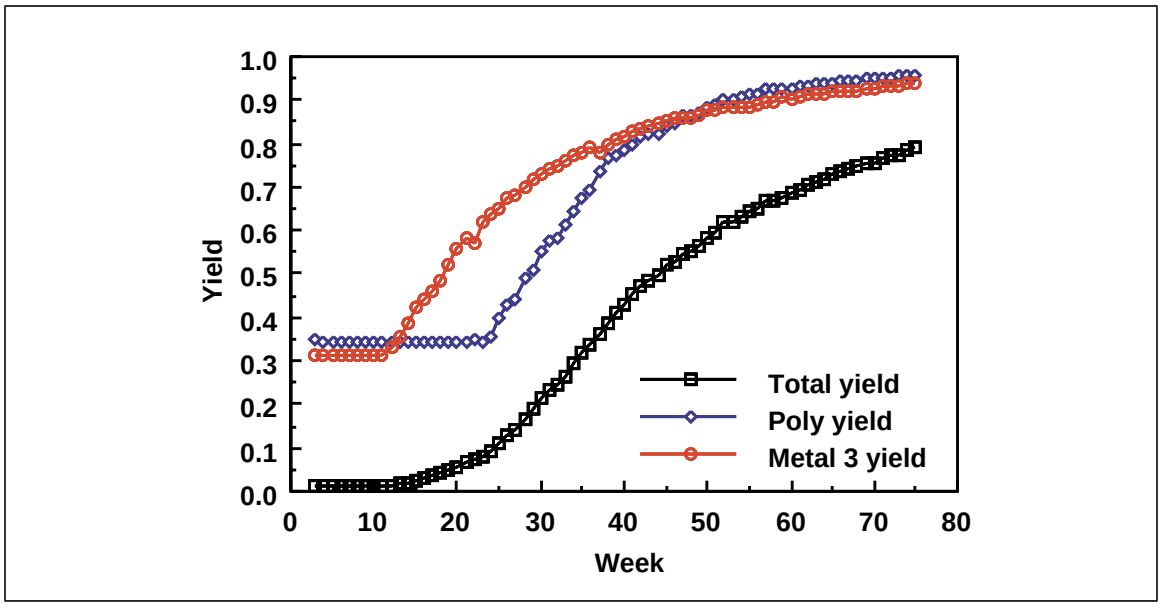

Figure 6.18 Yield vs. time trends of each defect type.

Cost estimations were also performed with the yield learning simulations. A total of 7.62 million good die are produced in the 75 weeks of simulation period at a cost of $72.52 per die. This die cost estimate includes costs arising out of failure analysis which accounts for 5.47% of the good die cost. This amounts to about 30 million dollars worth of failure analysis cost compared to about 522 million dollars for the rest of the factory over the simulation period. In the next chapter, further applications of Y4 in cost analysis will be presented.

6.7 Yield vs. Time Curve With Particle Monitoring

This section presents the results of yield learning simulations using particle monitors alone for local and short feedback for particle rate correction. The initial particle rate and cleaning function parameters are the same as in the previous section. Lots are sampled after each of the four steps where polysilicon, metal1, metal2 and metal3 defects are introduced. The sampling strategy is the same as in the example of particle monitors with wafer rejection (Section 6.5). Again, only one out of 20 lots can be sampled using one monitor and all lots can be sampled using five monitors. The threshold number of particles to initiate a corrective action is set at 30 particles per wafer pass.

It is expected that with increased rate of sampling, the feedback cycles will become shorter and thus yield learning rate should be higher. Figure 6.19 shows yield learning curves obtained with simulation as a function of increasing number of particle monitors. As expected the learning curves are increasingly steeper as the number of particle monitors is increased. Unlike the previous example of particle monitor simulation, here both yield and productivity (number of good die produced) are significantly increased thus increasing cost benefits. The cost and number of good die produced is illustrated in Figure 6.20. This simulation exaggerates the yield learning benefits of using particle monitors to some extent but illustrates a possible impact nevertheless.

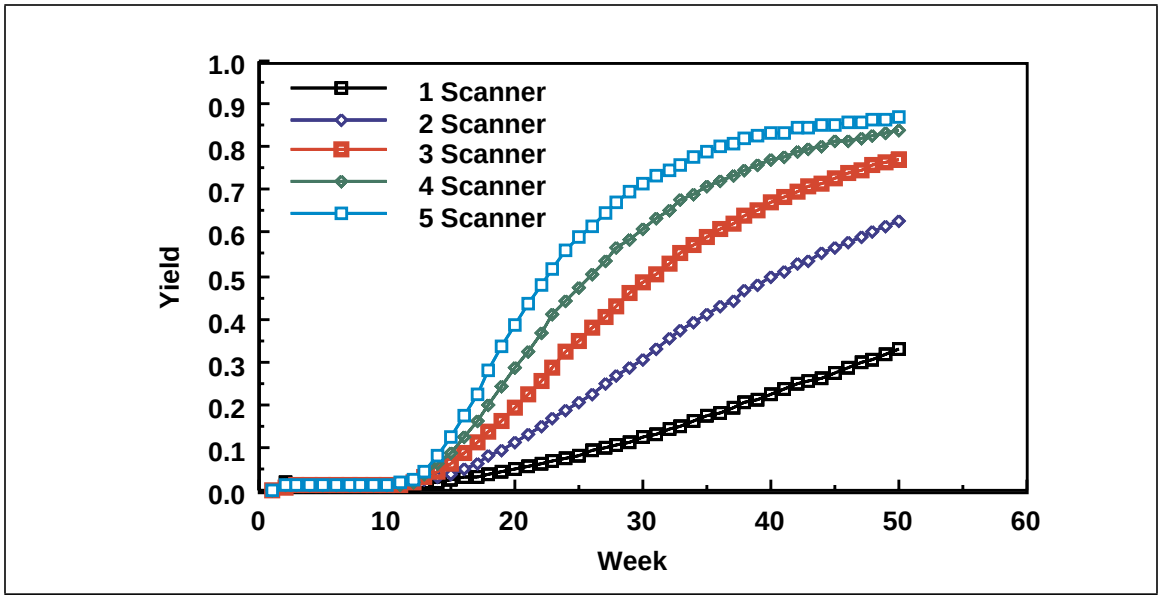

Figure 6.19 Yield vs. time curve simulation using particle monitors.

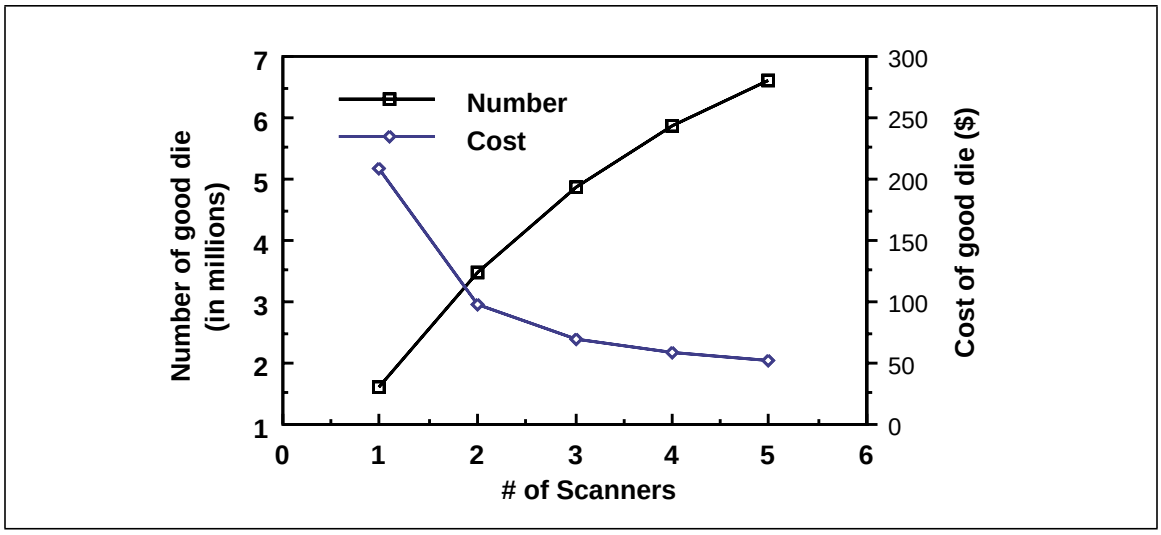

Figure 6.20 Cost and number of good die for particle monitor simulation.

6.8 Performance of Y4

Time and memory requirements of Y4 depend on a number of factors pertaining to the simulation setup. In WSIM the factors which affect performance are the wafer starts per week (WSPW), the number of process steps, number of pieces of equipment and the time period of simulation. Cycle time and cost of wafer simulations presented

in Sections 6.1 and 6.2 take less than 30 minutes and use approximately 7 Mb of memory. Performance of yield simulation such as the ones presented in Section 6.3 depends on the number of particle, defect and fault types that need to be simulated. Higher the rate of particles, the longer the time required for simulation. The time range is from about 1 to 8 hours for a single simulation over a period of 50 weeks. The memory requirement is between 8 and 10 Mb. The 75 week yield learning simulation presented in Section 6.6 takes about 14 hours using 11 Mb of memory. Simulations with PSIM and TSIM do not change the time and memory requirement significantly. These time and memory requirements are for a DEC station 5000 running Ultrix v4.3.

References

[1] W. Maly et. al., "Memory Chip for 24-Port Global Register File", *Proc. of IEEE Custom Integrated Circuits Conference*, San Diego, pp. 15.5.1-15.5.4, May 1991.

[2] M. Patyra and W. Maly, "Circuit Design for a Large Area High-Performance Crossbar Switch", *Proc. of IEEE Int. Workshop on Defect and Fault Tolerance in VLSI Systems*, pp. 32-43, Nov, 1991.

[3] P. K. Nag and W. Maly, "Hierarchical Extraction of Critical Area for Shorts in Very Large ICs", *Proc. of Int. Workshop on Defect and fault Tolerance in VLSI Systems (DFT)*, pp. 19-27, Nov. 1995.

Chapter 7

Applications of Y4

In Chapter 6, results of simulations using Y4 were presented to illustrate some of the basic capabilities. The examples used also can be viewed as possible applications of Y4. In this chapter, a spectrum of simulation results are presented which illustrate the potential applications of the software Y4. The examples presented here are primarily geared towards illustrating time domain response of a manufacturing line. The first aspect investigated is time domain changes in input wafer start rate and its impact on cycle time and cost of wafer. The rest of the four examples in this chapter are related to studying the impact of various manufacturing and strategic attributes on yield learning curves. These results are categorized as follows:

1. Effect of capacity of failure analysis facility on the yield learning rate.
2. Reaction of a manufacturing line to sudden increase in particle rates and consequent yield degradation.
3. Using a “diagnosable” product to aid in yield learning for an relatively undiagnosable product.
4. Delaying introducing a relatively undiagnosable product in a line partially “debugged” using a diagnosable product.

7.1 Cost of "Ad Hoc" Wafer Release Policies

In this section, the effect of non-uniform wafer release policies on the operational performance of a fabrication line will be considered. Under ideal conditions, the wafer start rate per week should be held constant at a certain value within the capacity of the fabrication line. In reality, the wafer start rate may need to be increased for a few

weeks to meet a special demand. Figure 7.1 shows such a surge of wafer starts per week (WSPW) as a function of time. A wafer surge is characterized by the number of additional wafers per week (height of the surge) and the duration of the surge. Of course, for any given number of additional wafers to be produced, there are a number of choices of surge length and duration.

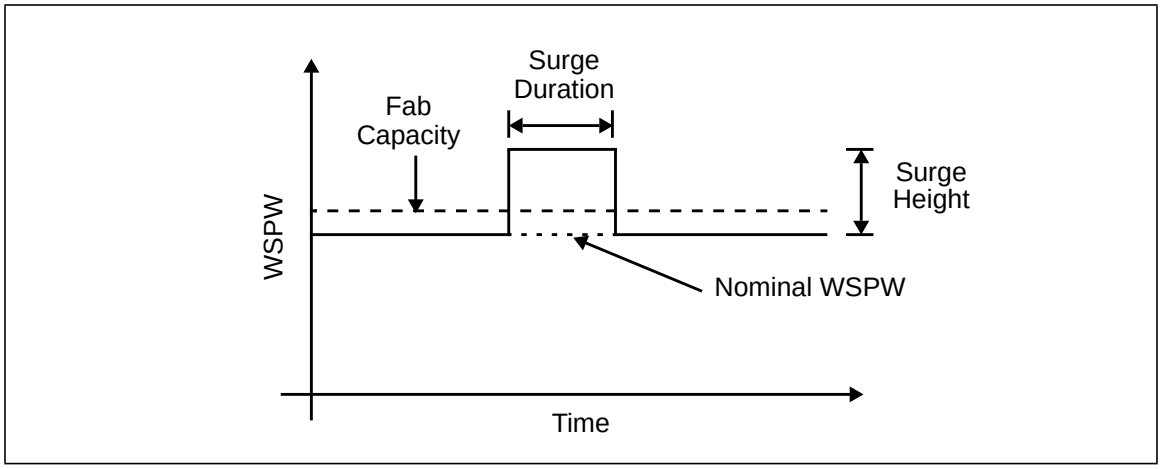

Figure 7.1 Graphical representation of a wafer surge.

An important impact of such wafer release policies is the estimate of possible additional revenue it might generate. Accurate estimation of such revenue requires a complete understanding of the complex interaction between market demands, pricing policy, and manufacturing efficiency and cost. Leaving this and other aspects aside, here the focus is on the change in manufacturing cost of wafer as a result of the change in operating conditions of the fabrication line.

When a fabrication line is operated under its rated capacity, any increase in wafer start rate leads to a decrease in cost of wafer, as amply illustrated in Figure 6.5. However, the more typical situation is that the fabrication line is operated near full capacity to achieve the lowest overall cost of wafer. In this situation, the consequences of a small increase in wafer starts may be dramatic. It has been shown, for instance in [1], that cycle times increase rapidly, and it takes several weeks for the cycle times to

return to equilibrium values. At the same time, the inventory size or WIP (Work in Progress) builds up but the throughput rate does not change much.

