The Dissertation Committee for Hyung Joo Lee Certifies that this is the approved version of the following dissertation:

Advanced Process Control and Optimal Sampling in Semiconductor Manufacturing

Committee:

Thomas F. Edgar, Supervisor

S. Joe Qin

Roger T. Bonnecaze

Glenn Y. Masada

Elmira Popova

John D. Stuber

Advanced Process Control and Optimal Sampling in Semiconductor Manufacturing

by

Hyung Joo Lee, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2008

Acknowledgements

First of all, I would like to thank Dr. Thomas F. Edgar, my Ph.D supervisor, who placed me in an excellent research environment where I could obtain valuable experiences. He allowed me the flexibility to pursue my career and educational goals simultaneously, while inspiring and encouraging me with patience throughout the duration of this work. I have worked with him for four years in my graduate education and his unending enthusiasm and dedication to the field of process control never ceases to amaze me.

I also thank my Ph.D. committee members, Dr. Qin, Dr. Bonnecaze, Dr. Masada, Dr. Popova, and Dr. Stuber for their helpful advice, technical support and for the interest and enthusiasm they have shown for my work. Their guidance was invaluable to me during my four years of graduate study at Austin.

I also thank the current graduate students in process control group, as well as the former students of my advisor, with whom I have worked with over the years. They include: Amogh Prabhu, Yang Zhang, Xiaoliang Zhuang, Sidharth Abrol, Dirk Thiele, Daniel B. Weber, Iván Castillo, Sepideh Ziaii, Blake Parkinson, Bhalinder Gill, Doug French, Ben Spivey, John Hedengren, Wonhui Cho, Lina Rueda, David Castiniera, Scott Harrison, Kevin Chamness, Terry Farmer, Sungbo Hwang, Alex Pasadyn, Scott Bushman, Kyehyun Baek and two undergraduates (Jimmy Park and Jamie Kim). I enjoyed working with and learned a lot from each of them.

This work was supported by Texas Instruments (TI) and Tokyo Electron America (TEL). In particular, I would like to thank John Stuber and Chris Harrison at TI for

offering me precious comments and insightful directions. Merritt funk, Dan Prager, Radha Sundararajan, Anita Viswanathan, and Asao Yamashita at TEL have given me a great deal of guidance and support as well.

For their great help with life at Austin, I want to thank Mikey Lin, Saul Lee, Kane Jen, Elizabeth Costner, Brandon Rowe, Jasmine Tam, Mehul Patel, and Danielle Smith. I would also like to thank all of Korean graduate colleagues in chemical engineering department especially, Eunsu Paek, Hyunwook Kwak, Bomi Nam, Byoungjoon Ko, Euikyoon Kim, Kwangseok Lee, and Dochang Lee.

Many people in current and former students of Seoul National University (SNU) have encouraged me in various ways during my graduate study. I am grateful to Prof. Moon, Prof. Rhee, Taekyung Yoo, Tekhyung Lee, Seoungwoo Lee, Youjin Lee, Eunjee Lee, Soohyung Choi, Minsuk Shim, Minsub Jung, Jaehyun Lee, Byoungwoo Yang, Sangwoo Lee, Kobum Kim, Nakwon Choi, Joonhyung Lee, Jonghoon Choi, Junsang Doh, Jujin Ahn, and Jiyeon Lee.

Finally, I would like to thank my family for bearing with me and giving constant emotional support throughout the time I've been involved with this work. In particular, the endless love and support from my parents and sister allowed me to maintain enough sanity to complete this work. I am forever grateful to everyone who has helped me during this time.

Advanced Process Control and Optimal Sampling in Semiconductor Manufacturing

Publication No._____

Hyung Joo Lee, Ph.D.

The University of Texas at Austin, 2008

Supervisor: Thomas F. Edgar

Semiconductor manufacturing is characterized by a dynamic, varying environment and the technology to produce integrated circuits is always shifting in response to the demand for faster and new products, and the time between the development of a new profitable method of manufacturing and its transfer to tangible production is very short. The semiconductor industry has adopted the use of advanced process control (APC), namely a set of automated methodologies to reach desired process goals in operating individual process steps. That is because the ultimate motivation for APC is improved device yield and a typical semiconductor manufacturing process can have several hundred unit processes, any of which could be a yield limiter if a given unit procedure is out of control. APC uses information about the materials to be processed, metrology data, and the desired output results to choose which model and control plan to employ. The current focus of APC for semiconductor manufacturers is run-to-run control. Many metrology applications have become key enablers for the conventionally labeled "value-added" processing steps in lithography and etch and are now integral parts

of these processes. The economic advantage of effective metrology applications increases with the difficulty of the manufacturing process. Frequent measurement facilitates products reaching its target but it increases the cost and cycle time. If lots of measurements are skipped, the product quality does not be guaranteed due to process error from uncompensated drift and step disturbance. Thus, it is necessary to optimize the sampling plan in order to quickly identify the sources of prediction errors and decrease the metrology cost and cycle time. The goal of this research intend to understand the relationship between metrology and advanced process control (APC) in semiconductor manufacturing and develop an enhanced sampling strategy in order to maximize the value of metrology and control for critical wafer features.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 ADVANCED PROCESS CONTROL IN SEMICONDUCTOR MANUFACTURING

Semiconductor manufacturing is characterized by a dynamic, varying environment, compared to conventional manufacturing in the chemical industries [1][2][3]. The technology to produce integrated circuits is always shifting in response to the demand for faster and new products, and the time between the development of a new profitable method of manufacturing and its transfer to tangible production is very short. This requires an industry that should be very flexible to adjust to constant changes in working procedures and challenges to produce new devices.

Thus, advances in process control will be necessary to meet future technical challenges in semiconductor manufacturing. Process control problems in microelectronics process can be divided into four categories [4]; fab management, contamination control, materials handling, and unit operations control. Much attention has been concentrated on coordinating the schedules of different unit operations, controlling the purity of the essential reactants, and monitoring the transport of wafers between equipments. Relatively less effort has been dedicated to improve the control of individual unit process.

However, the semiconductor industry has adopted the use of advanced process control (APC), namely a set of automated methodologies to reach desired process goals in operating individual process steps [5][6][7][8][9]. That is because the ultimate motivation for APC is improved device yield and a typical semiconductor manufacturing

process can have several hundred unit processes, any of which could be a yield limiter if a given unit procedure is out of control. It is difficult to estimate the potential yield for a given lot before the wafers arrive at the end of the production line. Therefore, it is critical that each one of the steps in the manufacturing process be operated as closely as possible to the process requirement for the operation.

For example, a series of process steps that can experience yield loss due to process variation is the sequence [1]: interlayer dielectric polish, photolithography mask and exposure, and contact etch. Polish process specifications on film thickness uniformity are determined by depth of focus requirements of the photolithography step. If the uniformity of the polish operation is not within requirement limits, it is possible that the contact etch process may fail because of a degraded mask and exposure step. If the etch fails to totally remove the interlayer dielectric, the metal interconnects may not give contact between the metal layers of the device, causing a short circuit or failed device. This type of process correlation is played out time and again all over the production line, leading to a need for exceedingly tight process control in each unit operation.

APC uses information about the materials to be processed, metrology data, and the desired output results to choose which model and control plan to employ. The current focus of APC for semiconductor manufacturers is run-to-run control [5][6][7][8]. If the run unit is a lot, it is called lot to lot (L2L) control, while if the run is a wafer, it is called wafer to wafer (W2W) control (Figure 1.1) [10]. Run-to-run control allows feedforward and feedback information between successive processes and is physically compatible with ex situ metrology, which has been the typical industrial practice [1]. Figure 1.2 shows a conceptual diagram of run-to-run control with feedforward and feedback.

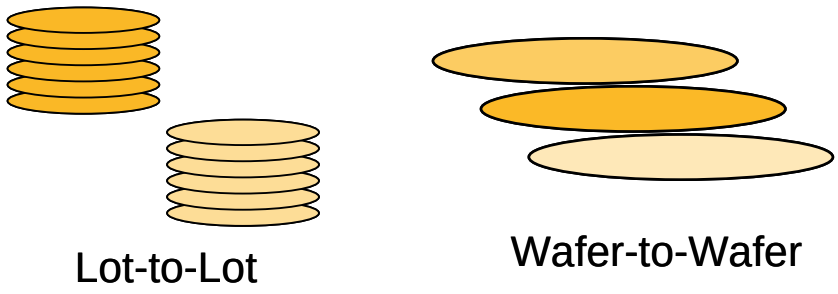


Figure 1.1 Lot-to-lot (L2L) control vs. wafer-to-wafer (W2W) control

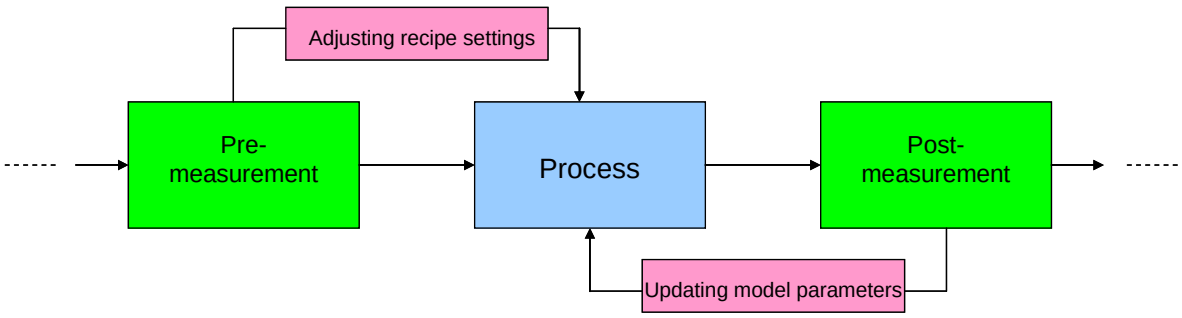


Figure 1.2 Run-to-run control with feedforward and feedback

In feedforward control the controller adjusts process recipe settings using pre-measurement and in feedback control model parameters are updated by the results from post-measurement [11][12].

Figure 1.3 shows run-to-run control applied to CD control [13][14][15]. Run-to-run controllers normally are model-based controllers, coupled with some types of observer. Although there are various ways to design an observer, the most broadly implemented design is based on an exponentially weighted moving average (EWMA) scheme, which can be shown to be equivalent to internal model control (IMC) [16]. The details of EWMA will be shown in section 5.3. The final component of the run-to-run controller is the control law which specifies how the recipe for the unit process should be updated [17].

Although statistical process control (SPC) has been a powerful technique for monitoring reducing variation in semiconductor manufacturing processes, it has certain limitations [3]. The basic assumption on which SPC is based is that the observations collected and plotted on control charts stand for a random sample from a stable probability distribution. However, this assumption does not hold for the dynamic characteristic in semiconductor manufacturing. Run-to-run control is additionally necessitated by the non-stationary nature of most semiconductor processes. While SPC is designed for stationary processes where output variations are independent, run-to-run control is able to compensate for drifting processes where output variations are interrelated. The variation correlation is normally caused by changes in the processing environment.