Figure 7.2 shows weekly averages of cycle times for an wafer surge height of 192 wafers per week. No special priority is given to the excess wafers. The nominal curve

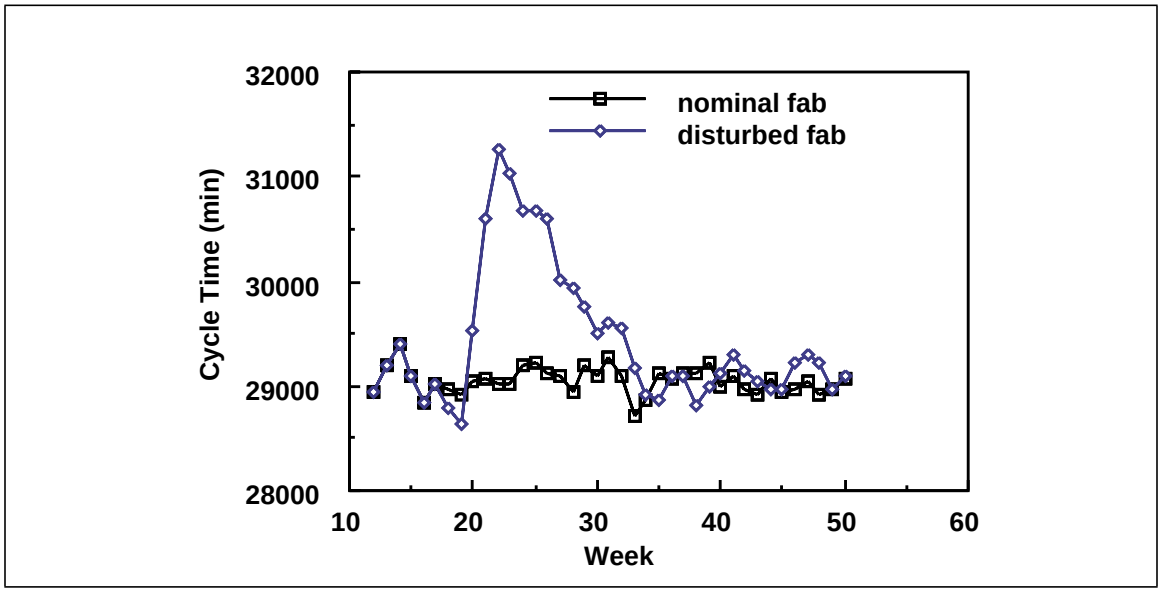

Figure 7.2 Weekly averages of cycle time for nominal factory and for a wafer surge.

of cycle time where no such surge is applied is also shown in the figure. As presented in [1], cycle time increases and returns to normal levels after a duration which is much larger than the input wafer surge duration. Cost estimations were also performed for both the cases with the assumption that the penalty cost K_{wait} (as in Equation 4.21) is \$0.01 per minute per wafer. Cost of wafer follows a similar trend and Figure 7.3 shows the difference in weekly wafer cost. Note that even the peak difference in wafer cost is a small fraction of the average wafer cost for the nominal case which is \$2850 (amounting to approximately \$0.16 per wafer per minute of actual processing time).

Both the duration of the cost surge and the maximum wafer cost difference depend on the input WSPW surge duration and height. Figure 7.4 shows the nature of the

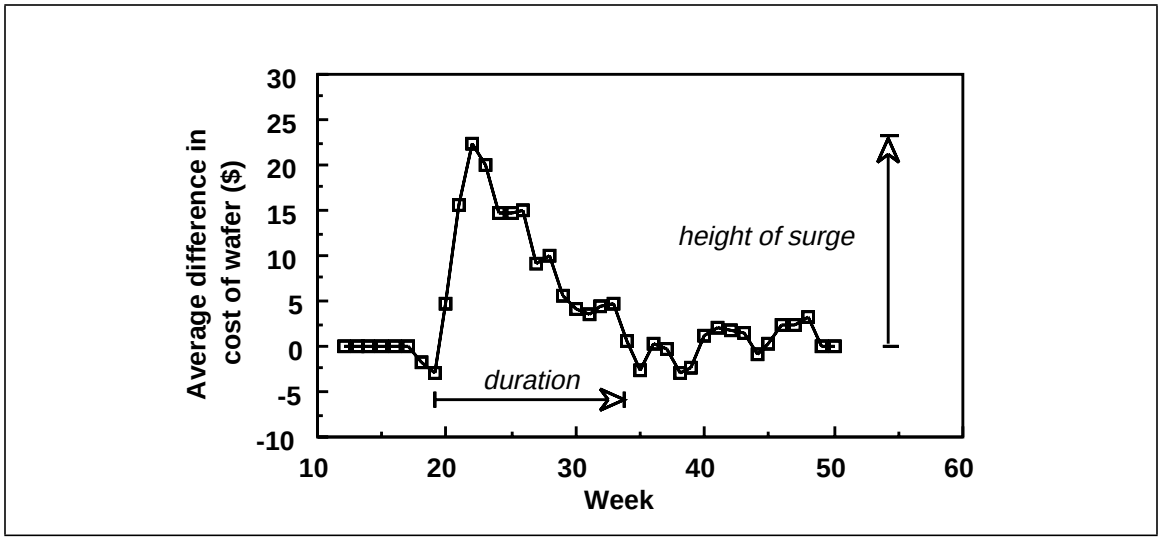

Figure 7.3 Difference in weekly wafer cost for nominal factory and for a wafer surge.

dependence of output surge duration for a range of input surge duration. The simulations were performed for three lengths (1, 2 and 3 weeks) of input surge duration. Linear regression was also performed to clearly illustrate the dependence (square of the regression coefficient are given with the legends). Notice that the duration is a stronger function of duration of the input surge. This suggests that for a given excess wafer requirement it is better to release it over a shorter duration of time.

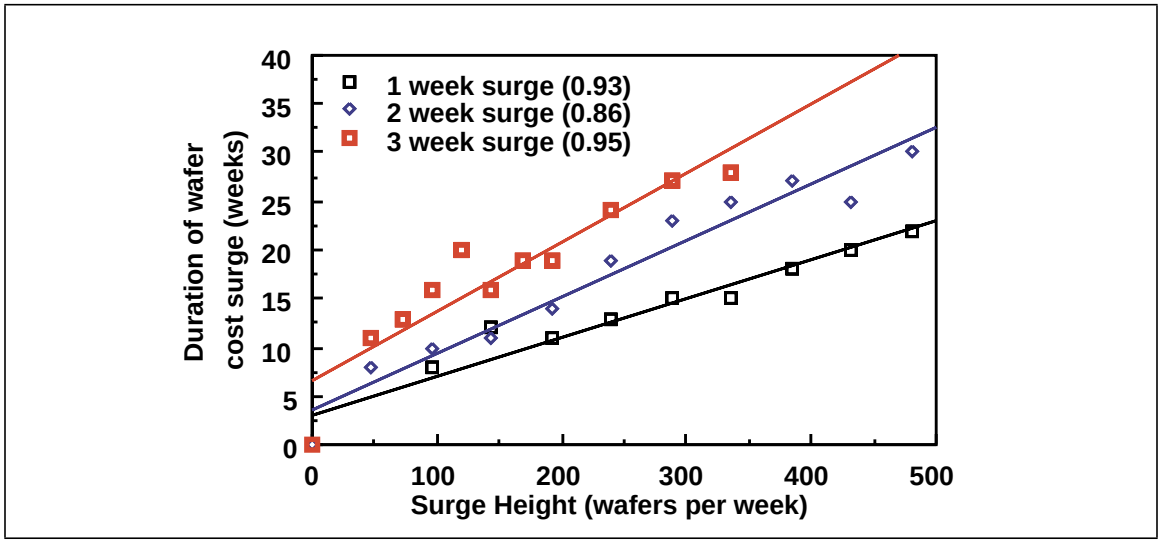

Figure 7.4 Duration of cost surge length vs. input surge height.

If one focuses on the difference in total cost of manufacturing, C_{total} (Equation 4.23), instead, then the trend looks different. Figure 7.5 shows the difference in manufacturing cost as a function of the height of the surge. Note that the regression curves shown in the figure show a quadratic growth and can easily be in the order of millions of dollars. This experiment illustrates the risk of applying a wafer surge to a fabrication line operating near capacity. One can justify such a risk by arguing that the extra wafers produced may enable a premium price or meet a key customer's demand. Such arguments can be placed in their proper perspective using simulation tools.

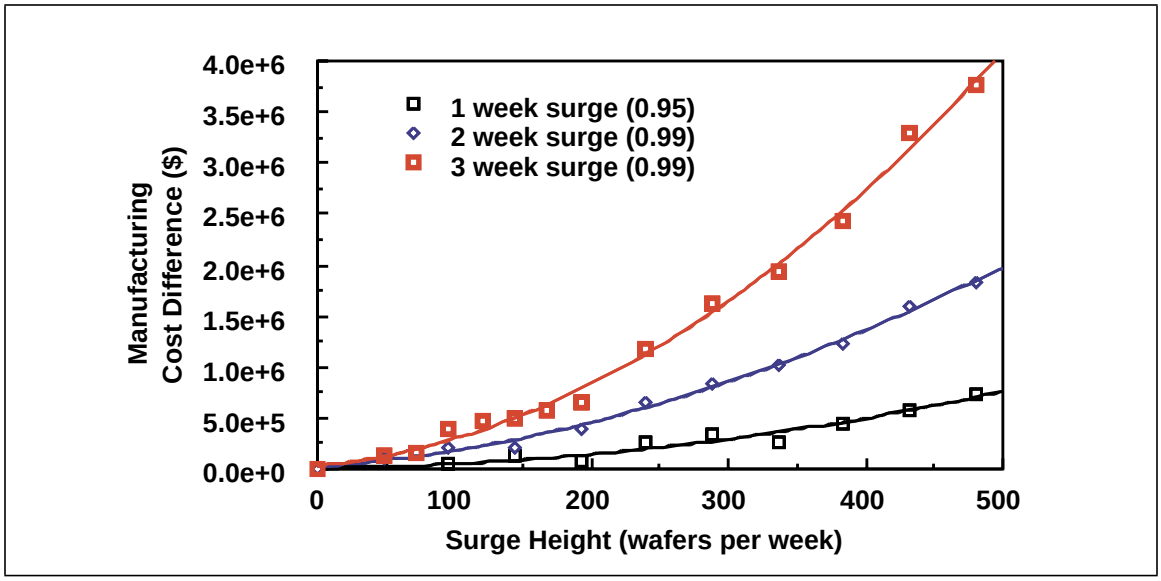

Figure 7.5 Difference in manufacturing cost vs. input surge length.

In the previous examples, we had assumed that partial loading of batch equipment is allowed. If one allows only full loading of the batch equipment, then the difference in cost of manufacturing is substantially reduced. This is not only due to better batch equipment utilization but also due to reduced sensitivity to cycle time build up. When the extra wafers during the surge are given hot lot priority (highest priority) then, as expected, the difference in manufacturing cost increases dramatically. Capacity of the fabrication line also plays an important role in this type of analysis. Naturally, a larger

fabrication line is better able to absorb a small surge in wafer starts and, consequently, the cost impact is less.

In summary, it has been demonstrated above that both cycle times and increase in wafer cost are strong functions of input wafer surge height and duration. Cost comparisons for the cases presented indicate that pricing policy of the additional wafers must be carefully evaluated. Scheduling rules play an important role in deciding the cost impact. Hot lots inherently introduce higher manufacturing cost and should be avoided unless short cycle times of additional wafers are of paramount importance. Finally, in smaller fabrication lines such disturbances can have adverse effects.

7.2 Effect of Failure Analysis Capacity on Yield Learning

In Section 6.6, simulation results were presented for a CMOS product and the yield learning curves obtained for polysilicon and metal3 defects are shown again in Figure 7.6. In this section, simulation results obtained by doubling the failure analysis capac-

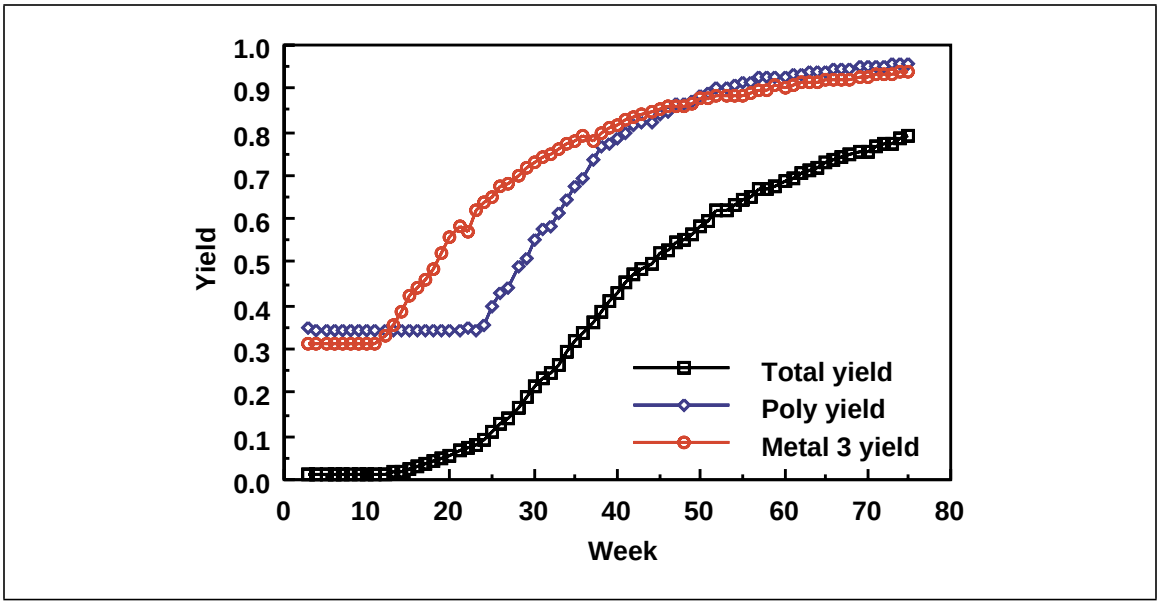

Figure 7.6 Yield learning curves for CMOS product.

ity will be presented and discussed.