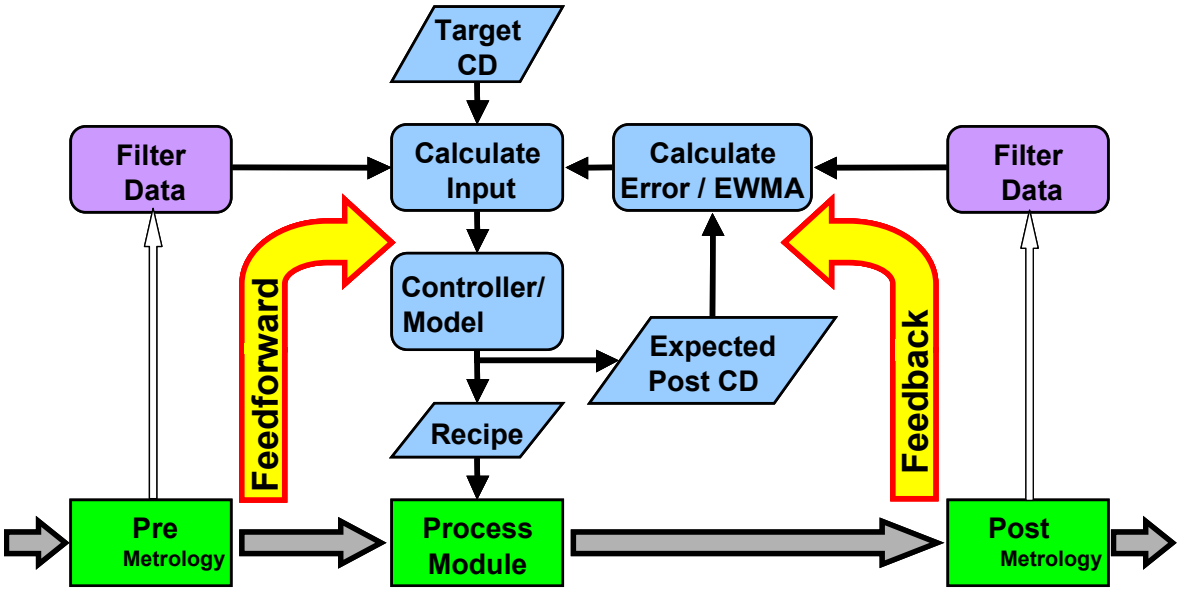


Figure 1.3 Run-to-run control example: CD control

For example, in a deposition process, the reactor walls may become fouled by deposition as many products are processed [1]. This slow drift in the reactor chamber state requires small changes to the batch recipe in order to make sure that the product outputs maintain on target. Eventually, the reactor chamber will be cleaned to remove the wall deposition, effectively causing a step disturbance to the process. As the run-to-run controller compensates for the gradual drifting process, it will also compensate for the step disturbance (abrupt shift) to return the process to target after an environment change.

Another limitation of SPC is that it is typically applied offline [3]. As a result, corrective procedures suggested by SPC alarms normally occur too long after process changes, potentially leading to considerable misprocessing. Run-to-run control is metrology-driven and control actions are triggered by metrology data from pre and post processes. These metrology data are compared to predictions calculated from process models. When model predictions differ significantly from measurements, corrective action is immediately initiated for the next run. The ultimate goal of run-to-run control is that of batch control for a lot or a single wafer. By analyzing the results of previous batches, the run-to-run controller should be able to adjust the batch recipe in order to reduce output variability.

1.2 VALUE-ADDED METROLOGY

Metrology operations are not always considered mandatory since metrology itself does not improve a physical property of product. However, this conventional idea that metrology is a "non-value-added necessary evil" is a misleading statement [18]. Metrology measurements and monitoring have become important for semiconductor manufacturing. Many metrology applications have become key enablers for the

conventionally labeled "value-added" processing steps in lithography and etch and are now integral parts of these processes. Metrology enables operators and engineers to identify problems early on to reduce their impact [3][19]. The economic advantage of effective metrology applications increases with the difficulty of the manufacturing process.

Thus, metrology is so important that a fab's overall performance is a strong function of the quality and quantity of metrology expertise on site [18]. Although metrology suppliers play an essential role in achieving the maximum performance of their equipment, it is frequently the experienced fab metrology engineer who allows a metrology tool's complete potential to be reached. Considering the lots of difficulties facing metrology today, getting this complete potential often means the distinction between a world-class manufacturing fab with leading process controls and yield and a fab struggling to make a profit.

The metrologies in manufacturing line consist of extremely complicated equipment that can be separated to tools characterizing the state of features on the wafers themselves and those that illustrate the status of the equipment operating on those wafers. Figure 1.4 shows a typical CMOS process flow [3]. Inserted into this flow diagram in various places is "M" symbols representing key measurement points. Clearly, CMOS technology involves many unit processes with high complexity and tight tolerances. This necessitates repeated and thorough inline process monitoring to guarantee high-quality final products.

The technology shift from a single development fab to multiple manufacturing fabs or foundries is a universal practice for the industry these days [18]. Technology transfer is much easier if all significant processes are characterized using precise metrology.

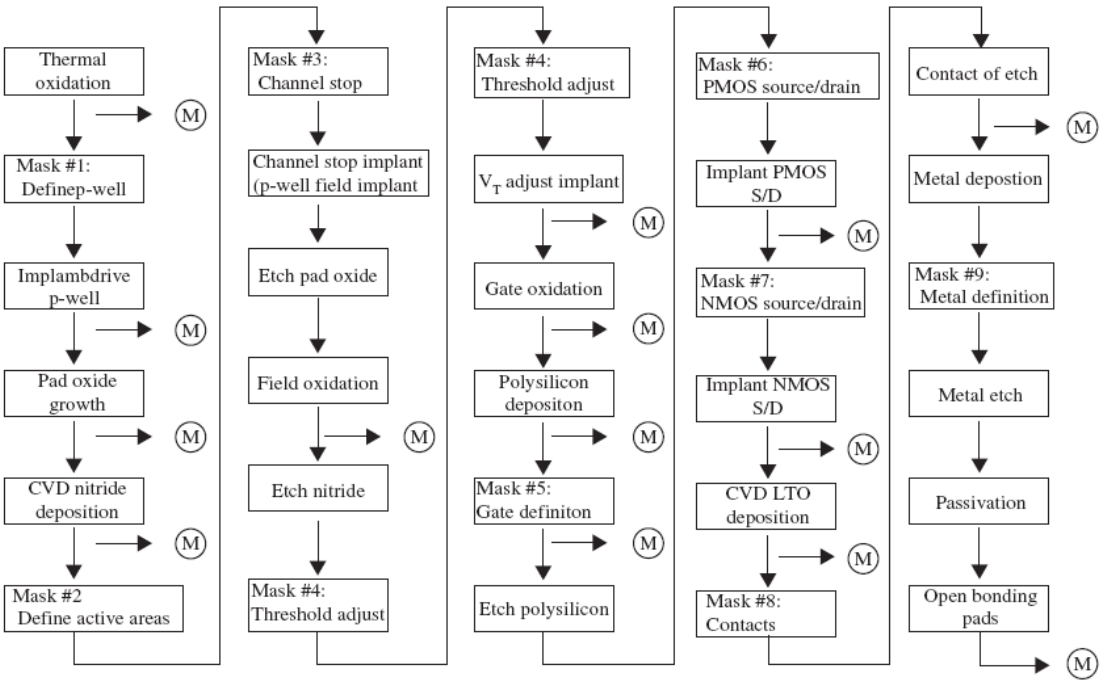


Figure 1.4 CMOS process flow showing key measurement points

Matching of a variety of metrology techniques within a single fab and between different fabs is an important part of a technology development and transfer. Rapid technology ramp-up and sustained production are critical for handling of the extremely high cost of modern technology development.

Eliminating measurement error is metrology's ultimate goal [18]. Measurement error or uncertainty is constantly present, but while it cannot be eliminated totally, it can be minimized. Therefore a major responsibility of the metrology engineering side is to analytically decrease sources of measurement error.

The actions to reduce measurement error contain the selection of equipment set, process development, employ of suitable metrics and techniques for gauging and maintaining equipment performance, measurement optimization, accuracy calibration to the process, and equipment stability and matching optimization. As error sources are minimized, the value of metrology can be maximized.

The measurements required illustrate physical quantities, such as film thickness, uniformity, and feature dimensions; or electrical parameters, such as resistance and capacitance [3]. Among those many parameters, critical dimension (CD) is one of the most significant ones to be controlled.

Current CD metrology equipment trends recognize two major types of online tools for process control—the imaging CD-SEM and optical scatterometry [18]. Each has pros and cons, but is complementary in the advanced process control environment. ISMI studies show that scatterometry tools are useful to around the 22nm node which represents the technology level from international technology roadmap for semiconductors (ITRS) [20] but may have unsolved issues beyond 22nm. This has been experimentally established for the 45nm node and feasible for the 32nm node. CD-SEM

is already applicable for the 45nm node and, in some ways, extends to the 32nm node. It is predicted to be extendible beyond these nodes.

The two fundamental criteria for successful CD metrology are precision and accuracy. Precision is a measurement of the uncertainty, or error bar of a measurement, and accuracy is the absolute targeting of a measurement, describing how well metrology trails the reality of a process as defined by a reference (currently CD-AFM). Several significant trends in CD-SEM and scatterometry have lately emerged that are escalating the value of metrology. Especially, scatterometry with smaller size and fast measuring capability is directly integrated into process equipment to make real time APC possible.

1.3 VARIOUS SAMPLING STRATEGIES

As integrated circuits are being produced with continuously shrinking dimension, accurate metrology along with feedback control have become more critical to producing parts with tight distributions of electrical parameters while maintaining high equipment availability. However, once a technology's yield has matured, the focus can quickly shift to cost reduction, which can often lead to metrology reduction or elimination.

Frequent measurement facilitates products reaching its target but it increases the cost and cycle time. If lots of measurements are skipped, the product quality does not be guaranteed due to process error from uncompensated drift and step disturbance. Thus, it is necessary to optimize the sampling plan in order to quickly identify the sources of prediction errors and decrease the metrology cost and cycle time.

As economic objective function can be developed as a decision metric that combines the cost of sampling with performance of the controller in terms of yield and cycle times. The details of economic objective function will be shown in section 2.2. By thoughtful selection of process measurements using this objective function, the fab-

wide sampling or unit process sampling strategies maximize the amount of information that is shared across different batches by capturing tool and product or combined characteristics in common parameters. An automatic algorithm is created which treats sampling plans as operating decisions so that their effect on other variables of interest can be identified.

Uniform rate of sampling is the easiest and most common algorithm used in industry currently. It measures wafer every N runs. The better performance (less mean square error (*MSE*) or higher process capability (*Cpk*)) can be expected when more runs are measured but this costs more (Figure 1.5). Even though this sampling strategy is not new and has a lot of problems, it is an important method to use since it is a basis of every advanced sampling strategy.

Multi-rate sampling or dynamic sampling matches the sampling rate to the process error in prior measurement. The basic concept is to increase sampling frequency when the process is more likely to go out of control and to decrease the frequency when the process is in control again during steady-state production. More details about dynamic sampling methodologies will be explained in chapter 2.

Random sampling chooses measurement runs randomly. According to simple simulations, it shows better performance than uniform rate of sampling. Although almost the same total number of measurement, random sampling shows better performance (much less mean square error) as shown in Figure 1.6. Figure 1.7 shows the results of Monte Carlo simulation of 100 runs to analyze random sampling case by case. To compare the results to uniform sampling with every three runs, the total number of measurements should be similar because the next measurement step varies randomly from one to five units of delay from the previous measurement step.

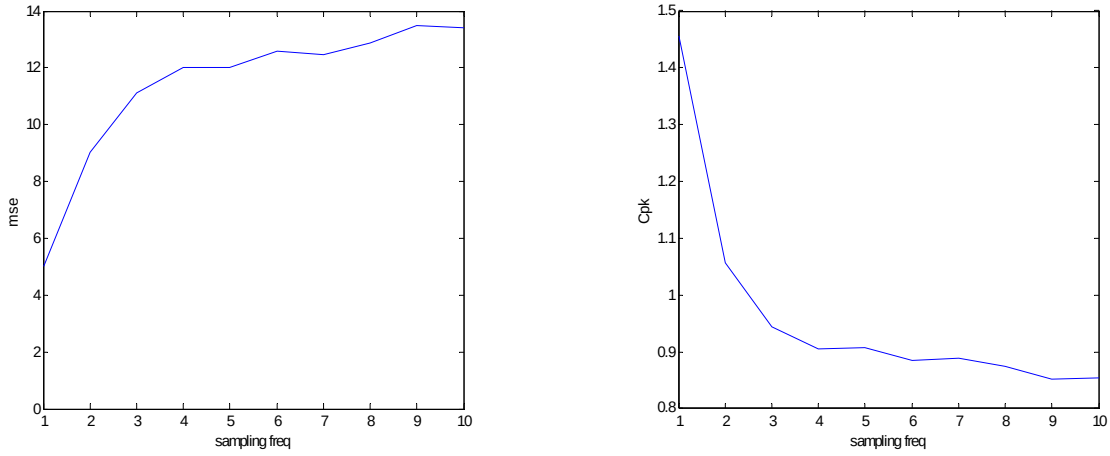


Figure 1.5 Simulation results of uniform rate of sampling

The average mean square error of Monte Carlo simulation is better than uniform rate of sampling as shown in Figure 1.7. That is because a number of consecutive measurements are critical for estimating subsequent runs. It can occur when the next measurement run is randomly chosen with a delay one in random sampling and supposed to reduce the mean square error of the process. The random sampling would not be applicable to real manufacturing because of its arbitrary nature but the idea of consecutive measurements will be addressed in dynamic sampling in chapter 2.

Hybrid sampling employs dynamic sampling (multi-rate) and random sampling. It uses uniform rate sampling (every run or every two runs) during transition periods and uses random sampling during stable periods. Though it shows the better performance than dynamic sampling in some cases, hybrid sampling also would not be applicable to real manufacturing for the same reason as random sampling.

1.4 RESEARCH OBJECTIVES AND DISSERTATION SUMMARY

The goal of this research intend to understand the relationship between metrology and advanced process control (APC) in semiconductor manufacturing and develop a noble sampling strategy in order to maximize the value of metrology and a new control methodology for critical wafer features to keep in control.

If the product or process is not so critical to measure every single wafer, lot-level control is enough to maintain the quality. Furthermore, if it is stable and well-known, not every lot has to be measured. Dynamic sampling can be appropriate for this case to deal with metrology information effectively in order not to sacrifice the quality of product due to infrequent measurement. Chapter 2 explains the best ways to build dynamic sampling method.

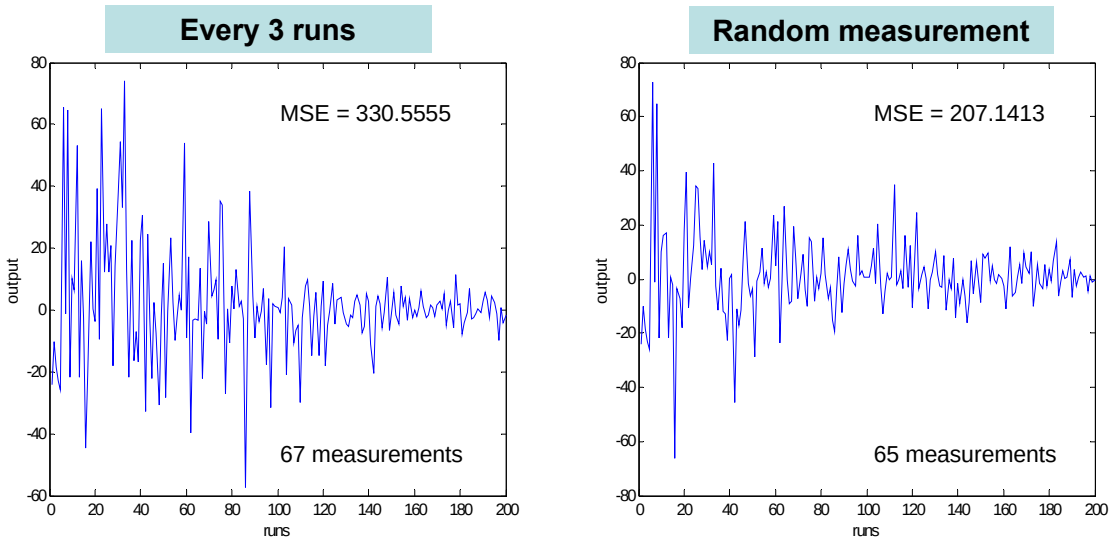


Figure 1.6 Simulation result – uniform sampling (every 3 runs) vs. random sampling

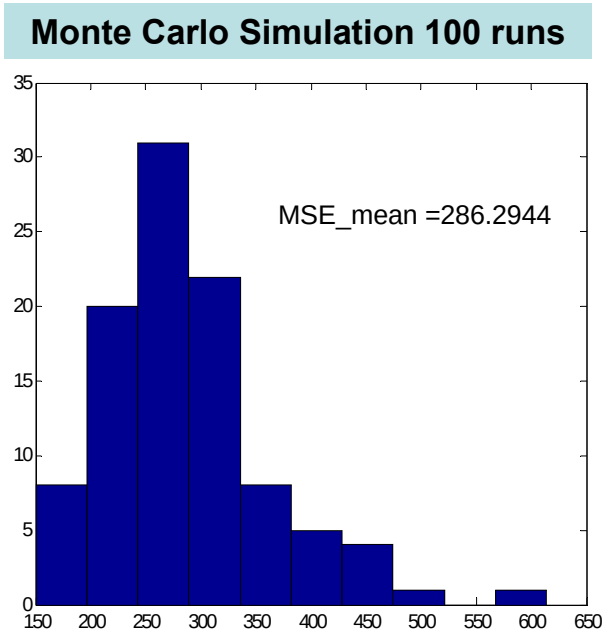


Figure 1.7 Monte Carlo simulation of random sampling

In semiconductor manufacturing it is becoming more common to make many different products by employing a variety of process sequences utilizing the same pieces of equipment, mainly because of the high capital costs associated with the tools and the limited capacity of the facility. In chapter 3, an optimal measurement sampling strategy under multiple product and process environments ("high-mix manufacturing") will be developed with the extension of dynamic sampling idea with qualification runs.

If 100% lot sampling is available, the measurement information should connect to immediate control actions in order to increase the value of metrology data. Using modified Bayesian exponentially weighted moving average (B-EWMA) and robust drift cancellation (RDC), advanced feedback control methodology will be developed in order to compensate both gradual drift and abrupt shift frequently occurs in semiconductor manufacturing in chapter 4.

For the critical process such as polysilicon gate etch, wafer level control is required to meet tight control specification as technical nodes shrink. Chapter 5 shows the advantage of wafer-to-wafer (W2W) control over lot-to-lot (L2L) control in terms of control performance. The role of integrated metrology (IM) in W2W is also evaluated in this chapter.

For better physical and electrical CD control, maintaining wafer uniformity and profile becomes critical. In order to control them simultaneously, multivariable control should be required rather than single variable control. In chapter 6, a multi-input multi-output (MIMO) control approach is evaluated. A commercial polysilicon gate etch process is used to demonstrate the effect of MIMO control.

The ultimate goal of this research is to better understand how to effectively use the data from metrology in various semiconductor manufacturing processes for process control. Understanding the relationship between metrology and APC for effective

control performance allows better development and integration with overall industry needs.

CHAPTER 2

Dynamic Sampling

2.1 BACKGROUND AND MOTIVATION

As integrated circuits are being produced with continuously shrinking dimensions, accurate metrology along with feedback control have become more critical to producing parts with tight distributions of electrical parameters while maintaining high equipment availability. However, once a technology's yield has matured, the focus can quickly shift to cost reduction, which can often lead to metrology reduction or elimination [21][22][23][24]. In order to balance the cost of taking a measurement against the cost of running a process off target, we propose an optimal sampling plan [23][25][26][27][28][29][30] can be developed where the net profit is maximized as shown in section 2.2.

Many wafer fabrication steps have unpredictable drifts and shifts, and thus there is a significant chance that valuable metrology data, i.e., points that would drive a process back to target via feedback, will be skipped if the sampling frequency remains fixed. In order to maintain product quality sampling strategies should be dynamic and adjusted as the process progresses by sampling more frequently when the process has moved away from target [22]. This enables a feedback controller to drive it back to target faster.

Dynamic sampling is a method of changing the sampling frequency based on prior observations [31][32][33][34][35][36]. The basic concept is to increase sampling frequency when the process is more likely to go out of control and to decrease the frequency when the process is in control again during steady-state production. Dynamic

sampling can be utilized to ensure that the risk to production is minimized and that the disturbances (drifts and shifts) can be detected and fixed with minimal impact and cost.

2.2 PROCESS PROFIT FUNCTION

The semiconductor manufacturing process is considered to be a complicated, extremely cost-intensive, and fast paced changing fabrication. There has been much previous work done both by industry and university labs to investigate the profitability of fab in terms of the usage of new tools and yield improvement. However, the economical model describing fab performance, such as yield, throughput, cycle time, and optimizing profitability based on present tools has been rare. We developed a new objective function for a unit process based off-line sampling plan on the metrics of cost of ownership (CoO).

Cost of ownership (CoO) was first developed by SEMATECH in 1990 for decisions during equipment purchase. CoO considers not only the equipment price with its lifetime, but also operation costs, throughput, utilization, and yield loss of the equipment [37]. Semiconductor Equipment and Materials International (SEMI) has since developed CoO publishing ensuing guidelines [38], [39], and [40]. These standards define not only the metrics of the CoO model, but also overall equipment effectiveness (OEE) and overall equipment efficiency to emphasize utilization of the equipment. Dance developed the CoO model for the inspection tools, emphasizing the separation of yield loss accompanied by inspection tools from that of processing tools prior to the inspection [41]. Dance also proposed other considerations that may occur in the inspection tools, such as the cost of discarding a good device, the cost of shipping a bad device [42], and the risks of inventory wafers queuing in front of the inspection tools [43].

According to the SEMI standards, E-35, the CoO model calculates the cost of processing one good wafer in a certain tool of interest through the following equation:

$$CoO = \frac{CF + CR + CY}{TPT \times Y \times U} \tag{2.1}$$

where, *CF*, *CV*, and *CY* are annualized fixed costs [$/yr], annualized recurring costs [$/yr], and annualized cost of yield loss [$/yr], respectively. *TPT*, *Y*, and *U* denote tool throughput [wafers/yr], composite yield [dimensionless], and utilization [dimensionless].