Besides the obvious fact that the yield learning rate should increase, two other effects are also apparent as shown in Figure 7.7. First, the polysilicon layer yield starts to increase around the 20th week, which is about 5 weeks sooner than the previous case. Second, this occurs when the metal 3 yield is 0.73 instead of 0.65 for the nominal case shown in Figure 7.6. Availability of more resources enables metal defects to be diagnosed quickly. More importantly, there is enough left over capacity to analyze polysilicon defects while the metal defects are being analyzed.

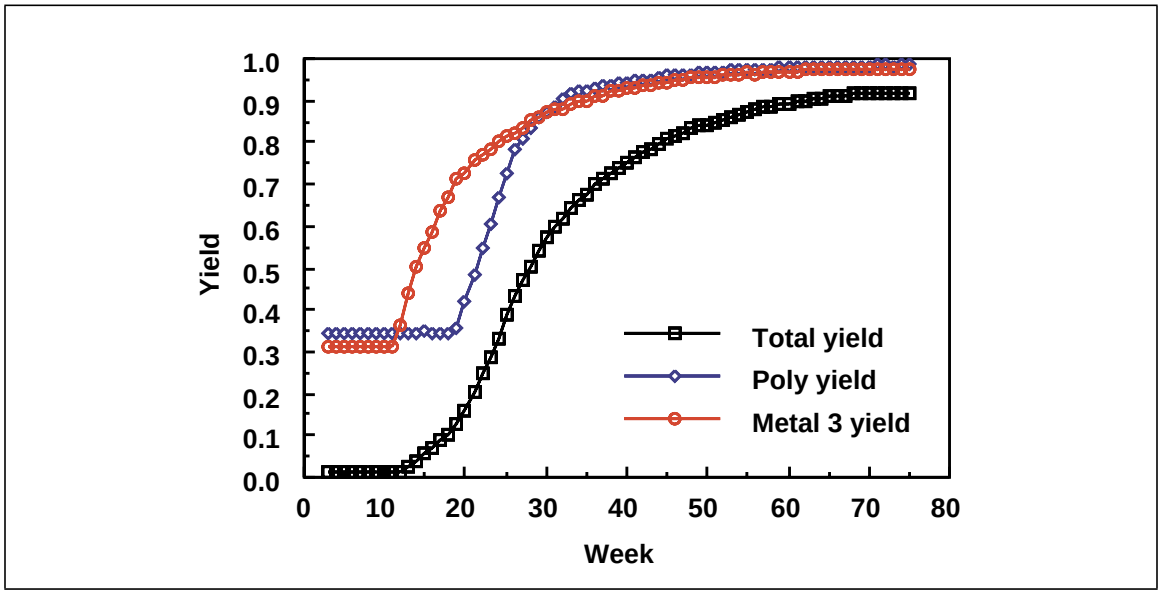

Figure 7.7 Yield learning with twice the failure analysis capacity.

Comparison of the two cases becomes more meaningful if the cost and number of good die produced are compared. In the first case, the estimated cost of good die is $72.52 and a total of 7.62 million good die are fabricated. In the second case, it is $51.13 and 11.54 million for the cost and number of good die, respectively. These estimates take into account the increased cost of failure analysis besides the cost of fabrication of wafers. Thus, an extra revenue of $247 million can be generated even when the ICs are sold at the cost price of the first case.

In this particular case, further increasing the capacity of failure analysis results in an additional smaller gain in the yield learning rate. This is because of the fact that the corrective feedback rate becomes too frequent. The corrective actions cannot be performed as fast because of constraints on scheduling the equipment being taken off-line and cleaned. This result is biased by the fact that only 4 types of defects were considered. However it illustrates that Y4 can be used to determine the required failure analysis capacity for attaining the maximum yield learning rate.

7.3 Effect of Sudden Degradation in Yield on Cost

In the last section, it is implicitly assumed that the particle rates and size distributions falls only when the corresponding equipment is cleaned. However, it is also possible that they may increase and degrade the yield. This could be due to some internal disturbance such as imprecise processing causing more particles to be released. Here, the effect of such a degradation in one of the sputtering tools is considered causing metal yield to degrade. Specifically, at the end of the 30th week, the mean of the particle number distribution for one of the seven sputtering tools is assumed to increase by a factor of five.

Figure 7.8 shows the result of the simulation illustrating the yield trend plots. Observe that, compared to the result shown in Figure 7.6, the net yield learning rate has decreased. The increase in metal defects causes metal yield to drop first. After a certain delay failure analysis catches up with this increase in defective die with metal defects and metal yield starts to increase again. But at the same time, polysilicon yield learning rate drops because failure analysis resources are consumed more in detecting metal defects.

Figure 7.9 illustrates the yield learning rates for the manufacturing line with double the failure analysis capacity as before. As expected, the yield learning rate is higher than in the one shown in Figure 7.7. But there is an important difference in the nature

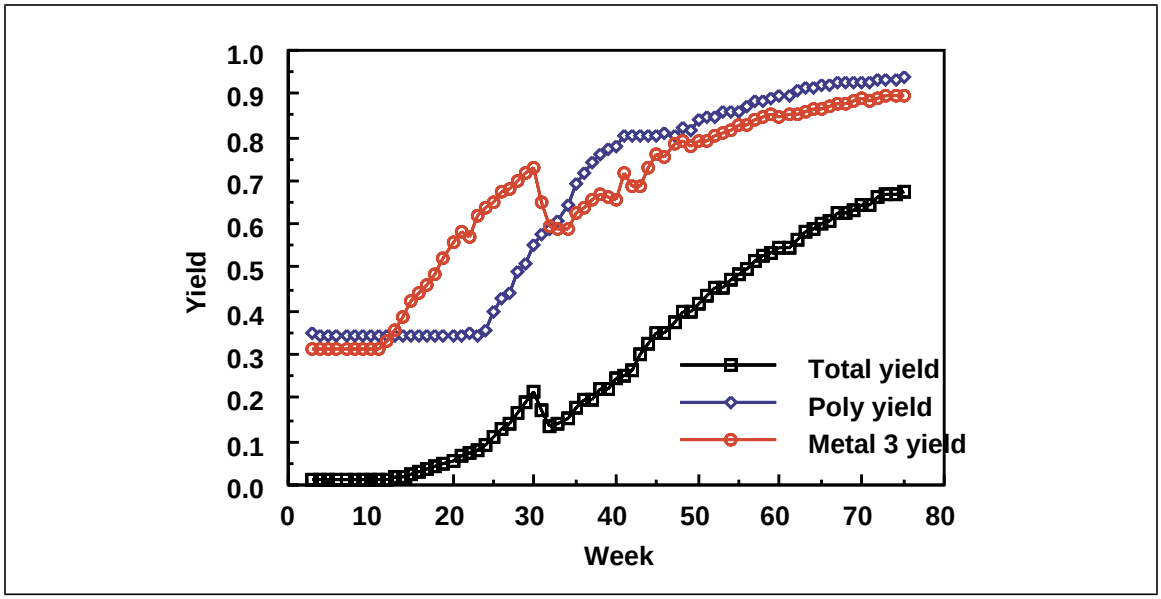

Figure 7.8 Yield learning with sudden increase in defect rates.

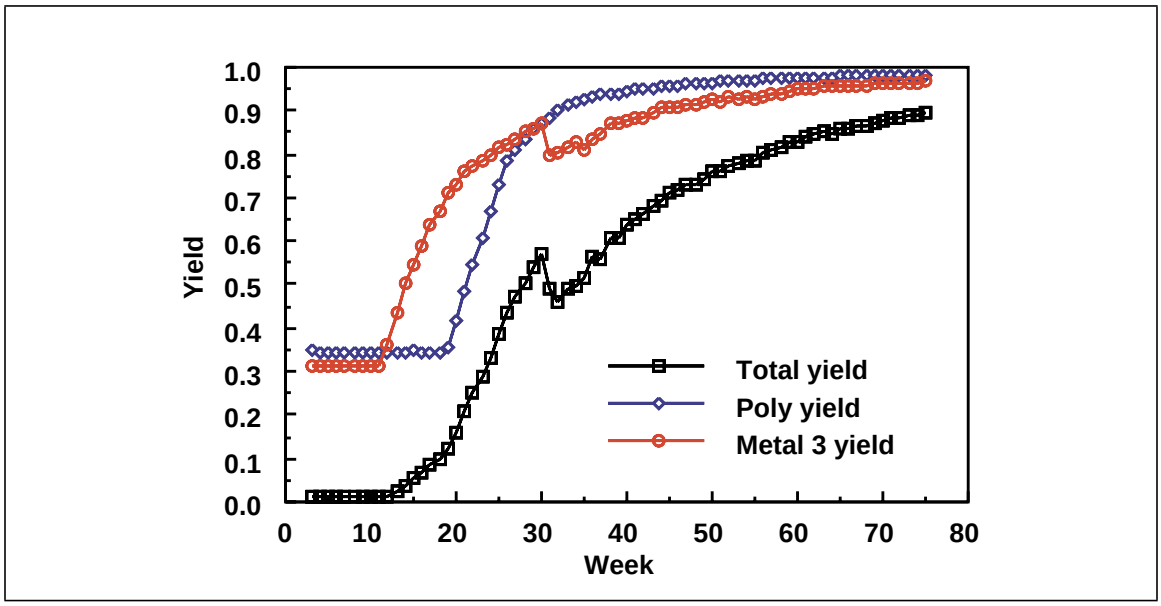

Figure 7.9 Effect of increased failure analysis capacity in the event of yield degradation.

of the yield learning curves. In the second case, the yield learning rate of the polysilicon layer remains essentially unaffected. This result again illustrates how the extra capacity helps by continuing to perform analysis on polysilicon defects in spite of

occurrence of an increased number of defective die with metal defects. Observe that the time the yield problem occurs, the metal yield is high enough so that the number of defective die being sampled for analysis is already low. Thus, the failure analysis facility has little trouble in absorbing the small increase in number of defective die with metal defects.

Again a cost analysis of the two cases helps to put the comparison in proper perspective. With normal capacity of failure analysis, the cost and number of good die produced is $94.92 and 5.81 million, respectively. Doubling the capacity of the failure analysis facility causes cost and number of good die to be $56.90 and 10.49 million, respectively. This means a net gain of $399 million if all the IC's could be sold for the cost price ($94.92) of the first case.

It is useful to compare the two manufacturing lines - one with normal capacity and the other with double the capacity of failure analysis - from the perspective of sensitivity to yield degradation. Table 7.1 summarizes the results for the two manufacturing lines. The first and second rows of the table shows the number and cost of good die

	Normal capacity			Double capacity		
	Undisturbed fab	With yield degradation	% change	Undisturbed fab	With yield degradation	% change
Number of good die (in millions)	7.62	5.81	-23.75	11.54	10.49	-9.1
Cost of die ($)	72.52	94.92	+29.92	51.13	56.90	+11.28
% of cost from failure analysis	5.47	5.32	-2.74	11.5	12.44	+8.17

Table 7.1 Cost comparison.

produced. The fraction of the cost of good die attributable to failure analysis is given in the third row. Notice that obviously the manufacturing line with more failure analysis capacity is much less sensitive to the yield problem than the normal case. Thus, any loss incurred due the yield problem illustrated earlier is much less in the second manufacturing line. To reiterate, the manufacturing line with more failure analysis capacity is less sensitive and more productive, as borne out by these simulation.

When yield degradation occurs for polysilicon defects, the situation is as illustrated by the layer yield curves in Figure 7.10. In this case, metal3 yield learning remains unaffected by the disturbance and polysilicon yield also recovers quickly. This is because, the metal yield is high enough that there is extra capacity available to analyze polysilicon defects. Although not shown, the cost impact is also less severe.

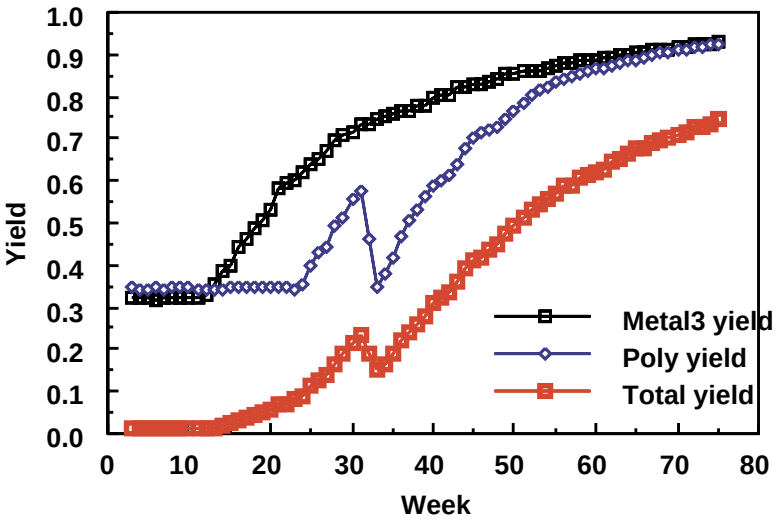

Figure 7.10 Layer yield trends for polysilicon yield degradation.

7.4 Yield Learning Dependence on Product Design

In this section, the two product factory designed for CMOS and DRAM processes presented in Section 6.1 will be considered to illustrate yield learning rate dependence on product attributes. This factory is designed to operate for 832 and 1664 WSPW for

the CMOS and DRAM products, respectively. The same defect types i.e., polysilicon and metal shorts, are considered as in previous cases. The die size is also assumed to be the same as for the CMOS product, i.e. 1.4 cm^2. However, several important differences in attributes of the CMOS and DRAM products are assumed. These are:

1. Defects in DRAM are more diagnosable than in the CMOS product. This is modeled by assuming a smaller mean search area, A_s, for DRAM 0.08 cm^2, than CMOS, A_s being 0.5 cm^2 (variances are 0.0002 and 0.008).
2. DRAM is more sensitive to polysilicon shorts and is comparable to the sensitivity of CMOS to metal1 shorts. Sensitivity to metal shorts in both products is assumed to be similar.
3. There are two metal levels in the DRAM compared to three in the CMOS product. The critical areas assumed for each defect type are given in Appendix C.

All other assumptions for cleaning model and defect diagnosis equipment parameters are the same as in the previous examples of yield learning.

Figure 7.11 shows the yield learning curves for the CMOS and the DRAM products when the CMOS product alone is sampled for performing defect diagnosis. The final yield attained in 75 weeks of simulation is 0.48 for CMOS and 0.41 for DRAM. Note that the yield of the DRAM product is less than the CMOS entirely because of significantly lower polysilicon yield for DRAM. Although the CMOS product has one more metal layer, the higher density assumed for the polysilicon defects more than compensates for it.