For our sampling plan, we combined metrology into an integrated tool, described in Figure 2.1, that is subject to the CoO model. Also, we omitted the costs that are not directly dependent on the sampling interval, such as fixed costs, utilities, maintenance, and labor, obtaining equation (2.2). Equation (2.2) describes the costs that are directly dependent on the sampling interval and occurring in the integrated tool in Figure 2.1.

$$SamplingCoO = \frac{MC + TL}{TPT \times Y} \tag{2.2}$$

where, *MC* and *TL* are metrology cost, which are the cost of test lots [$/yr] and Taguchi yield loss [$/yr] (Figure 2.2) [44], respectively, while *TPT* and *Y* are throughput and yield, respectively. The denominator, consequently, represents the number of production lots.

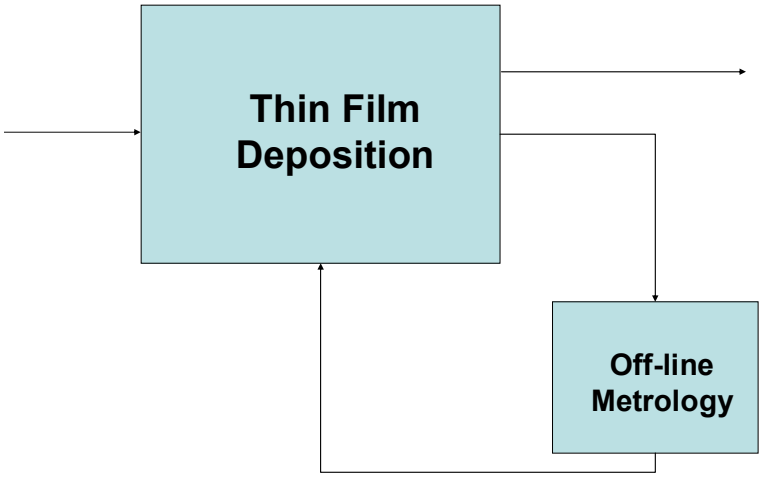


Figure 2.1 Thin film deposition with off-line metrology

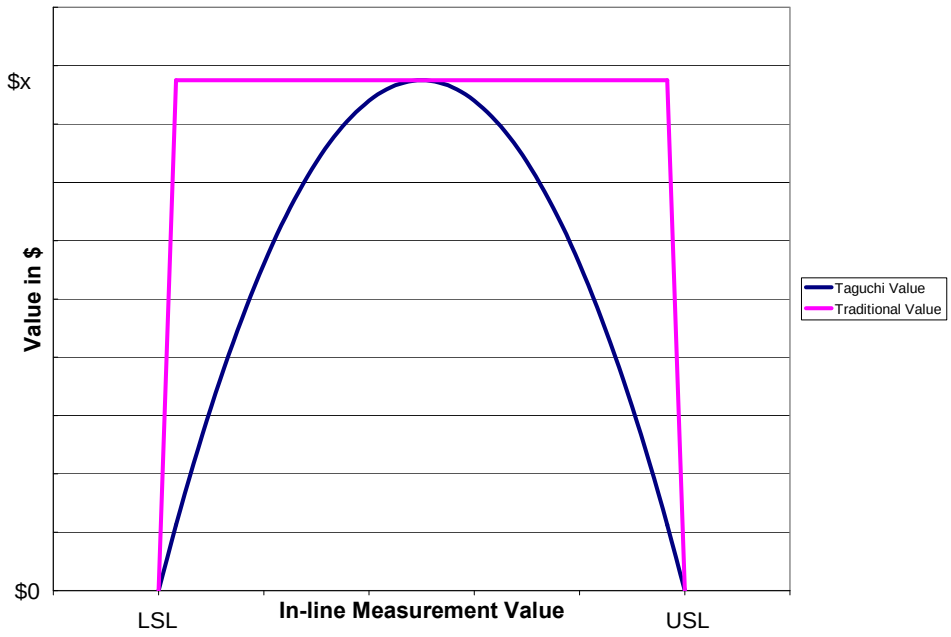


Figure 2.2 Taguchi yield loss function

However, minimizing this sampling CoO does not ensure maximizing the revenue. The revenue also depends on the number of production lots over a specific time period, which is also a function of the sampling interval. The CoO model intrinsically has an effect of good wafers in the denominator calculating the costs needed to make one good wafer, however, this model does not tell how much revenue we can get as the number of productive lots changes. Therefore, it is necessary for the objective function of our sampling plan to include the revenue of selling productive lots.

The modified CoO model had been changed to another form to include the effect of productive lots as follows:

$$Revenue\ per\ lot = Price\ per\ lot - \frac{MC + TL}{TPT \times Y} \tag{2.3}$$

The number of productive lots must be multiplied to revenue per lot, resulting in equation (2.4):

$$\begin{aligned} &(Revenue\ per\ lot) \times (\#\ of\ productive\ lots) \\ &= [(Price\ per\ lot) - \frac{MC + TL}{TPT \times Y}](\#\ of\ productive\ lots) \end{aligned} \tag{2.4}$$

The number of productive lots are equivalent to $TPT \times Y$:

$$\begin{aligned} &(Revenue\ per\ lot) \times (\#\ of\ productive\ lots) \\ &= (Price\ per\ lot) \times (\#\ of\ productive\ lots) - (MC + TL) \end{aligned} \tag{2.5}$$

Consequently, we get the objective function, equation (2.6).

$$NP = PP - (MC + TL) \tag{2.6}$$

where *NP* is net profit and *PP* is production profit calculated by multiplying the number of productive lots and the price of productive lots of deposition cluster, while *MC* and *TL* denote metrology cost and Taguchi yield loss, respectively. The difference between equation (2.2) and (2.6) is the usage of the number of salable lots. Equation (2.2) is a cost function that has the number of good lots in the denominator and has units of [cost $ / one good lot], while equation (2.6), a profit function, multiplies it by the price of productive lots, generating expected net profit from the process of our interest. Maximizing net profit by using equation (2.6) can be a more realistic objective than minimizing the cost of equation (2.2).

The schematic description of calculating this objective function is shown in Figure 2.3. Figure 2.4 shows the effect of sampling frequency on profit using this process profit function (economic optimal sampling [25]).

The basic assumptions for this calculation are as follows:

- There is no contamination yield loss by metrology tool because of the usage of test lots (productive lots are not inspected).
- Test lots are not recycled.
- The salable lots have the same price even though they have different quality.
- The time for maintenance, standby, and engineering is already considered in *TPT*.

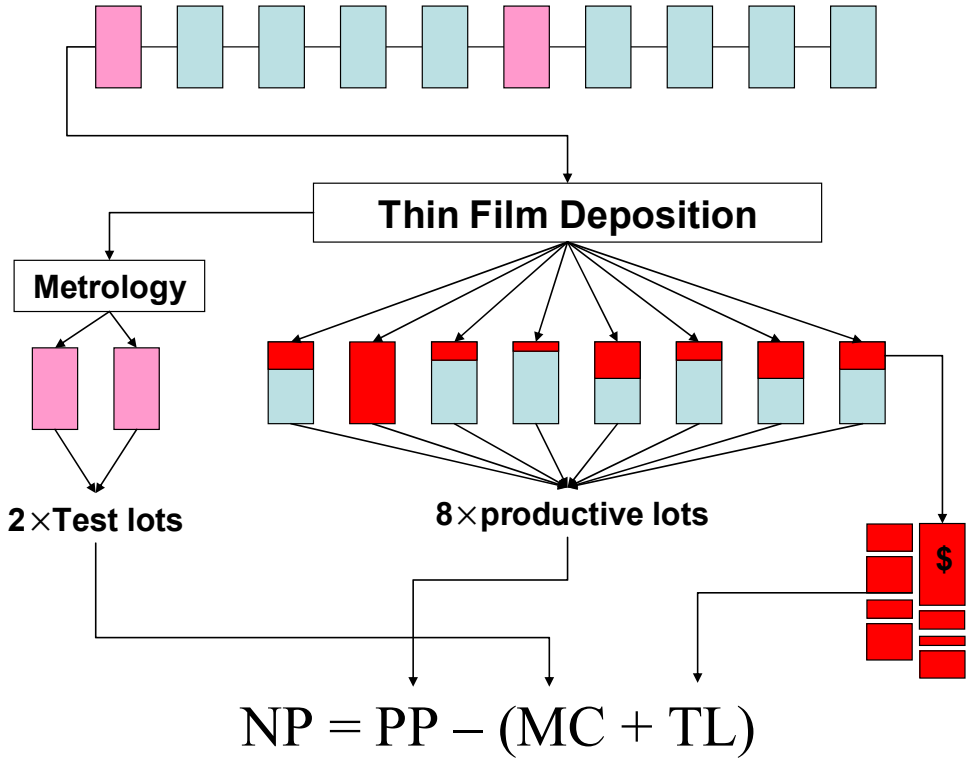


Figure 2.3 Calculation of process profit function for deposition process

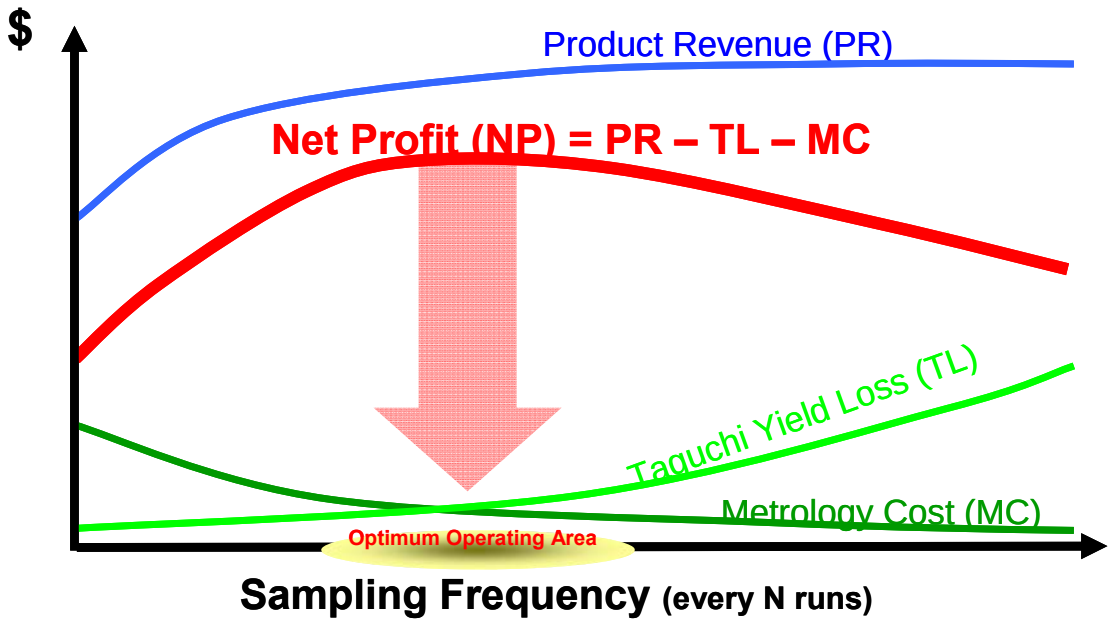


Figure 2.4 Economic optimal sampling

2.3 SAMPLING RATE SELECTION METHODOLOGY

The optimal sampling frequency can be found by using process profit function as an objective function. Figure 2.5 shows the simulation results as a profit chart like Figure 2.4 with the sampling frequency varied from two to ten. Selecting appropriate economic values such as production profit per lot, metrology cost per lot, and upper and lower control limit for calculating a loss function is required to obtain the right answer. Note that this is a sharp optimum instead of a flat optimum.

This method is easy to implement and applicable to a stabilized process using historical data. Since it uses economic values, it can find the contribution of optimal sampling frequency for cost reduction of process. However, it is not useful for rapidly changing process with significant process dynamics and sensitivity to noise. Also, it always requires accurate economic values and historical data. In other words, this selection method can be done offline. Thus, selecting optimal frequency by process profit function will be used to set a baseline sampling rate in online sampling selection algorithms.

The ideas of dynamic sampling are from a selecting sampling schedule proposed by Pasadyn et al. [45]. They extracted an uncontrolled signal from an historical data set and took an offline simulation with that uncontrolled signal. By defining a performance index *J(P)* in terms of a metric for determining the apparent value of a given data set (e.g., the trace of state error covariance matrix) and adding several constraints (e.g., the number of samples) in order to prevent unrealistic solutions, enumeration was used to evaluate all possibilities in order to find sampling schedule that optimizes *J(P)*.

Figure 2.6 shows the simulation result of simple tracking set point change. The rerformance index *J(P)* is used in this simulation to minimize mean square error of output (MSE). The number of runs sampled is fixed to 20 (20% sampling of 100 runs).

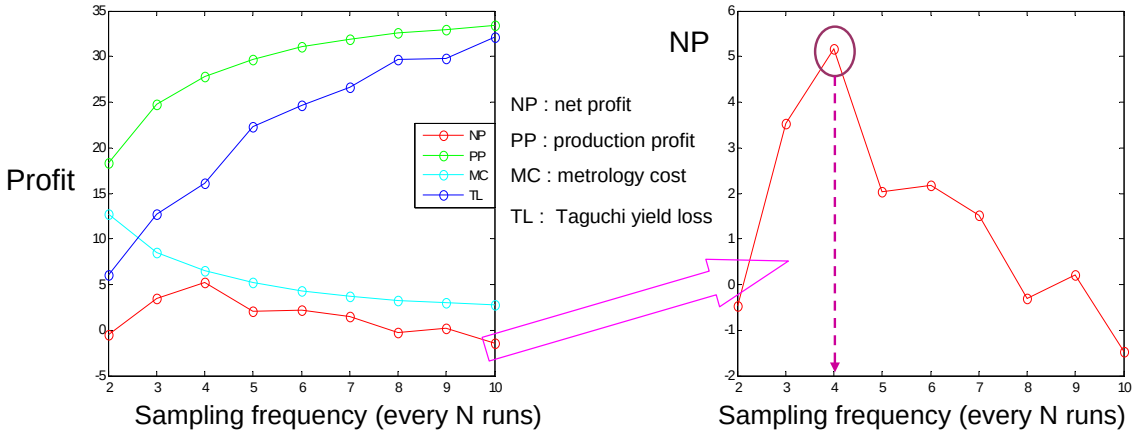


Figure 2.5 Selecting optimal frequency using process profit function

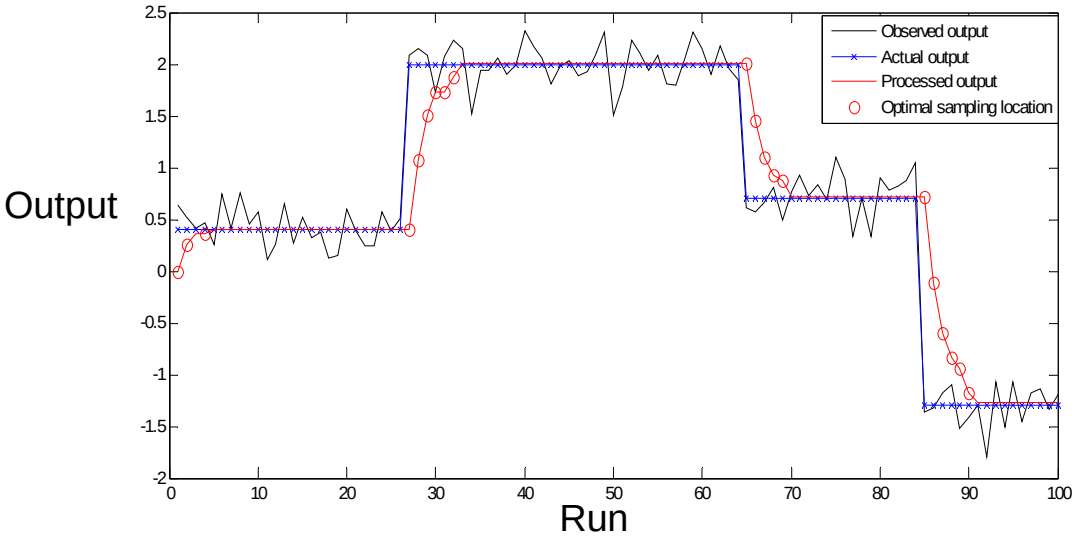


Figure 2.6 Simulation result of optimized sampling

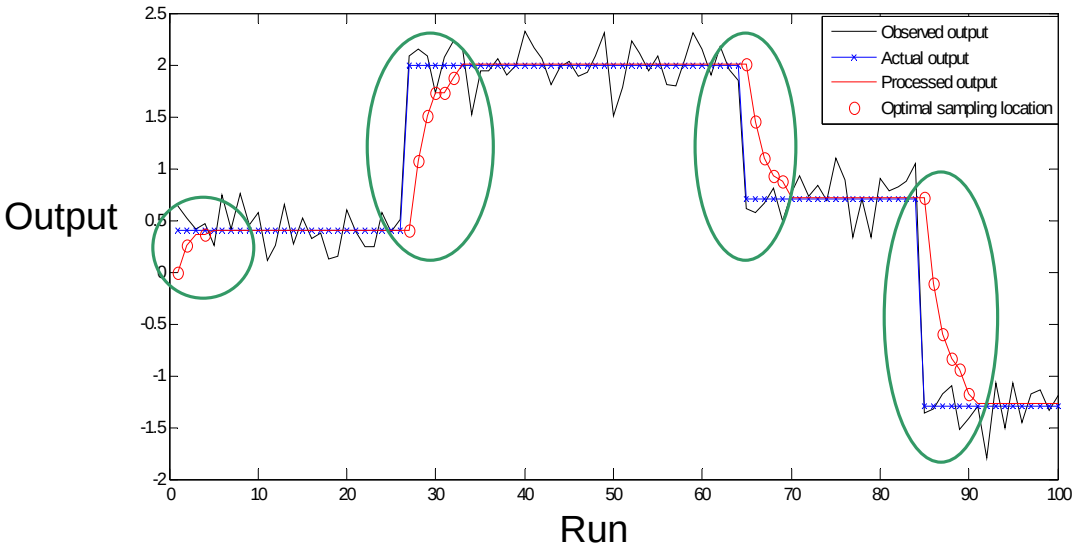


Figure 2.7 Selecting sampling schedule

As shown in Figure 2.7, sampling is more frequent during transition periods of disturbance to minimize the performance index *J(P)*. Therefore, dynamic adaptive sampling where the sampling rate (frequency) is adjusted according to process dynamics can be considered. The sampling rate increases as process error increases and decreases as process error decreases.

Dynamic sampling is expected to be applicable to any processes online since it is flexible and adjustable to process dynamics. Thus, it would have better economical performance than a uniform rate of sampling (higher profit). But several process parameters should be pre-determined for the best performance. As mentioned in section 2.2, the baseline sampling rate can be determined by offline sampling rate selection using process profit function with a historical data set. Also, the magnitude of error to switch sampling rate should be selected from the ideas of Bayesian disturbance detection which will be explained in the following section.

2.3.1 Bayesian Detection Approach

Bayesian detection was first proposed by Wang and He [46][47] to improve the state estimation performance. The Bayesian theorem [48][49][50] is the basic principle in Bayesian detection. Suppose we are interested in estimating the values of a set of parameters Θ for some data set $\boldsymbol{D}$ in some underlying model of the data. For any given model, one can write down an expression for the likelihood function $P(\boldsymbol{D} \mid \Theta)$ of obtaining the data vector $\boldsymbol{D}$ given a particular set of values for the parameter, Θ. In addition to the likelihood function, one may impose a prior distribution $P(\Theta)$ on the parameters, which represents our state of knowledge regarding the values of the

parameters before analyzing the data ***D***. By Bayes' theorem, the posterior probability is then calculated as:

$$P(\Theta \mid D) = \frac{P(D \mid \Theta)P(\Theta)}{P(D)} \tag{2.7}$$

which gives the posterior distribution $P(\Theta \mid \boldsymbol{D})$ in terms of the likelihood, the prior and the evidence $P(\boldsymbol{D})$ [51].

Wang and He proposed predefined data windows for detecting various types of disturbance [46]. The term "pre-change data" denotes the samples prior to the pre-assumed disturbance onset and "post-change data" denotes the samples after and including the pre-assumed disturbance onset. A directly measured observation which has just gone beyond a certain threshold is defined as onset location. After any onset location is determined, pre- and post-change windows are built. A pre-change window is a vector consisting of the past several data points prior to the onset location. A post-change window consists of several data points, one at the onset location and some points right after the onset location. Both pre- and post-change windows keep updating in a simple first-in/last-out (FILO) stack fashion by moving these windows as process runs.

For a step disturbance, the posterior probability is generated by computing the joint posterior probability for each subset of the post-change window, X_k, where $X_k \equiv \{x_1, x_2, ..., x_k\}$ for $k = 1, 2, \ldots, l_{w_2}$. Assuming that the mean of samples in the pre-change window is zero, the step magnitude, μ_D, is calculated as the mean of X_k by the following equation (2.8).

$$\mu_D = \frac{\sum_{i=1}^{k} x_i}{k} \tag{2.8}$$

If Gaussian distribution is assumed, the probability density function for normal and shifted process states are denoted by $N(0,\ \sigma^2)$ and $N(\mu_D, \sigma^2)$, respectively, where σ is the process standard deviation. The likelihood function of a step disturbance for a single observation, x_i, is calculated by the equation (2.9):

$$p(x_i \mid \Theta_D) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x_i - \mu_D)^2}{2\sigma^2}] \tag{2.9}$$

Assuming all samples in X_k are independent and identically distributed, the likelihood function for X_k is

$$p(X_k \mid \Theta_D) = \prod_{i=1}^{k} p(x_i \mid \Theta_D) \ . \tag{2.10}$$

By substituting the equation (2.9) into (2.10), the function is derived as

$$p(X_k \mid \Theta_D) = \frac{1}{(\sqrt{2\pi\sigma^2})^k} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} (x_i - \mu_D)^2] \ . \tag{2.11}$$

Similarly, when the normal process data is assumed, the likelihood function is,

$$p(X_k \mid \Theta_D) = \frac{1}{(\sqrt{2\pi\sigma^2})^k} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} x_i^{\ 2}] \ . \tag{2.12}$$

Combining all the equations mentioned above, the joint posterior probability, $P(\Theta_D \mid X_k)$, in equation (2.13), can be obtained as the following equation (2.14):

$$P(\Theta_D \mid X_k) = \frac{P(X_k \mid \Theta_D)P(\Theta_D)}{P(D)} = \frac{P(X_k \mid \Theta_D)P(\Theta_D)}{P(\Theta_D)P(X_k \mid \Theta_D) + (1 - P(\Theta_D))P(X_k \mid \Theta_N)} \tag{2.13}$$

$$p(\Theta_D \mid X_k) = \frac{P_0}{P_0 + (1 - P_0)\exp[-\frac{(\sum_{i=1}^{k} x_i)^2}{2k\sigma^2}]} \tag{2.14}$$

If the disturbance observed is a step disturbance, the posterior probability of the first point in the post-change window, $P(\Theta_D \mid x_1)$ is

$$p(\Theta_D \mid x_1) = \frac{P_0}{P_0 + (1 - P_0)\exp[-\frac{1}{2}(\frac{x_1}{\sigma})^2]} \quad . \tag{2.15}$$

In order to detect step disturbances, the posterior probability, $P(\Theta_D \mid X_k)$, derived from the post-change window is required to be greater than a confidence level *C*. The prior probability, $P(\Theta_D)$, and process standard deviation σ, can be obtained from historical data.

$$P(\Theta_D \mid x_1) \geq C \tag{2.16}$$

A Bayesian detection approach for selecting sampling frequency is applied using this idea.

$$\frac{P_0}{P_0 + (1 - P_0)\exp[-\frac{1}{2}(\frac{x_1}{\sigma})^2]} \geq C \tag{2.17}$$

where P_0 is the probability of out of control (OOC), C is the confidence level, x is the deviation from target, and σ is the process standard deviation. A dynamic sampling frequency based on (2.17) is shown in Figure 2.8. According to the value of the posterior probability, $P(\Theta_D \mid X_k)$, sampling frequency varies from one to three.