Instead, if only the DRAM product is sampled for defect diagnosis then the maximum yields attained are 0.68 and 0.60 for the CMOS and DRAM products, respectively. For comparison, the yield learning curves for DRAM product for the two defect diagnosis cases discussed is shown in Figure 7.12. This difference in learning rates amounts to a significant gain in terms of productivity and cost of good die, and is shown in Table 7.2. This analysis not only illustrates the importance of developing

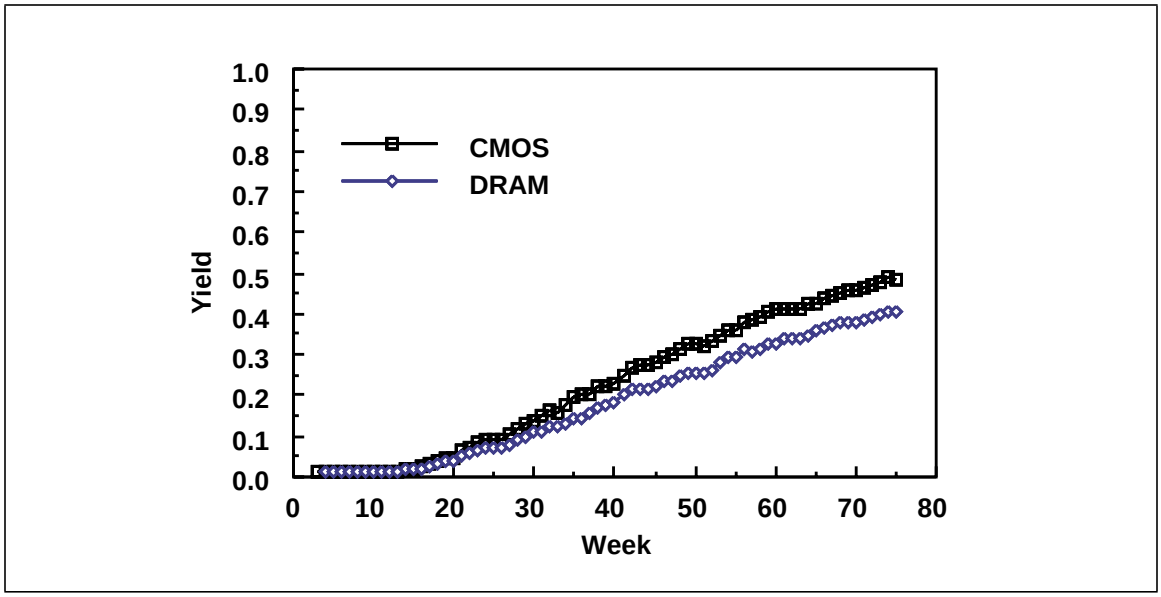

Figure 7.11 Yield learning curves when CMOS product is sampled for analysis.

proper models to differentiate diagnosability of products, but also that such analysis can be applied to quantify differences in cost benefits.

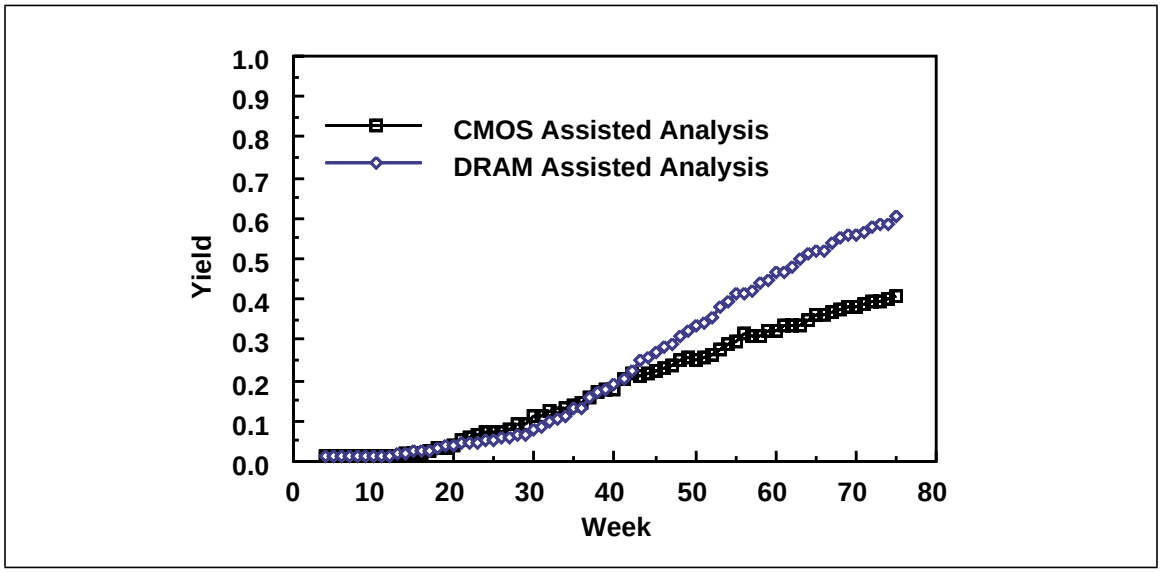

Figure 7.12 Comparison of yield learning curves of DRAM.

The advantage of using a product with high diagnosability was illustrated by setting the area of search for defects to be very low i.e. mean = 0.08 cm^2 and variance = 0.0002

	CMOS Assisted Analysis		DRAM Assisted Analysis	
	CMOS	DRAM	CMOS	DRAM
Number of good die (millions)	1.518	2.804	2.107	3.605
Cost of good die ($)	125	161	95	126
% of cost from failure analysis	5.5	4.27	6.08	4.57

Table 7.2 Productivity and cost comparison,

for A_s. One can also explore a spectrum of diagnosability conditions by varying A_s. The results obtained through such experiments are shown in Table 7.3 to further illustrate the dependence of productivity and cost on efficiency of failure analysis. The table indicates that the case with mean A_s = 0.16 cm^2 results in higher productivity and lower cost than that for mean A_s = 0.08 cm^2. Further investigation revealed that the yield learning rate for the polysilicon defects is faster for the case when mean A_s = 0.16 cm^2 as shown in Figure 7.13 for the DRAM product. This is because, the variance in A_s results in a higher probability of occurrence of chips with too low a diagnosability value. This results in preference being given to some of the diagnosable polysilicon defects. In the case where mean A_s = 0.08 cm^2, the higher rate of occurrence of diagnosable metal defects results in much less capacity available for diagnosing polysilicon defects. This result, of course, is an artifact of the chosen model and it illustrates the need to verify such relationships in practice.

Increasing failure analysis capacity shows exactly the same trend with a higher level of productivity and a lower level of good die cost. As presented in Section 7.2, the change is not proportional to the increase in failure analysis capacity. These experi-

	mean As = 0.08 var As = 0.0002		mean As = 0.16 var As = 0.0008		mean As = 0.32 var As = 0.0032		mean As = 0.4 var As = 0.005	
	CMOS	DRAM	CMOS	DRAM	CMOS	DRAM	CMOS	DRAM
Number of good die (millions)	2.017	3.605	2.086	4.400	1.758	3.900	1.364	3.007
Cost of good die ($)	95	126	91	103	108	117	139	151
% of cost from failure analysis	6.08	4.57	5.54	4.89	5.87	5.41	5.23	4.80

Table 7.3 Productivity and cost comparison for different diagnosability conditions.

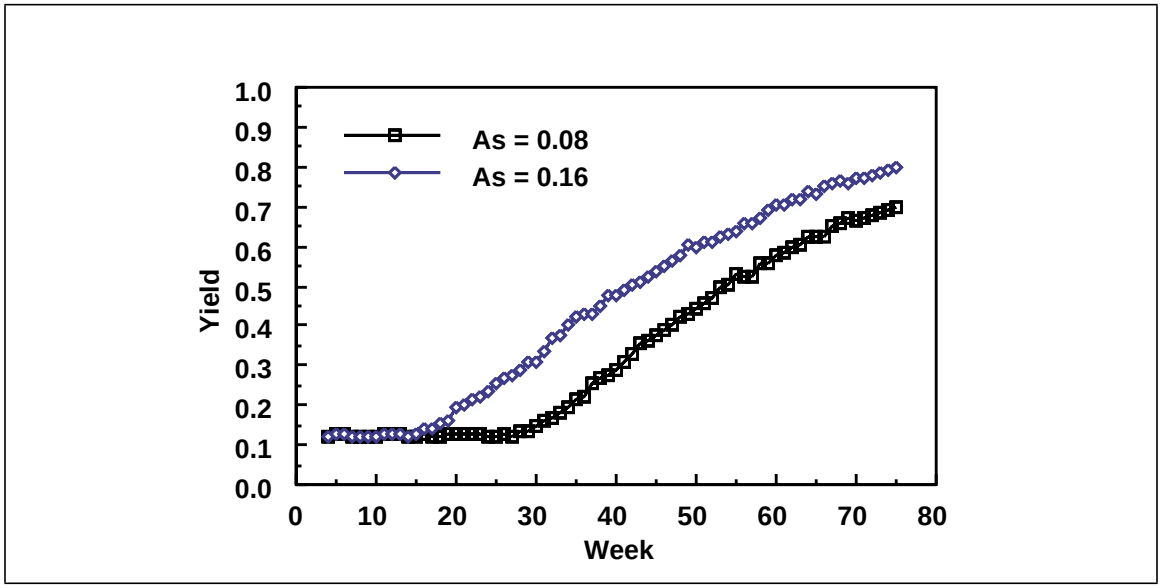

Figure 7.13 Polysilicon yield comparison for A_s = 0.08 and A_s = 0.16 cm for DRAM[2].

ments also illustrate that for a given capacity, the dominating factor in determining yield learning rate is the proportion of diagnosable to undiagnosable defects of each type.

7.5 Effect of Delayed Product Introduction on Productivity

In a two product factory, such as the one illustrated in the previous section, one has the option to enhance productivity by properly delaying the introduction of the second product in a fabrication line. One can also view this as introducing a second product in a single product line at an "optimal" time. In this way one can take advantage of the fact that during the initial low yield period, a larger sample of diagnosable dies are available. To illustrate this, the experiment has been set up in the following way.

First, failure analysis related product attributes are kept exactly the same as in the previous section for the case where the mean and variance of A_s for DRAM are 0.16 cm^2 and 0.0008. The difference is that the wafer start rate now depends on time and is illustrated in Figure 7.14. Initially, the DRAM WSPW is set near capacity of the line. After a time interval, T_i, the CMOS product is introduced at 832 WSPW, and DRAM WSPW is ramped down to 1664 WSPW. these values being the design point of the fabrication line. A set of experiments were conducted with three values of T_i from 30 to 40 weeks in steps of 5 weeks each. Note that in the previous section all the results presented were for $T_i = 0$.

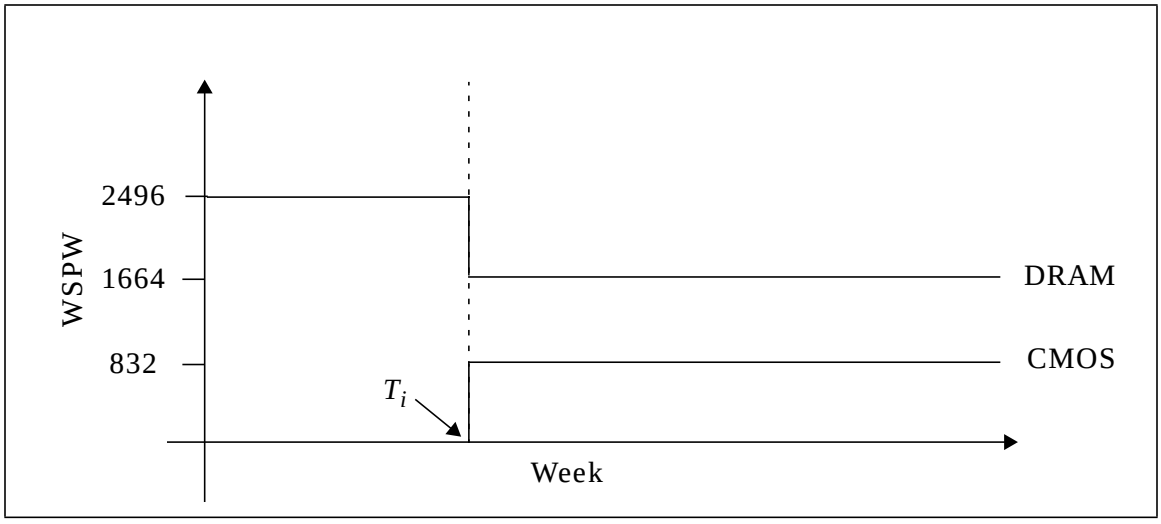

Figure 7.14 Wafer start rate setup.