2.3.2 Offline Analysis using Statistical Process Control Chart

As mentioned in the previous section, the prior probability, P_0, and the process standard deviation, σ, should be predetermined from the historical data set. Appropriate values of P_0 and σ are required for better performance. The statistical process control (SPC) chart with open-loop data gives suitable information to estimate the prior probability and the process standard deviation [3][31].

Figure 2.9 shows the prior probability, P_0 (out of control probability) estimation. Since almost every fab has SPC charts of any processes for monitoring, it is not difficult to obtain them. P_0 is simply selected from the probability out of the upper and lower spec limit. If it is desired to tighter control, the control spec limit can be adjusted. Usually, the goal of process capability (*Cpk*) is known to be 1.33.

$$\frac{|CSL - \mu|}{3\sigma} = 1.33 \tag{2.18}$$

Because the target is zero in this example, the mean should be close to zero,

$$\frac{|CSL - 0|}{3\sigma} = 1.33 \tag{2.19}$$

Equation (2.19) is simply expressed

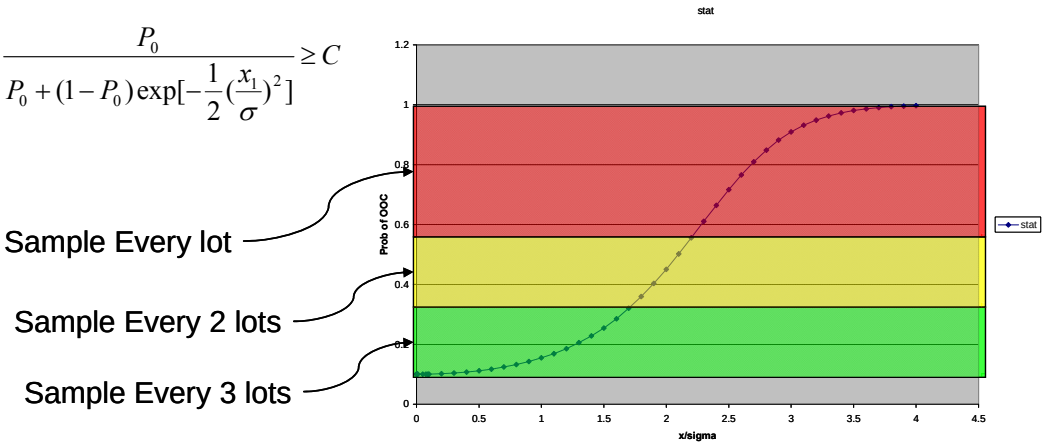
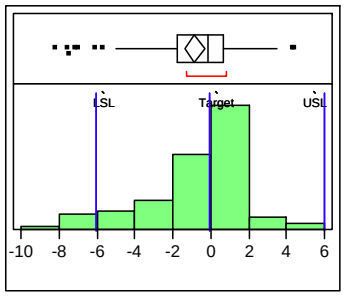
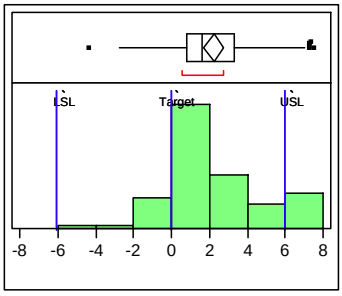


Figure 2.8 Sampling frequency vs. the probability of an out of control (OOC) process.



Specification	Value	Portion	% Actual
Lower Spec Limit	-6	Below LSL	6.8182
Upper Spec Limit	6	Above USL	0.0000
Spec Target	0	Total Outside	6.8182

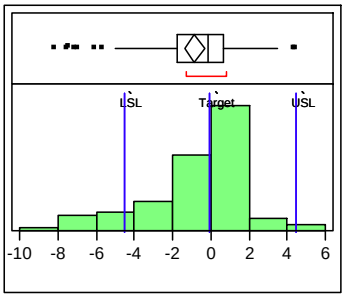
P0 = 0.068



Specification	Value	Portion	% Actual
Lower Spec Limit	-6	Below LSL	0.0000
Upper Spec Limit	6	Above USL	12.6437
Spec Target	0	Total Outside	12.6437

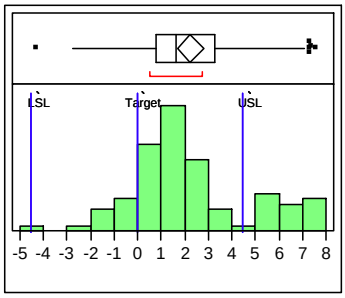
P0 = 0.126

Figure 2.9 The prior probability, P_0 (out of control probability) estimation



Specification	Value	Portion	% Actual
Lower Spec Limit	-4.5	Below LSL	11.3636
Upper Spec Limit	4.5	Above USL	0.0000
Spec Target	0	Total Outside	11.3636

P0 = 0.114



Specification	Value	Portion	% Actual
Lower Spec Limit	-4.5	Below LSL	0.0000
Upper Spec Limit	4.5	Above USL	21.8391
Spec Target	0	Total Outside	21.8391

P0 = 0.218

Figure 2.10 The prior probability, P_0 (out of control probability) estimation for tighter control

$$CSL \cong 4\sigma \tag{2.20}$$

Therefore 3σ will be

$$3\sigma = \frac{3}{4} CSL \tag{2.21}$$

Since we set the control spec limit at ± 6, the new control spec limit for tighter control would be ± 4.5. Figure 2.10 shows the prior probability estimation for tighter control.

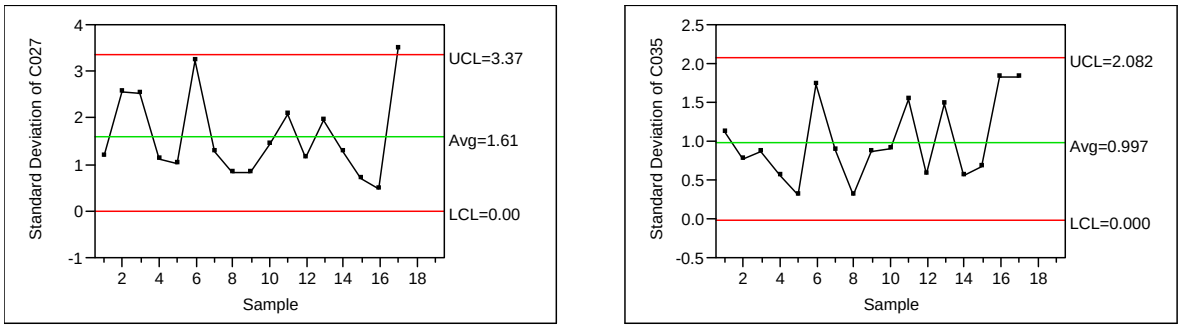
Figure 2.11 shows the process standard deviation, σ, estimation. Since there are drift or step disturbances occurring in the historical open-loop data set, an appropriate estimation interval would be important.

If the process is quite mature, the information from the SPC chart with historical data would be useful to determine the parameters of dynamic sampling. However, since semiconductor manufacturing environment is very dynamic, even if the process was stable, it changes over time due to a variety of reasons. Therefore, online updating should be considered in order not to degrade the dynamic sampling performance. The next section will address this question.

2.3.3 Online Update with Variance Estimation

The process standard deviation, σ, can be updated online simply using prior measurement with the estimation equation (2.22) for standard deviation [3][52][53].

$$s_n^{\,2} = \frac{n-2}{n-1} s_{n-1}^{\,2} + \frac{(y - TGT)^2}{n-1} \tag{2.22}$$



σ = 1.61 σ = 1.00

Figure 2.11 The process standard deviation, σ, estimation

where *TGT* is the target and s_n is estimate of the process standard deviation of *n* runs.

Note that if dynamic sampling results in measuring every run or every two runs, the points should be excluded for updating the standard deviation until returning to baseline sampling frequency.

Other options would be applicable for an online update. Since equation (2.22) goes to zero as *n* goes to infinity, renewal of starting points should be required. A good option would be to keep estimating until after taking several points (e.g., five points). Also, in order to prevent the estimates from being sensitive to noise, a confidence interval for variance can be implemented.

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2),n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2),n-1}} \tag{2.23}$$

2.4 SIMULATION RESULTS

Simulations are focused on the online dynamic sampling algorithm rather than information from an historical data set. Even if appropriate settings of several parameters are required to make the algorithm perform well, the online dynamic sampling algorithm described above is quite robust so we can obtain an excellent result with reasonable selections of those parameters.

Prior to simulations with industrial data, a simple test signal with step disturbances is simulated. The process model is a simple linear equation (2.24).

$$y = x + b \tag{2.24}$$

where *y* is output, *x* is input, and *b* is offset with step changes (Figure 2.12).