First the case where T_i = 30 is compared with the DRAM product yield trends where T_i = 0 as presented in previous section. The two yield learning curves are presented in Figure 7.15, and it is observed that the DRAM product shows a small increase (~4%) in yield learning rate for T_i = 30. As an illustration, the CMOS yield learning curve is

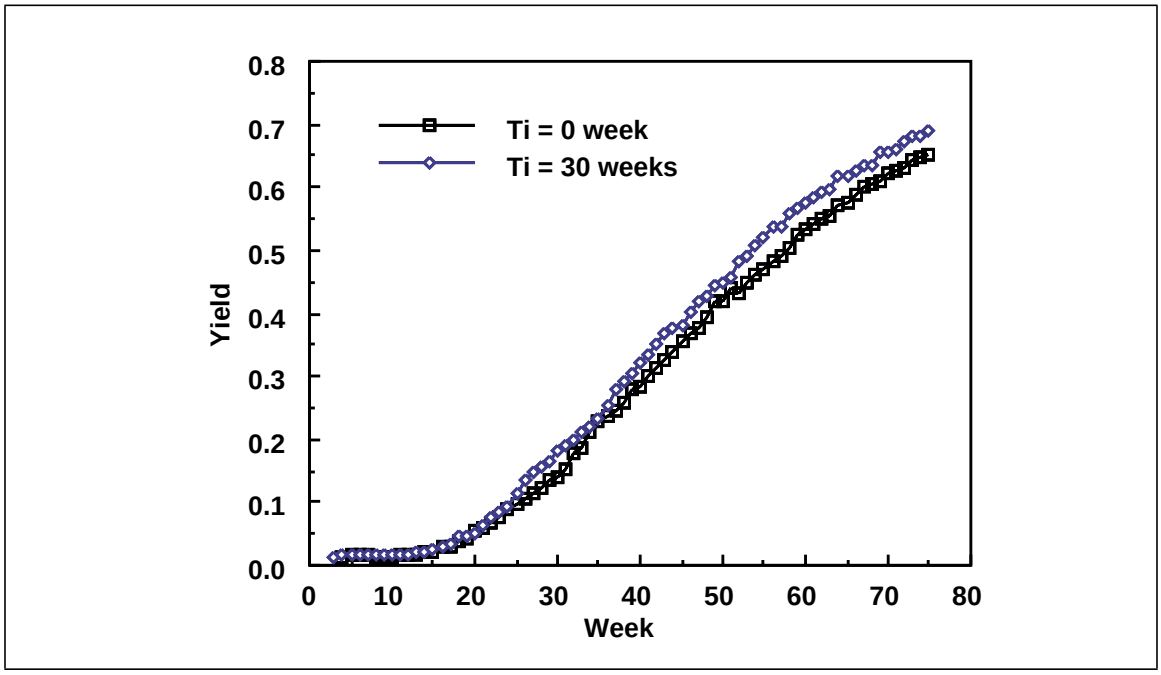

Figure 7.15 Comparison of the yield learning curves for DRAM.

shown along with the DRAMs in Figure 7.15. The impact on the productivity and cost of good die for both CMOS and DRAM products is significant. A summary of the results for the number and cost of good die is presented in Table 7.4. For T_i = 30, the number of good die produced is slightly smaller for CMOS but the savings in cost is significant. DRAM product on the other hand shows a small increase in both productivity and cost of good die. This experiment shows that there is a possibility to optimize the overall cost performance of a fabrication line by quantifying the trade-offs in such strategic decisions.

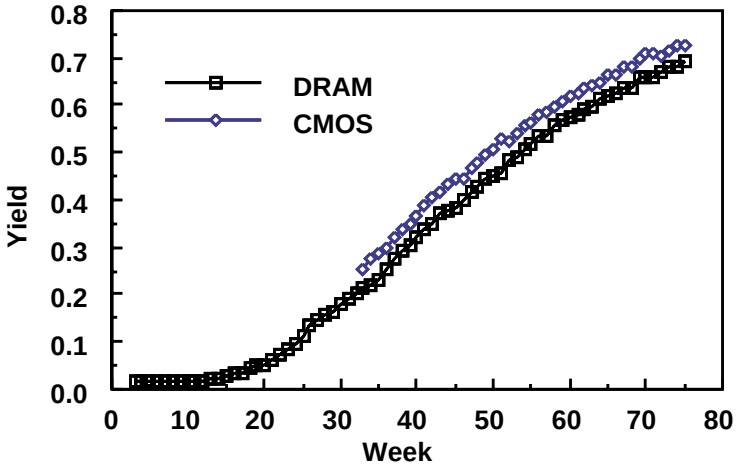

Figure 7.16 Illustration of yield trends for delayed product introduction.

	Ti = 0 week		Ti = 30 weeks		Ti = 35 weeks		Ti = 40 weeks	
	CMOS	DRAM	CMOS	DRAM	CMOS	DRAM	CMOS	DRAM
Number of good die (millions)	2.086	4.400	1.942	4.722	1.818	4.869	1.658	5.059
Cost of good die ($)	91	103	60	114	57	113	55	112
% of cost from failure analysis	5.54	4.89	8.33	4.3	8.56	4.34	8.92	4.35

Table 7.4 Productivity and cost comparison for different T_i values.

7.6 Summary

It has been shown that Y4 is capable of simulating scenarios which are relevant to cost-revenue trade-off studies. Specifically, a sudden increase in wafer start rates can lead to instabilities in the factory and the cost impact of such a disturbance should be studied in its proper perspective. Secondly, increasing capacity of failure analysis for the simulated factory can increase the productivity. Further, a higher capacity failure

analysis facility is more quickly able to absorb the shock of a sudden degradation in yield. In a multi-product factory, there may exist various possibilities to improve yield learning rate. First, cost benefit can be significant enough to justify allocating resources to increase the diagnosability of a product. Second, one can affect the yield learning rate and cost performance by strategically delaying the manufacturing of undiagnosable products.

Judging cost impact purely from previous experience may be difficult if not impossible since a number of interacting factors affect the cost evaluation of a given manufacturing line. For the sake of simplicity, many factors such as operator interaction, applicability of particle scanners in yield learning, etc., have been ignored. It must be noted that any two factories are unlikely to be the same and that the results presented here are specific to the factories considered and the assumptions made. However, there is reason to believe that the trends observed in our simulations should be replicable in other situations.

References

[1] Rick McKiddie, “Some No-Panic Help for Wafer-Start Surges”, *Semiconductor International*, pp. 115-120, June 1995.

[2] P. K. Nag and W. Maly, “Cost of Ad Hoc Wafer Release Policies”, *Int. Symp. on Semiconductor Manufacturing (ISSM)*, pp. 97-102, Nov. 1995.

Chapter 8

Future Work

The research presented in this thesis was mainly aimed at understanding the nature of the yield learning process. A great portion of the effort went into capturing the primary factors leading to yield learning in a manufacturing line. An important outcome of this research is the realization that predicting yield learning curves is complex task. It is so complex that research in yield learning should be continued. This chapter attempts to indicate some of the directions which should expand the domain of the presented work.

8.1 Model Development

First, let us analyze the shortcomings of the models used to mimic the fabrication phase. The primary factor which has not been addressed in development of WSIM are the operators in the line. In industry, variability in the line introduced by operators is an important concern. One can mimic the scheduling of operators in Y4 but the implementation is rudimentary.

Secondly, operating rules studied in this research were limited to only a subset of the possibilities. For example, rules for merging and re-assigning wafers to new sets of lots has not been considered. In reality, scheduling of wafers is also guided by expected due dates and by estimating equipment loading before and after the piece of equipment where rules are applied [1]. Such dynamic estimates of performance of fabrication lines and adjusting scheduling policies have not been considered.

A manufacturing line undergoes a lot of changes during its lifetime. Equipment are removed or added as old products are phased out or new products added. Similarly, process recipes also undergo changes or evolve as the manufacturing line properties change. New pieces of equipment may have different contamination properties which alter the rate of yield learning. For example, a new source of contamination may be introduced in the line just by incorporating a new piece of equipment in a line. The evolutionary nature of a manufacturing line has not been dealt with in development of Y4.

The yield models used in this work are adapted from existing models presented and used by various researchers. As noted earlier, a given defect type can be caused by many different types of particles whose size distributions are not known or available. For the purpose of this work they were was assumed to have the same form as the defect size distributions. But the cumulative effects of randomly distributed particle sizes is unlikely to retain the same distribution for defects. This aspect needs to be studied and quantified further to achieve an understanding between relationships between particle size and defect size distributions.

Spatial distributions of particles and defects on the wafer surface cannot be addressed using the present models. Though simple die-to-die variatons can be incorporated easily, intra die variations cannot be taken into account. Appropriate models must be developed to reflect the spatial variation of defects.

Variability in contamination related yield cannot be modeled well in Y4. For example, etching variations can lead to variability in critical areas from wafer to wafer [2]. As the yield is sensitive to such variations, yield learning rates could be affected by these process excursions. Hence, it is important that models be implemented to mimic yield variability and simulation experiments be conducted to quantify its relationship with yield learning rates.

The models to mimic the effect of equipment cleaning are rather simple. Further, there is no representative data available. Thus, research in this area must be directed

towards better modeling of equipment and collecting data from the industry. As clear from the presented work, the models for mimicking the effect of cleaning/repairing must be consistent with models for particle rates and size distributions. Thus, any changes in models for particles in the line must also be reflected in equipment cleaning models.

Efficiency and accuracy of defect diagnosis are modeled in this work using the concept of a diagnosability measure. The model development was mainly guided by the fact that one must take into account the product, defect and analysis equipment characteristics. The actual functional form presented only reflects this objective. In order to reflect reality, research must be directed towards quantifying the relationship between product and diagnosis attributes using controlled experiments. This can be a complex task since failure analysis is essentially an “ad hoc” process which heavily relies on human expertise and the history of the fabrication line. Thus, experiments must be designed to decouple product, defect, equipment and human attributes and to quantify their individual effects on accuracy and efficiency of defect diagnosis process.

The particle monitor model of Y4 is not able to mimic several of the possibilities. First, models are needed for less than 100% accuracy of particle monitors in determining both size and location of defects. Second, a model must be built to predict the probable faults that result from detected particles. Simulation experiments must then be conducted to determine the correlation between predicted and detected faults. This will provide some quantification of the usefulness of employing particle monitors for local feedback control of the fabrication line.

Defect diagnosis can be far more accurate if a correlation exists between predicted and detected faults. One can use wafers that have been scanned for defect diagnosis purposes and quantify the relationship with accuracy and speed of diagnosis. The trade-off analysis capability of Y4 could be vastly improved by taking into consider-

ation that several mechanisms may be concurrently available to speed up the defect diagnosis process.

One important drawback of cost modeling in Y4 is that it does not consider any of the non-volume dependent costs such as the cost of IC design and the cost of test generation. To obtain a correct perspective of cost impact one must consider these in cost calculations since they can have a significant contribution.

Lastly, one needs to identify simpler macro models which mimic the combined effects of more detailed micro models. The models of Y4 are not currently organized in any fashion. Models can be developed and implemented hierarchically in such a way that one can perform accurate detailed simulation or a coarse simulation or a mixture of both as deemed necessary.

Verification must be carried out not only at the individual model level but also from the perspective of an entire manufacturing line. Y4 can be used as a base simulator to mimic the properties of a manufacturing line and compare with real industrial data. This serves two purposes: first, to verify that the models consider the relevant factors appropriately and second, to identify the model parameters which have significant impact on observed criteria such as cost.

8.2 Statistical Tools for Tuning Model Parameters

Modern manufacturing lines have integrated data collection capability directly connected to a central network. An ideal situation when Y4 is directly coupled to a manufacturing line data acquisition system. The data so gathered could be used to extract model parameters for Y4. Similar data could then be collected from Y4 and the corresponding parameters extracted and compared. Such a setup can be used in conjucntion with statistical tools to measure and possibly correct any error between a manufacturing line and Y4. In such a setup, one can think of using Y4 for short term

forecasts with higher confidence and use the forecast measures to apply “corrective actions” on the manufacturing line.

Tuning of the six submodules in Y4 must be undertaken in a sequence of steps. WSIM should be tuned first and can be done independently of other modules followed by COSIM. YSIM should be tuned next followed by TSIM. Tuning of YSIM and TSIM may need to performed together and iteratively depending on the available data. PSIM should be tuned next and again it may require retuning the YSIM module. Lastly, FASIM should be tuned in a manner similar to WSIM.

WSIM needs to be tuned to resemble the properties of a given fabrication line by matching the mean and variance of the cycle times obtained through simulations to those measured in a fabrication line. Tuning the cycle time essentially means that the distribution of time to equipment breakdown and repair must first be correctly extracted. Secondly, operator timing distribution must be extracted from measured data in a fabrication line.

Tuning YSIM requires that all the possible sources of particle types be known first. The parameters of particle number and size distributions can be tuned by matching the observed means and variances of the different fault types. In the extreme case only the total yield data may be available without any classification; in this case individual parameters can only be approximated. It is important that the data used for tuning be obtained from a stable fabrication line as much as possible. The changes in particle parameters must be isolated in such a way that the model parameters for corrective actions can be determined.

Tuning TSIM can be easier if the real fault coverage and corresponding cycle time data is available. This should be obtained directly from the manufacturing line for a particular product. Since the cycle time is dependent on the defect levels to a certain extent, tuning of TSIM may require simultaneous retuning of YSIM.

PSIM can be tuned only after the parameters of models in WSIM and YSIM have been tuned to a first approximation. The parameters of the model for accuracy and efficiency of particle monitors can be extracted directly from measured data in a fabrication line. But more importantly, the measured data can also be used to fine tune the model parameters of a subset of the particle sources. The subset of particle sources that can be tuned depends on the particular steps at which the monitors are employed.

Tuning FASIM requires that experiments be conducted to track the time required to analyze a defective die in different equipment. A first approximation of the model parameters can be obtained in this way. It is important that each measured data on timing for a defective die be associated with the layer of the defect and the approximate area of search for each die. The model parameters should be fine tuned in such a way that the predicted cycle time for analyzing a wafer match the measured data.

In summary, based on past experience in tuning Y4, it has often been necessary to observe statistics on various factors such as cycle time, yield, diagnosis rate, corrective feedback rate, etc., simultaneously. This is useful to correctly determine reasonable first approximations for model parameters.

8.3 Y4 Enhancements

Currently, the user interface of Y4 is rudimentary, achieved through a number of files. A window based interface must be implemented to facilitate the editing of input data and, analysis and visualization of performance measures. A graphical editor for designing a manufacturing floor, process sequencing and equipment assignment will be useful additions to Y4. Layout parameters like critical area are input parameters to Y4. One can enhance the interface by having Y4 acquire these parameters by sending appropriate events to a layout parameter extraction tool like MAPEX [3] or appropriately designed version of CREST.

One must be able to stop the simulation at an intermediate state and save the state variables. Such a feature will enable Y4 to restart a simulation from an equilibrium condition without going through the warm-up period. One can also randomize the initial conditions of simulations and avoid possible systematic dependency in simulation. An even more important outcome of such a feature is that one can stop and restart a simulation with altered conditions. Such alterations could be the addition of a product or a piece of equipment with new contamination levels, etc. A visual interface could greatly facilitate this since one can stop the simulation at a certain point depending on the dynamically updated feedback of measures such as cost, yield, etc.

Lastly, effort should also be directed towards improving the speed and reducing the memory requirements of Y4. Simulation time requirement can grow rapidly if one considers hundreds of products, large capacity, many different particle sources, etc. Resource requirements of Y4 have not been characterized well and need to be evaluated in future development.

References

[1] L. F. Atherton and R. W. Atherton, *Wafer Fabrication: Factory Performance and Analysis*, Kluwer Academic Publishers, 1995.

[2] I. Bubel, W. Maly, T. Waas, P. K. Nag, H. Hartmann, D. Schmitt-Landsiedel and S. Griep, “AFFCCA: A Tool for Critical Area Analysis with Circular Defects and Lithography Deformed Layouts”, *in Proc. of Int. Workshop on Defect and Fault Tolerance in VLSI Systems*, pp. 10-18, Nov. 1995.

[3] H.T. Heineken and W. Maly, “Manufacturability analysis environment - MAPEX,” *Proc. Custom Integrated Circuits Conference*, pp. 309-312, May 1994.

Chapter 9

Conclusions

With the cost of building a new manufacturing line nearly doubling every generation, it has become necessary to carefully analyze cost-revenue trade-offs before any decisions are taken. One of the significant contributors to cost in a modern manufacturing line is yield loss due to contamination and the time required to ramp-up the yield to profitable levels. Thus, it is important to not only understand the reasons for yield loss, but also to quantify the various attributes of fabrication, product and failure analysis that determine the yield learning rate.

In the past, evaluation of manufacturing performance was mainly focussed on characterizing attributes like product performance, diagnosability, testability, etc., independently of each other. But looking at the manufacturing process as a system consisting of a number of interacting components makes one realize that yield learning is tied to this complex inter-dependence. Hence, the first contribution of this research is the documentation and discussion of this inter-dependence, and the understanding gained of the cross-disciplinary nature of yield learning. The deficiencies of earlier simplistic models of yield learning are discussed in the context of both the timing of yield improvement cycles, and the change in yield as a result of corrective actions.

A methodology to predict contamination related yield learning curves for a multi-product manufacturing line has been developed. This methodology is based upon the observation that the yield learning process can be viewed as the super-imposition of a

number of yield improvement cycles for each particle and defect type considered. In this thesis models have been developed for:

1. Yield loss as a function of particle, defect, fault and layout attributes.
2. Product and defect attributes which decide diagnosability.
3. Failure analysis attributes which decide timing of the diagnosis process.
4. Effect of corrective actions on the change in yield loss.
5. Cost contribution of fabrication, testing and failure analysis.

These models have been implemented in a discrete event-based prototype simulator - Y4.

In order to test the functionality of Y4 and its models, a number of simulation experiments were conducted. The basic experiment consisted of testing the operational aspects of a fabrication line demonstrated through cycle times and cost estimations for a single and a two-product factory. Yield simulation capability was demonstrated for a spectrum of defect characteristics showing that known dependencies can be replicated. Imperfect test quality simulation was illustrated through simple experiments considering a range of fault coverage values. Defect diagnosis simulation has been illustrated by computing yield learning curves for several defect types. It was shown that the model has the capability to distinguish product attributes like uncertainty in fault-location and defect attributes like size and layer of IC. Particle monitoring simulation capability has been demonstrated through estimating the impact of yield due to wafer rejection and corrective actions.

Several possible areas where Y4 can be used for evaluating cost-revenue trade-off experiments have been identified and illustrated. It was shown that introducing wafer surges, even for a short duration, can have a strong impact on the stability and cost performance of the line. It was also concluded that, depending on the scheduling rules employed, assigned lot priorities and the designed capacity of the fabrication line can have an impact on the cost.

The possibility of a strong impact of increasing failure-analysis capacity on the yield learning rate has been demonstrated: to the extent that the cost of failure analysis is more than compensated by the improvement in productivity. It was demonstrated that a higher failure analysis capacity better absorbs yield disturbances as a result of a sudden increase in particle rates. This is an important consideration since new sources of particles are often introduced in the line and are not easily identified. Extra capacity of failure analysis may increase short-term cost, but in the long run, it can reduce manufacturing cost.

Y4 has been applied to evaluate yield learning curves for a two product manufacturing line employing a DRAM and a CMOS process respectively. It was shown that products can be distinguished based on their diagnosability attributes. Later it was shown that a highly diagnosable product like DRAM can be effectively used for increasing the yield learning rate. The importance of understanding and modeling of products and failure analysis equipment from a diagnosability point of view was demonstrated. Considerable cost gains were shown to be achieved by improving product diagnosability.

A lot of the strategic activities in industry involves introducing new products in time for the market and also removing or ramping down outdated products. For the former scenario it was shown that it is possible to “optimize” the time of introduction of a new product for maximization of productivity. Simulation experiments were presented to show that a highly diagnosable product can be used to ramp up the yield faster before introducing a second undiagnosable product. In reality, one needs to plan ahead for the allocation of necessary resources like equipment, personnel, etc., in time to be ready for the next product. Although this aspect is not illustrated but it is clear that such a capability can be very helpful in planning the allocation of resources.

In retrospect, this work clearly demonstrates the importance of modeling manufacturing line attributes and their interdependence, in spite of a number of limitations of its models and assumptions.

Appendix A

Process Recipes

In this appendix, the steps for the CMOS and DRAM process recipes used in the simulation examples are presented. The steps are defined and numbered sequentially and there are four fields describing each step:

1. *Description of step*: This field gives a brief description of the nature of the process step.
2. *Feature*: This field describes the feature of IC that is affected during the process steps. This serves as a rough classification of the steps.
3. *Time*: This field specifies the time required, in minutes, to process one lot (24 wafers). For batch equipment this is equal to the actual processing time for one load.
4. *Workstation*: This field specifies the work-station to be used for performing the necessary step.

A.1 CMOS Process Recipe

The 0.5 micron, 3-metal CMOS process recipe consists of 145 steps and is given in Table A.1.

Step #	Description of step	Feature	Time (minutes)	Workstation
1	begin_process	begin	20.0	Begin
2	mark_lines	scribe	23.0	Scriber1
3	init_clean	scribe	52.0	TClean1

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
4	init_oxide	pwell	524.0	Boven1
5	litho_pwell	pwell	200.0	Litho1
6	etch-ox-nitr	pwell	100.0	Strip2
7	pwell_implant	pwell	22.0	Implant2
8	ox-etch-strip	pwell	100.0	Strip2
9	clean_surf	pwell	71.0	TClean1
10	well_oxide+drive	pwell	398.0	Boven2
11	ox_nit-etch	pwell	153.0	Strip2
12	clean_surf2	pwell	51.0	TClean1
13	sec-implant	pwell	22.0	Implant2
14	clean_surf3	pwell	72.0	TClean1
15	sec-drive-in	pwell	403.0	Boven6
16	ox-etch1	pwell	37.0	Strip3
17	clean-surf3	pwell	51.0	TClean1
18	nitrid-ox-dep	pwell	282.0	Boven1
19	poly-sil-dep	pwell	202.0	Boven7
20	nitrid-ox-dep2	pwell	241.0	Boven1
21	isol-litho	field-ox	212.0	Litho1
22	isol-etch	field-ox	268.0	Strip5
23	isol-back-etch	field-ox	112.0	Strip2
24	clean-surf4	field-ox	51.0	TClean1
25	isol-ox-dep	field-ox	430.0	Boven5
26	isol-nit-etch	field-ox	94.0	Strip7
27	clean-surf5	field-ox	51.0	TClean1
28	aox-drive	field-ox	464.0	Boven6
29	aox-etch	field-ox	47.0	Strip2

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
30	clean-surf	field-ox	51.0	TClean1
31	streu-ox	field-ox	190.0	Boven2
32	part-meas	field-ox	20.0	Surf1
33	init-implant	vt-adjust	21.5	Implant2
34	ox-surf-etch	gate	36.0	Strip7
35	clean-surf7	gate	54.0	TClean1
36	thin-ox	gate	182.0	Boven2
37	poly-dep	poly	246.0	Boven7
38	poly-reflow	poly	199.0	Boven9
39	etch-poly-ox	poly	35.0	Strip9
40	poly-back-etch	poly	49.0	Strip6
41	clean-surf8	poly	51.0	TClean1
42	teos-dep	poly	210.0	Boven10
43	sputter-a-sil	poly	105.0	Dep1
44	meas-part2	poly	20.0	Surf1
45	poly-litho	poly	211.0	Litho2
46	poly-a-sil-etch	poly	91.0	Strip11
47	poly-res-etch	poly	112.0	Strip4
48	poly-etch	poly	114.0	Strip1
49	poly2-res-etch	poly	191.0	Strip10
50	poly-clean-surf	poly	51.0	TClean1
51	poly-teos-dep	poly	214.0	Boven10
52	poly-meas-part	poly	20.0	Surf1
53	polyox-etch	poly	206.0	Strip5
54	clean-surf9	poly	51.0	TClean1
55	post-ox-ldd	poly	160.0	Boven11

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
56	ntran-litho	source-drain	200.0	Litho2
57	ntran-impl-ldd	source-drain	50.0	Implant2
58	ntran-res-etch	source-drain	100.0	Strip8
59	ntran-litho2	source-drain	200.0	Litho2
60	ntran-impl-ldd2	source-drain	39.0	Implant1
61	ntran-res-etch2	source-drain	129.0	Strip8
62	clean-surf10	source-drain	51.0	TClean1
63	ox+Teos-dep	source-drain	295.0	Boven10
64	tran-etch-ox	source-drain	194.0	Strip5
65	clean-surf11	source-drain	51.0	TClean1
66	tran-post-ox	source-drain	200.0	Boven11
67	ntran-litho3	source-drain	200.0	Litho2
68	ntran-implant3	source-drain	48.0	Implant2
69	ntran-res2-etch	source-drain	195.0	Strip8
70	clean-surf12	source-drain	51.0	TClean1
71	ntran-bake	source-drain	200.0	Boven8
72	ptran-litho	source-drain	200.0	Litho2
73	ptran-implant	source-drain	105.0	Implant2
74	ptran-res-etch	source-drain	158.0	Strip8
75	ptran-back-etch	source-drain	31.0	Strip6
76	clean-surf13	source-drain	51.0	TClean1
77	cvd+teos	Foxide	135.0	Boven12
78	ox-reflow	Foxide	82.0	Boven14
79	ox-planar	Foxide	317.0	Strip5
80	fox--etch	Foxide	53.0	Strip18
81	fox-post-allign	Foxide	23.0	Scriber1

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (min-utes)	Workstation
82	clean-surf14	Foxide	51.0	TClean2
83	meas-part-ox	Foxide	20.0	Surf1
84	contact-litho	metal1	212.0	Litho2
85	ox-res-etch	metal1	207.0	Strip18
86	defect-cont1	metal1	13.0	Mikro1
87	hand-clean	metal1	56.0	TClean2
88	clean-surfx	metal1	18.0	MClean1
89	metal1-dep	metal1	106.0	Sputter1
90	met-part-meas	metal1	20.0	Surf1
91	metal1-dep2	metal1	222.0	Sputter2
92	metal1-clean	metal1	28.0	MClean1
93	met-part-meas2	metal1	20.0	Surf1
94	metal1-dep3	metal1	137.0	Sputter2
95	met-part-meas3	metal1	20.0	Surf1
96	metal1-litho	metal1	212.0	Litho2
97	metal1-etch	metal1	127.0	Strip13
98	metal1-clean2	metal1	18.0	MClean1
99	metal1-etch2	metal1	120.0	Strip13
100	metal1-part-meas	metal1	20.0	Surf1
101	imox-sog	Soxide	392.0	Boven15
102	planar-backetch	Soxide	107.0	Strip5
103	imox-sec	Soxide	112.0	Boven17
104	imox-part-meas	Soxide	20.0	Surf1
105	met2-cont-litho	metal2	212.0	Litho2
106	cont-res-oxetch	metal2	355.0	Strip18
107	cont-met-etch	metal2	49.0	Strip14

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
108	cont-polymer	metal2	26.0	Dep1
109	cont-def-cont	metal2	13.0	Mikro1
110	met2-con-dep	metal2	119.0	Sputter1
111	met2-part-meas	metal2	20.0	Surf1
112	met2-con-dep2	metal2	233.0	Sputter1
113	met2-con-clean	metal2	28.0	MClean1
114	met2-part-meas2	metal2	20.0	Surf1
115	met2-dep	metal2	142.0	Sputter2
116	metal2-litho	metal2	212.0	Litho2
117	metal2-etch	metal2	127.0	Strip14
118	metal2-clean	metal2	18.0	MClean1
119	metal2-etch2	metal2	120.0	Strip14
120	metal2-def-cont	metal2	13.0	Mikro1
121	imox-sog-2	Toxide	392.0	Boven15
122	planar-backetch-2	Toxide	107.0	Strip5
123	imox-sec-2	Toxide	112.0	Boven17
124	imox-part-meas2	Toxide	20.0	Surf1
125	met3-cont-litho	metal3	212.0	Litho2
126	cont2-res-oxetch	metal3	355.0	Strip18
127	cont2-met-etch	metal3	49.0	Strip14
128	cont2-polymer	metal3	26.0	Dep1
129	cont2-def-cont	metal3	13.0	Mikro1
130	met3-con-dep	metal3	119.0	Sputter1
131	met3-part-meas	metal3	20.0	Surf1
132	met3-con-dep2	metal3	233.0	Sputter1
133	met3-con-clean	metal3	28.0	MClean1

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

Step #	Description of step	Feature	Time (min-utes)	Workstation
134	met3-part-meas2	metal3	20.0	Surf1
135	met3-dep	metal3	142.0	Sputter2
136	metal3-litho	metal3	212.0	Litho2
137	metal3-etch	metal3	127.0	Strip14
138	metal3-clean	metal3	18.0	MClean1
139	metal3-etch2	metal3	120.0	Strip14
140	metal3-def-cont	metal3	13.0	Mikro1
141	passv-dep	passv	144.0	Boven19
142	passv-litho	passv	200.0	Litho1
143	passv-res-etch	passv	250.0	Strip5
144	passv-bake	passv	150.0	Boven16
145	end-process	end	57.0	End

Table A.1 0.5 micron 3-metal CMOS Recipe Steps.

A.2 DRAM Process Recipe

The recipe for a 0.5 micron, 2-metal trench capacitor DRAM process consists of 174 steps and is given in Table A.2.