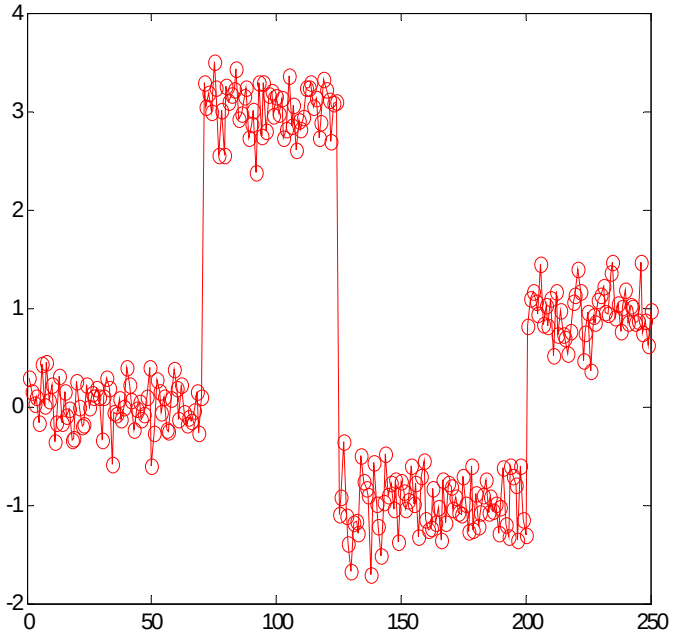


Figure 2.12 Test signal with step disturbance and noise

An EWMA filter is used to update offset and process capability (*Cpk*) is selected as a metric for control performance. The target is assumed to be zero. For the predefined values of dynamic sampling, prior probability, P_0, is set to 0.1, the initial process standard deviation, σ, is 0.25. If the left hand side in equation (2.17) is greater than 0.6, dynamic sampling is turned on to measure every run, and if it is less than 0.6 but greater than 0.3, measurements are taken every two runs. Otherwise, the sampling rate is kept at the baseline rate – measured every third run. As mentioned in section 2.2, the baseline sampling rate can be selected by a process profit function with an historical data set. Then dynamic sampling selection rules can be more complicated with more categories or simpler such as just dual sampling rate. The standard deviation keeps updating with five point-moving windows but excludes the points if dynamic sampling is turned on to measure every run or every two runs, as mentioned in section 2.3.3. The simulation result is shown in Figure 2.13. The comparison of the result of uniform sampling with measuring every 3 runs is also shown in Figure 2.14. Even if dynamic sampling measures additional runs, the *Cpk* value is much better than uniform sampling. In other words, a few more measurements from the online dynamic sampling algorithm make the control performance better. The following industrial examples will show the effect of dynamic sampling for semiconductor manufacturing process.

2.4.1 Industrial Example 1 : Chemical Vapor Deposition

The first industrial example is a chemical vapor deposition (CVD) process of TiN deposition for contact barrier from the DMOS6 in Texas Instruments Inc [54][55]. The goal is deposition thickness control of the TiN layer by adjusting deposition time.

A linear process model is used by the run-to-run controller

$$y_k = rate_k \times u_k + b_k \tag{2.25}$$

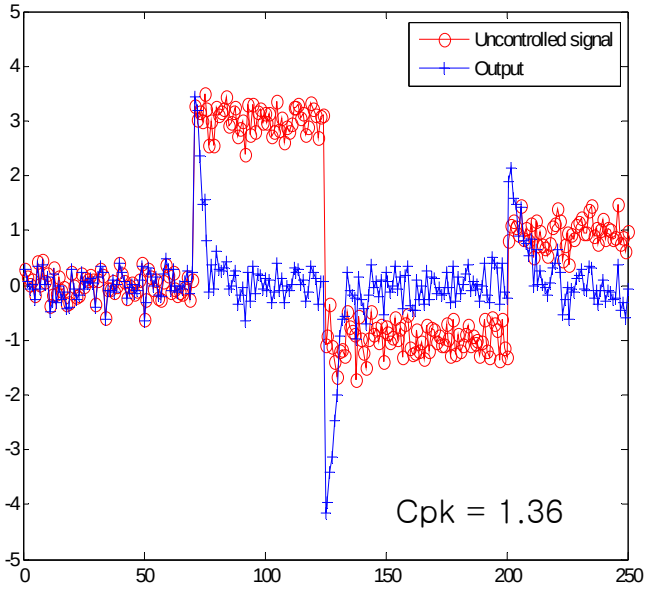


Figure 2.13 Simulation result of dynamic sampling (the number of measurement = 97)

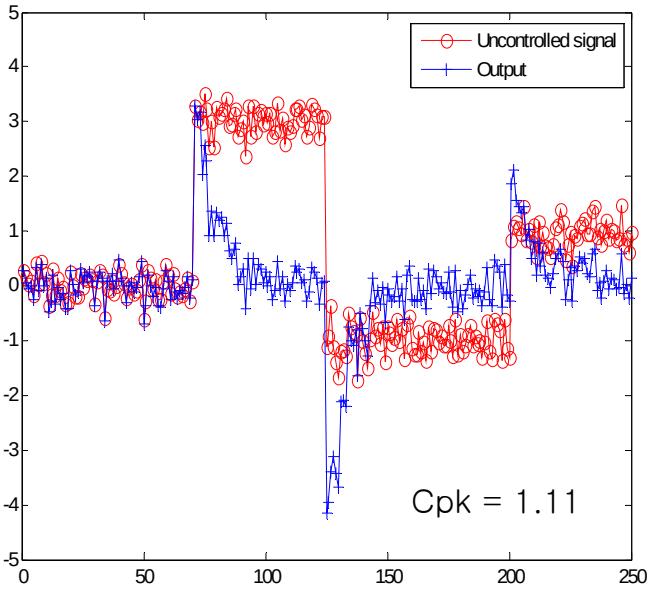


Figure 2.14 Simulation result of uniform sampling (the number of measurement = 84)

where y_k is the thickness of TiN layer, u_k is the deposition time, and b_k is a metrology offset for the specific target. There are two different offsets for two different thicknesses (two different products).

EWMA estimator is used to update $rate_k$

$$rate_{k+1} = rate_k + \lambda \times [(y_k - b_k)/u_k - rate_k] \tag{2.26}$$

Deposition time can be calculated using the following process model inversion with target:

$$u_{k+1} = (TGT - b_k)/rate_{k+1} \tag{2.27}$$

where TGT is the target of thickness.

The prior probability, P_0, and the initial process standard deviation, σ, are from the historical data set. Like the simple simulations above, if equation (2.17) is greater than 0.6, dynamic sampling is turned on to measure every run, if less than 0.6 but greater than 0.3, we measure every two runs. Otherwise, the sampling rate is kept at the baseline rate, i.e., measured every three runs. The standard deviation also keeps updating with five point-moving windows but excludes the points if dynamic sampling turns on to measure every run or every two runs.

The simulation result of both products and each *Cpk* are shown in Figure 2.15. For comparison, the result of uniform sampling with measuring every 3 runs and its *Cpk* are shown in Figure 2.16. All simulation results are normalized and actual scales are not reported because this is Texas Instrument's proprietary data set. As shown in Figures 2.15 and 2.16, dynamic sampling works properly for both products.

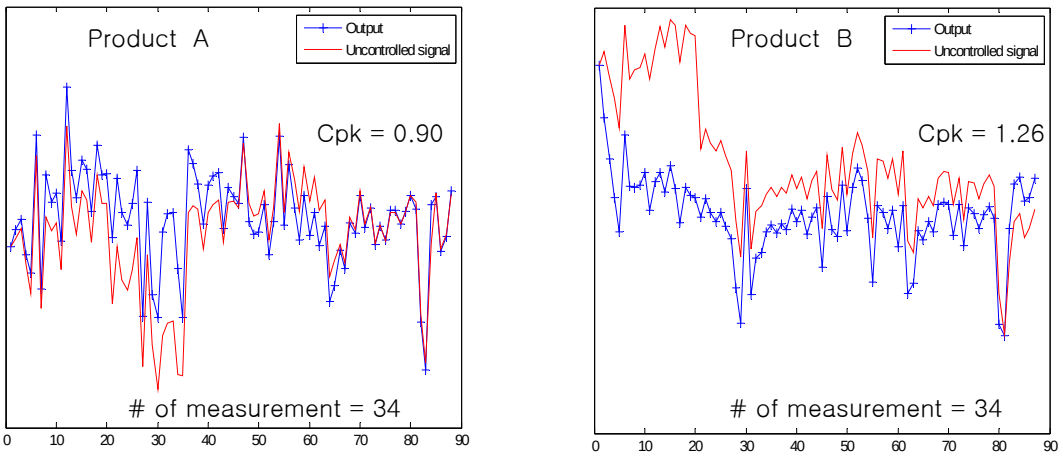


Figure 2.15 Simulation result of dynamic sampling of CVD process

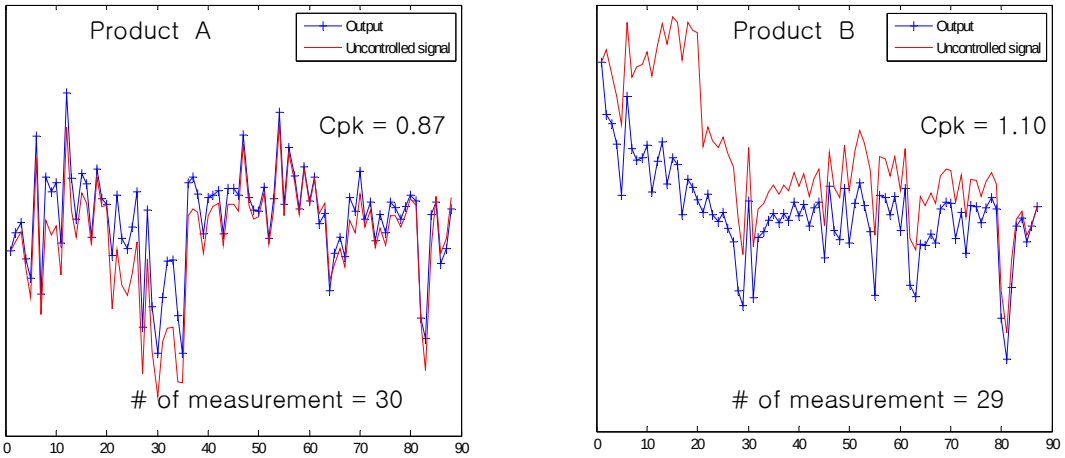


Figure 2.16 Simulation result of uniform sampling of CVD process

2.4.2 Industrial Example 2 : Physical Vapor Deposition

The second industrial data set is a physical vapor deposition (PVD) process of Ta barrier deposition at 40kw from the DMOS6 wafer fab in Texas Instruments Inc. Ta is a precursor to copper seed deposition. The control objective is deposition thickness control of Ta layer in order to get an appropriate resistivity by adjusting deposition time.

A linear process model is used by the run-to-run controller

$$y_k = rate_k \times u_k \tag{2.28}$$

where y_k is the Ta thickness, u_k is the deposition time, and an EWMA filter is used to update $rate_k$

$$rate_{k+1} = rate_k + \lambda \times (y_k / u_k - rate_k) \tag{2.29}$$

Deposition time can be calculated using the following deadbeat control law with target.

$$u_{k+1} = TGT / rate_{k+1} \tag{2.30}$$

where TGT is the target of the thickness.

Unlike in the CVD process, the rate decreases linearly during the process. Thus, other model formats such as double EWMA can be considered to compensate drift. However, dynamic sampling also could have positive effects on this PVD process. The prior probability, P_0, and the initial process standard deviation, σ, are from the historical data set. Other settings are similar to the CVD process in section 2.4.1. The simulation results of dynamic and uniform sampling every 3 runs are shown in Figures 2.17 and 2.18.