Step #	Description of step	Feature	Time (min-utes)	Workstation
1	begin_process	begin	20.0	Begin
2	mark_lines	scribe	23.0	Scriber1
3	init_clean	scribe	52.0	TClean1
4	init_oxide	blayer	524.0	Boven1
5	meas-part1	blayer	20.0	Surf1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
6	litho_blayer	blayer	218.0	Litho3
7	blay_ox_etch	blayer	363.0	Strip1
8	clean-surf	blayer	71.0	TClean1
9	blayer_ox	blayer	168.0	Boven2
10	blayer_implant	blayer	31.0	Implant1
11	blayer_clean	blayer	71.0	TClean2
12	blayer-diff	blayer	356.0	Boven3
13	ox-etch-strip	blayer	196.0	Strip2
14	clean_surf	blayer	71.0	TClean2
15	blayer-implan2	blayer	16.0	Implant2
16	blayer-clean1	blayer	71.0	TClean2
17	blayer-anneal	blayer	270.0	Boven4
18	blayer-etch	blayer	40.0	Strip3
19	blayer-clean2	blayer	53.0	TClean2
20	epitaxy-dep	epitaxy	346.0	Epi1
21	epitaxy-clean	epitaxy	51.0	TClean1
22	epitaxy-nit-ox	epitaxy	261.0	Boven1
23	epitaxy-ox-etch	epitaxy	29.0	Strip3
24	epitaxy-clean2	epitaxy	51.0	TClean1
25	epitaxy-nit-ox2	epitaxy	487.0	Boven1
26	meas-part2	epitaxy	20.0	Surf1
27	nwell-litho	epitaxy	218.0	Litho3
28	ox_nit-etch	epitaxy	175.0	Strip1
29	nwell-implant	nwell	31.0	Implant2
30	nwell-res-strip	nwell	308.0	Strip4
31	nwell-clean	nwell	71.0	TClean1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (min-utes)	Workstation
32	pwell-oxide	pwell	844.0	Boven5
33	nit-oxide-etch	pwell	211.0	Strip2
34	pwell-clean	pwell	51.0	TClean1
35	pwell-litho	pwell	218.0	Litho3
36	pwell-implant	pwell	31.0	Implant2
37	pwell-res-etch	pwell	147.0	Strip4
38	pwell-clean2	pwell	116.0	TClean1
39	pwell-drive	pwell	766.0	Boven6
40	pwell-etch	pwell	37.0	Strip3
41	pwell-clean3	pwell	51.0	TClean1
42	trench-pad-ox-nit	trench	260	Boven2
43	trench-pad-teos	trench	260	Boven13
44	trench-litho	trench	212	Litho1
45	trench-ox-nit-etch	trench	211	Strip9
46	trench-res-etch	trench	147	Strip8
47	trench-etch	trench	600	Strip19
48	trench-teos-etch	trench	150	Strip7
49	cap-sac-ox	capacitor	180	Boven11
50	cap-sac-etch	capacitor	120	Strip7
51	cap-node-ox	capacitor	350	Boven20
52	cap-poly-fill	capacitor	180	Dep1
53	cap-poly-anneal	capacitor	370	Boven8
54	cap-poly-etch-back	capacitor	220	Strip6
55	cap-teos	capacitor	280	Boven10
56	cap-etch	capacitor	300	Strip5
57	cap-poly2-fill	capacitor	80	Dep1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
58	cap-poly2-anneal	capacitor	160	Boven8
59	cap-back-etch	capacitor	60	Strip6
60	cap-etch2	capacitor	180	Strip5
61	iso-litho	isolation	212	Litho1
62	iso-etch	isolation	120	Strip5
63	iso-res-etch	isolation	100	Strip8
64	iso-teos-fill	isolation	340	Boven10
65	iso-litho2-planar	isolation	212	Litho1
66	iso-res-planar	isolation	200	Strip8
67	part-meas	isolation	20.0	Surf1
68	sac-ox-dep	vt-adjust	222.0	Boven20
69	sac-ox-etch	vt-adjust	32.0	Strip7
70	clean-surf5	vt-adjust	105.0	TClean1
71	thin-ox-dep	vt-adjust	176.0	Boven13
72	vt-litho	vt-adjust	218.0	Litho2
73	vt-impl1	vt-adjust	31.0	Implant2
74	vt-etch	vt-adjust	147.0	Strip8
75	vt2-litho	vt-adjust	218.0	Litho2
76	vt2-impl	vt-adjust	91.0	Implant1
77	clean-surf7	gate	54.0	TClean1
78	thin-ox	gate	182.0	Boven2
79	poly-dep	poly	246.0	Boven7
80	poly-reflow	poly	199.0	Boven9
81	etch-poly-ox	poly	35.0	Strip9
82	poly-back-etch	poly	49.0	Strip6
83	clean-surf8	poly	51.0	TClean1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
84	teos-dep	poly	210.0	Boven10
85	sputter-a-sil	poly	105.0	Dep1
86	meas-part3	poly	20.0	Surf1
87	poly-litho	poly	211.0	Litho2
88	poly-a-sil-etch	poly	91.0	Strip11
89	poly-res-etch	poly	112.0	Strip20
90	poly-etch	poly	114.0	Strip1
91	poly2-res-etch	poly	191.0	Strip10
92	poly-clean-surf	poly	51.0	TClean1
93	poly-teos-dep	poly	214.0	Boven10
94	poly-meas-part	poly	20.0	Surf1
95	polyox-etch	poly	206.0	Strip5
96	clean-surf9	poly	51.0	TClean1
97	post-ox-ldd	poly	160.0	Boven11
98	ntran-ldd-litho	source-drain	200.0	Litho2
99	ntran-impl-ldd	source-drain	50.0	Implant2
100	ntran-res-etch	source-drain	100.0	Strip8
101	ptran-ldd-litho2	source-drain	200.0	Litho2
102	ptran-impl-ldd2	source-drain	39.0	Implant1
103	ptran-res-etch2	source-drain	129.0	Strip8
104	clean-surf10	source-drain	51.0	TClean1
105	ox+Teos-dep	source-drain	295.0	Boven11
106	tran-etch-ox	source-drain	194.0	Strip5
107	clean-surf11	source-drain	51.0	TClean2
108	tran-post-ox	source-drain	200.0	Boven11
109	ntran-litho3	source-drain	200.0	Litho2

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
110	ntran-implant3	source-drain	48.0	Implant2
111	ntran-res2-etch	source-drain	195.0	Strip20
112	clean-surf12	source-drain	51.0	TClean2
113	ntran-bake	source-drain	200.0	Boven8
114	ptran-litho	source-drain	200.0	Litho2
115	ptran-implant	source-drain	105.0	Implant2
116	ptran-res-etch	source-drain	158.0	Strip20
117	ptran-back-etch	source-drain	31.0	Strip6
118	clean-surf13	source-drain	51.0	TClean2
119	local-int-nit	local-interc	150.0	Boven12
120	local-int-litho	local-interc	212.0	Litho1
121	local-int-nit-etch	local-interc	100.0	Strip9
122	local-int-poly-dep	local-interc	70.0	Boven7
123	local-int-anneal	local-interc	180.0	Boven8
124	local-int-etch	local-interc	230.0	Strip11
125	local-int-clean	local-interc	51.0	TClean2
126	cvd+teos	Foxide	135.0	Boven12
127	ox-reflow	Foxide	82.0	Boven14
128	ox-planar	Foxide	317.0	Strip5
129	fox--etch	Foxide	53.0	Strip18
130	fox-post-allign	Foxide	23.0	Scriber1
131	clean-surf14	Foxide	51.0	TClean2
132	meas-part-ox	Foxide	20.0	Surf1
133	contact-litho	metal1	212.0	Litho2
134	ox-res-etch	metal1	207.0	Strip18
135	defect-cont1	metal1	13.0	Mikro1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
136	hand-clean	metal1	56.0	TClean2
137	clean-surf15	metal1	18.0	MClean1
138	metal1-dep	metal1	106.0	Sputter1
139	met-part-meas	metal1	20.0	Surf1
140	metal1-dep2	metal1	222.0	Sputter2
141	metal1-clean	metal1	28.0	MClean1
142	met-part-meas2	metal1	20.0	Surf1
143	metal1-dep3	metal1	137.0	Sputter2
144	met-part-meas3	metal1	20.0	Surf2
145	metal1-litho	metal1	212.0	Litho2
146	metal1-etch	metal1	127.0	Strip13
147	metal1-clean2	metal1	18.0	MClean1
148	metal1-etch2	metal1	120.0	Strip13
149	metal1-part-meas	metal1	20.0	Surf1
150	imox-sog	Soxide	392.0	Boven15
151	planar-backetch	Soxide	107.0	Strip5
152	imox-sec	Soxide	112.0	Boven17
153	imox-part-meas	Soxide	20.0	Surf1
154	met2-cont-litho	metal2	212.0	Litho2
155	cont-res-oxetch	metal2	355.0	Strip18
156	cont-met-etch	metal2	49.0	Strip14
157	cont-polymer	metal2	26.0	Dep1
158	cont-def-cont	metal2	13.0	Mikro1
159	met2-con-dep	metal2	119.0	Sputter1
160	met2-part-meas	metal2	20.0	Surf1
161	met2-con-dep2	metal2	233.0	Sputter1

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Step #	Description of step	Feature	Time (minutes)	Workstation
162	met2-con-clean	metal2	28.0	MClean1
163	met2-part-meas2	metal2	20.0	Surf1
164	met2-dep	metal2	142.0	Sputter2
165	metal2-litho	metal2	212.0	Litho2
166	metal2-etch	metal2	127.0	Strip14
167	metal2-clean	metal2	18.0	MClean1
168	metal2-etch2	metal2	120.0	Strip14
169	metal2-def-cont	metal2	13.0	Mikro1
170	passv-dep	passv	144.0	Boven19
171	passv-litho	passv	200.0	Litho1
172	passv-res-etch	passv	250.0	Strip5
173	passv-bake	passv	150.0	Boven16
174	end-process	end	57.0	End

Table A.2 0.5 micron 2-metal DRAM Recipe Steps.

Appendix B

Equipment Set

The equipment set in this appendix is described in terms of the workstation that they belong to. Table B.1 describes the basic features of the work-stations and equipment set designed for the two product factory with 2496 wafer starts per week. Note that for each workstation only one piece of equipment is shown; the rest are not shown in this table for clarity. The five fields in this table are:

1. *Equipment Id*: The identifying number of the piece of equipment.
2. *Example equipment name*: This field contains the name of a representative piece of equipment in the work-station.
3. *Work-station name*: The name of the work-station referred to in process recipe steps.
4. *Capacity*: This field gives the equipment capacity in number of wafers. Note that the minimum capacity is given as 24 wafers (lot size).
5. *No. of Equipment*: The number of pieces of equipment in the named work-station.
6. *Capital Cost*: Cost of equipment purchase in 1000's of dollars. The depreciation rate is assumed to be 40% (over a year) for all equipment.
7. *Usage Cost*: Cost of using the equipment for processing normalized to a year of 100% utilization.

Equip-ment Id	Example equipment name	Worksta-tion name	Capac-ity	No. of Equip-ment	Capital Cost in 1000 $	Usage Cost in 1000 $
10101	begin1	Begin	24	1	0	0
10201	scriber1	Scriber1	24	1	180	40
10301	tclean1_1	TClean1	48	5	280	50
10401	tclean2_1	TClean2	48	4	280	50
10501	boven1_1	Boven1	96	4	700	200
10601	boven2_1	Boven2	72	3	450	200
10701	boven3_1	Boven3	96	2	450	200
10801	boven4_1	Boven4	96	1	250	100
10901	boven5_1	Boven5	96	3	550	250
11001	boven6_1	Boven6	96	3	450	200
11101	boven7_1	Boven7	96	1	300	400
11201	boven8_1	Boven8	96	3	450	200
11301	boven9_1	Boven9	96	1	450	200
11401	boven10_1	Boven10	96	3	450	200
11501	boven11_1	Boven11	96	3	400	200
11601	boven12_1	Boven12	96	1	500	250
11701	boven13_1	Boven13	96	2	220	120
11801	implant1_1	Implant1	24	2	5100	700
11901	implant2_1	Implant2	24	4	1800	400
12001	epi1_1	Epi1	24	4	1850	400
12101	strip1_1	Strip1	24	7	3000	700
12201	strip2_1	Strip2	24	5	300	100
12301	strip3_1	Strip3	24	2	100	40
12401	strip4_1	Strip4	24	5	300	100

Table B.1 Equipment set description.

Equip-ment Id	Example equipment name	Worksta-tion name	Capac-ity	No. of Equip-ment	Capital Cost in 1000 $	Usage Cost in 1000 $
12501	strip6_1	Strip6	24	4	570	200
12601	strip7_1	Strip7	24	4	200	80
12701	strip8_1	Strip8	24	10	800	200
12801	strip9_1	Strip9	24	4	200	60
12901	strip10_1	Strip10	24	3	800	300
13001	strip11_1	Strip11	24	4	1500	700
13101	strip15_1	Strip15	24	2	800	400
13201	boven20_1	Boven20	96	2	420	100
13301	strip19_1	Strip19	24	7	3000	700
13401	strip20_1	Strip20	24	5	600	200
20101	boven14_1	Boven14	96	1	500	200
20201	boven15_1	Boven15	48	4	1200	400
20301	boven16_1	Boven16	48	1	100	30
20401	boven17_1	Boven17	24	2	100	30
20501	boven18_1	Boven18	48	1	820	200
20601	boven19_1	Boven19	48	1	800	200
20701	mclean1_1	MClean1	24	2	100	30
20801	sputter1_1	Sputter1	24	7	3800	600
20901	sputter2_1	Sputter2	24	6	2500	500
21001	dep1_1	Dep1	24	5	500	200
21101	strip12_1	Strip12	48	4	1600	400
21201	strip13_1	Strip13	24	3	1100	400
21301	strip14_1	Strip14	24	6	2300	600
21401	strip16_1	Strip16	24	4	800	300
21501	strip17_1	Strip17	24	4	1000	400
21601	end	End	24	2	0	0

Table B.1 Equipment set description.

Equip-ment Id	Example equipment name	Worksta-tion name	Capac-ity	No. of Equip-ment	Capital Cost in 1000 $	Usage Cost in 1000 $
21701	strip18_1	Strip18	48	5	1500	600
30101	litho1_1	Litho1	48	6	3100	1000
30201	litho2_1	Litho2	48	13	5000	1500
30301	surf1_1	Surf1	24	3	1500	300
30401	mikro1_1	Mikro1	24	1	25	8
30501	strip5_1	Strip5	24	19	3000	600
30601	litho3_1	Litho3	48	4	3100	1400
30701	surf2_1	Surf2	24	1	1500	400

Table B.1 Equipment set description.

Appendix C

Product Attributes

In this appendix, yield related attributes of the CMOS and DRAM products used for simulation is presented. First, the critical area for the CMOS design with three values of minimum feature sizes (0.6, 0.5 and 0.4 microns) are presented followed by the 0.4 micron DRAM critical areas.

C.1 CMOS Product

The die size of the CMOS product with 0.6 micron minimum feature size is 3.00 cm^2 and the number of dies per wafer is 50. The critical area for shorts in polysilicon, metal1, metal2 and metal3 are shown in Figure C.1.

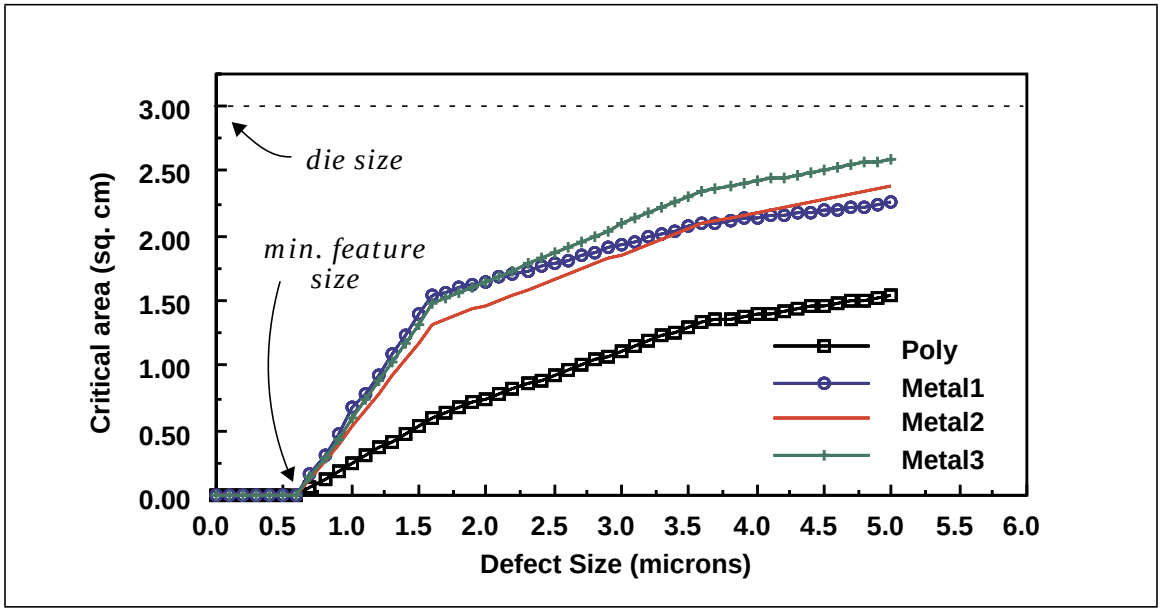

Figure C.1 Critical area vs. defect size for 0.6 micron CMOS design.

The die size of the CMOS product with 0.5 micron minimum feature size is 2.11 cm^2 and the number of dies per wafer is 73. The critical area for shorts in polysilicon, metal1, metal2 and metal3 are shown in Figure C.2.

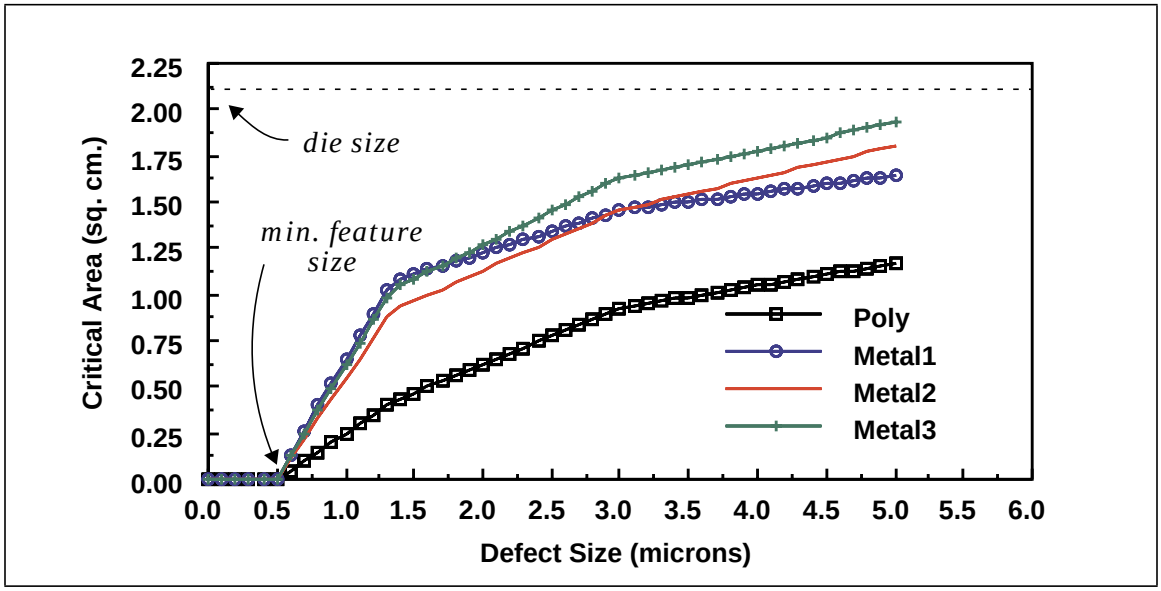

Figure C.2 Critical area vs. defect size for 0.5 micron CMOS design.

The die size of the CMOS product with 0.4 micron minimum feature size is 1.4 cm^2 and the number of dies per wafer is 110. The critical area for shorts in polysilicon, metal1, metal2 and metal3 are shown in Figure C.3.

The scaling of critical area for polysilicon as a result of shrinking from 0.6 micron to 0.4 microns is illustrated in Figure C.4. The scaled critical areas are obtained in the following way. Let the critical area function, f_1, for a minimum feature size, m_1, be given by:

$$f_1 = A_c(R) \tag{C.1}$$

where, R is the defect size. Then the critical area function, f_2, for a minimum feature size of, m_2, is given by:

$$f_2 = s^2 A_c(sR) \tag{C.2}$$

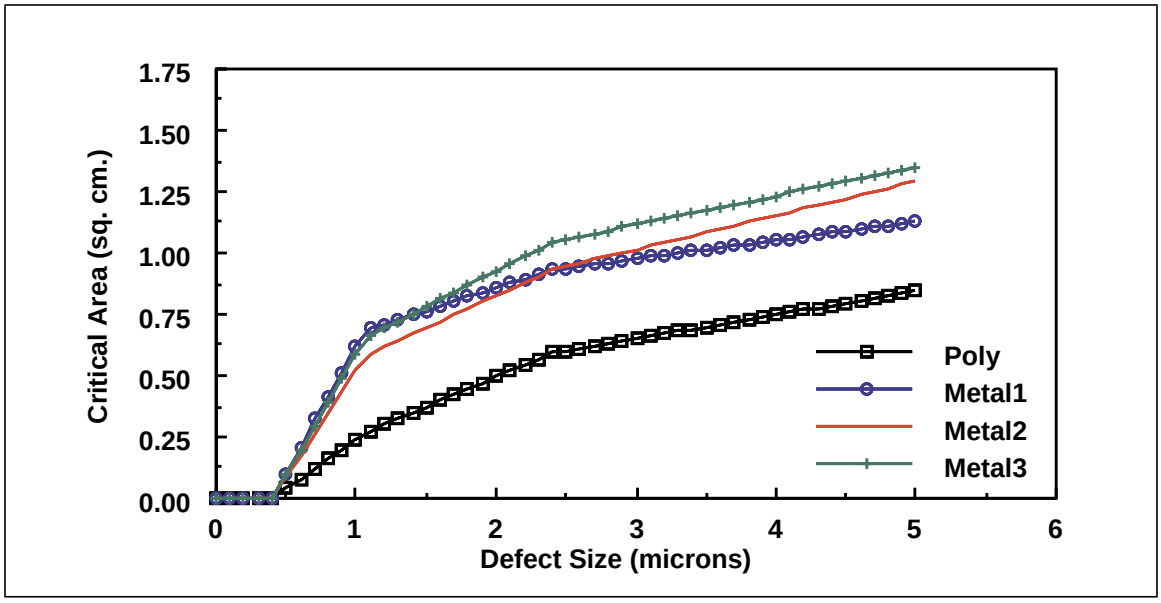

Figure C.3 Critical area vs. defect size for 0.4 micron CMOS design.

where, s is the scaling factor and is equal to m_2/m_1. The critical area for the original product with 0.6 micron minimum feature size is given up to 5 micron defect size. Critical obtained after scaling is thus valid up to a defect size 5s which is less than 5 micron. From 5s to 5 microns defect size, the critical area function is extrapolated linearly.

C.2 DRAM Product

The die size of DRAM product with 0.4 micron minimum feature size is 1.4 cm^2 and the number of dies per wafer is 110. The critical area for shorts in polysilicon, metal1, metal2 and metal3 are shown in Figure C.5. Here, the critical areas are assumed to be comparable for polysilicon, metal1 and metal2 as shown. Metal1 is most sensitive to defects and metal2 is the least sensitive although the difference is noticeable only for small defect sizes.

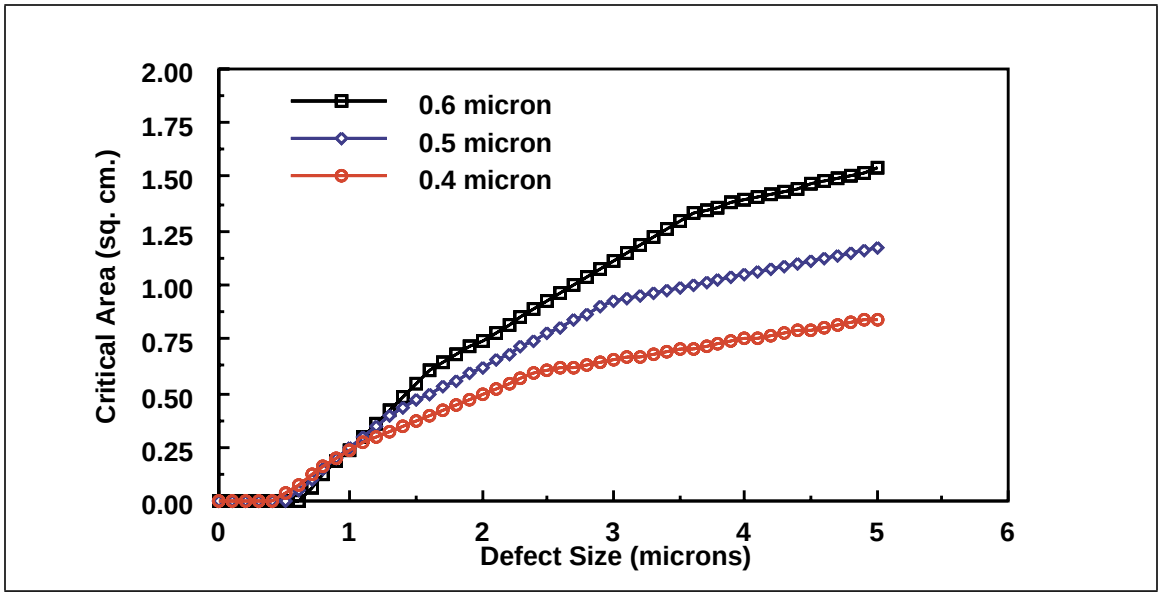

Figure C.4 Critical area scaling for polysilicon shorts for the three designs.

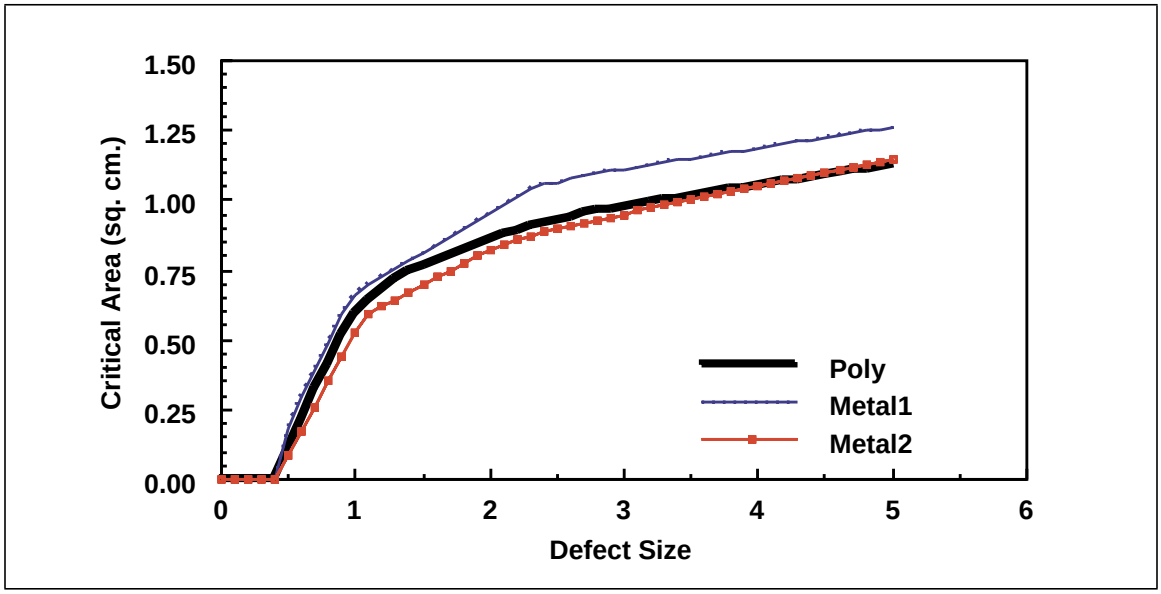

Figure C.5 Critical area vs. defect size for 0.4 micron DRAM design.