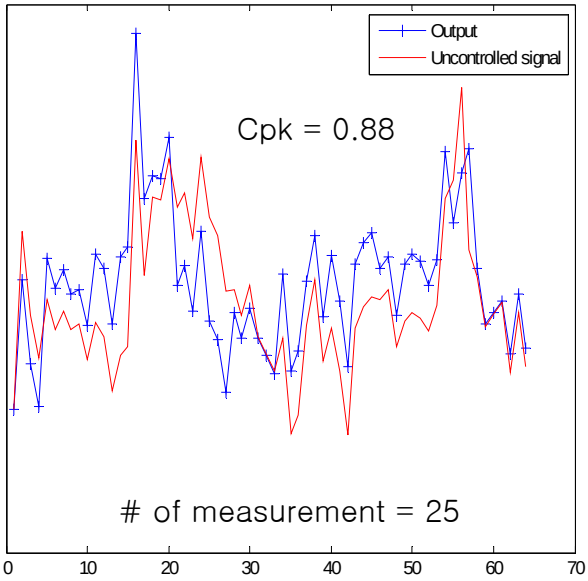


Figure 2.17 Simulation result of dynamic sampling of PVD process

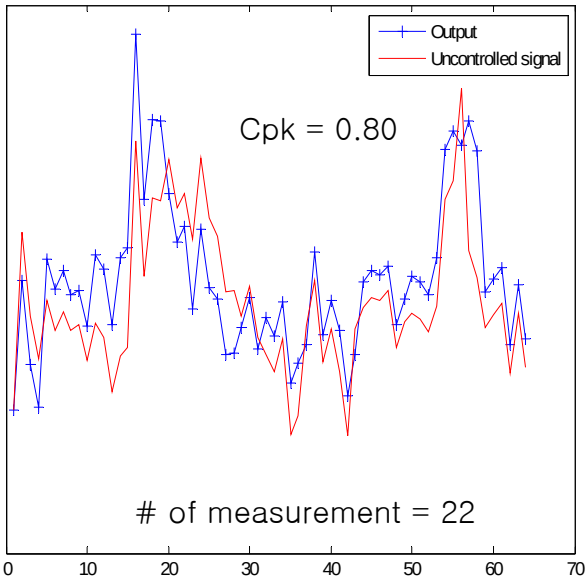


Figure 2.18 Simulation result of uniform sampling of PVD process

2.4.3 Industrial Example 3 : Photolithography Process

The third industrial data set is a photolithography process for shallow trench isolation (STI) of the critical dimension (CD) controller from DMOS6 in TI [56]. CD should be controlled to make mask patterns for following etch processes by adjusting the exposure time. The process model has a linear format as an equation (2.31).

$$y_k = s \times u_k + b_k + c_k \tag{2.31}$$

where y_k is the CD, u_k is the exposure time, s is slope which is constant, b_k is the reticle offset which is also constant with the values from the data set, and c_k is the scanner offset which is updated by an EWMA estimator.

$$c_{k+1} = c_k + \lambda \times (y_k - s \times u_k - b_k - c_k) \tag{2.32}$$

Exposure time can be calculated using the plant inversion control law with target.

$$u_{k+1} = (TGT - b_{k+1} - c_{k+1}) / s \tag{2.33}$$

where TGT is target of the CD.

There are two different lithography equipments and two different products in this data set. Thus four processing threads are considered in simulations. The prior probability, P_0, and the initial process standard deviation, σ, are assumed to be reasonable values. Other settings are also similar to the CVD process in section 2.4.1. The simulation results of dynamic sampling and uniform sampling with measuring every 3 runs are shown in Figures 2.19 and 2.20.

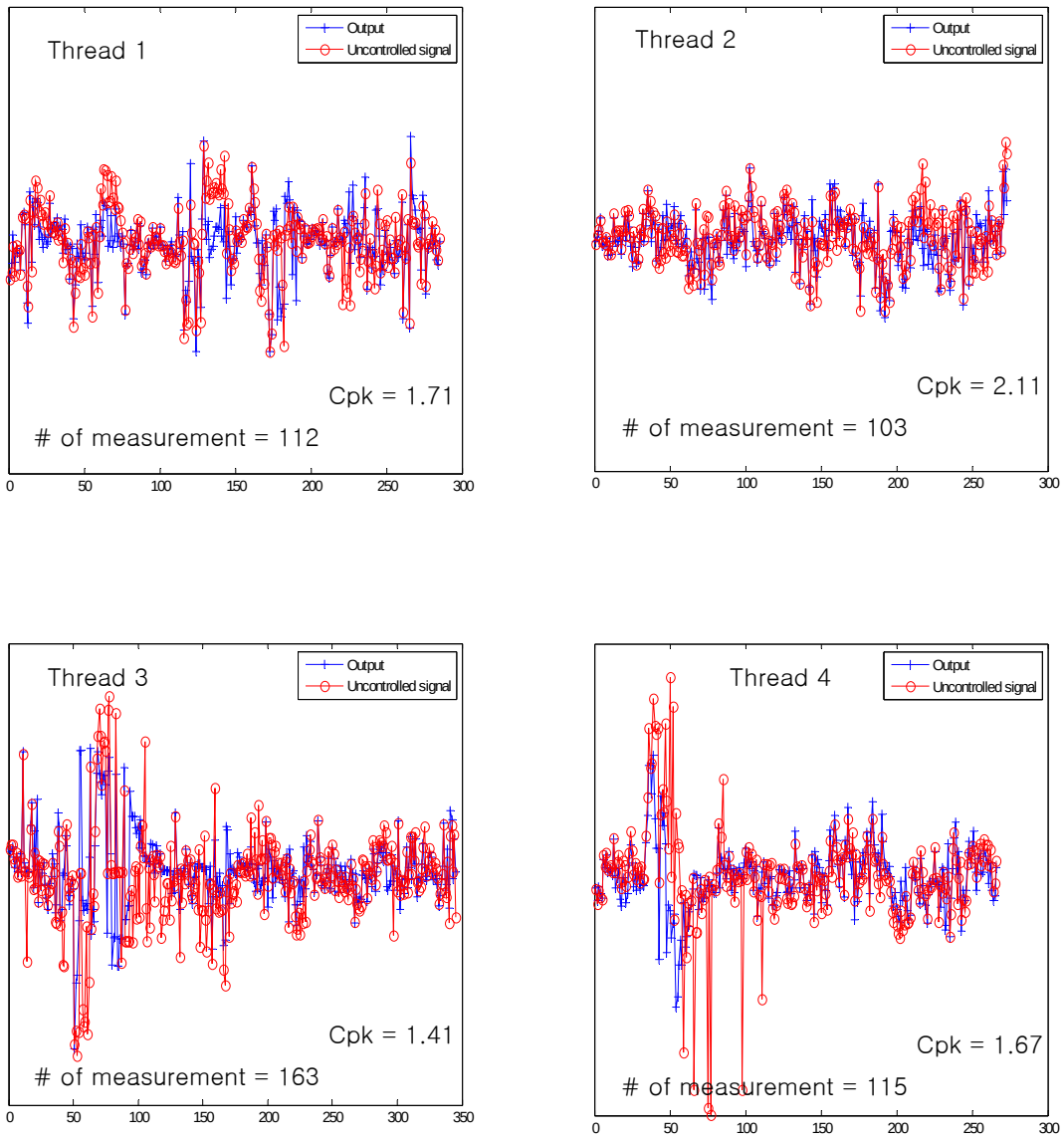


Figure 2.19 Simulation result of dynamic sampling of photolithography process

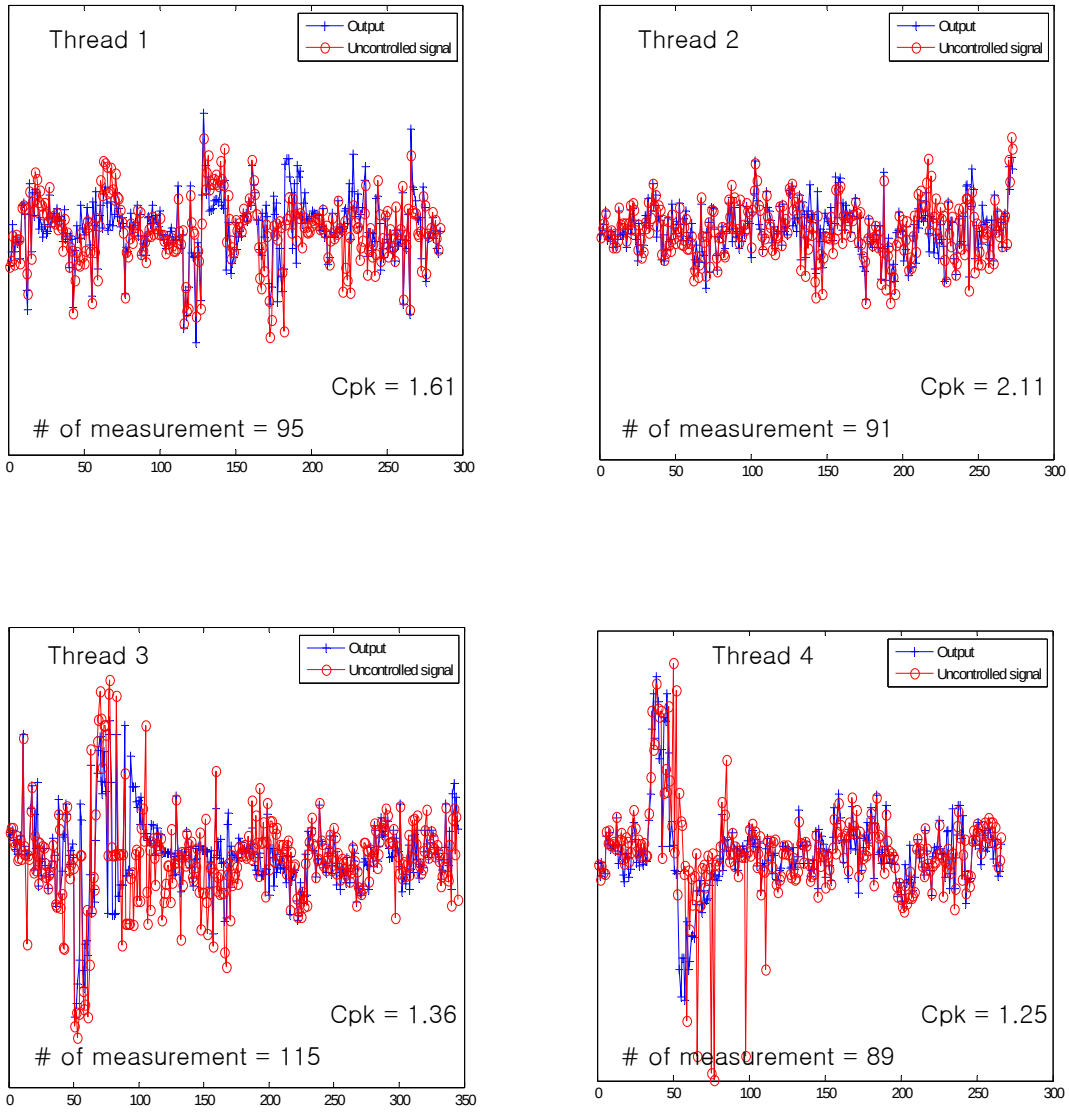


Figure 2.20 Simulation result of uniform sampling of photolithography process

According to the simulation results of the photolithography process, the performance of dynamic sampling depends on the characteristic of the process. For thread 2, this process is very stable so every run is within the control limits. In this case, dynamic sampling has no effect. But it does not degrade the control performance and thus the online dynamic sampling algorithm would not give negative effects to a stable process.

However, there might be some problems for a data set with a large amount of noise. For thread 3, even though dynamic sampling shows better control performance than uniform sampling, it measures many more wafers than uniform sampling. That is because the dynamic sampling algorithm believes every noise with a large magnitude is a step change, so it measures the consecutive runs. In order to solve this problem, the sampling selection rules need to be changed. If the value for sampling every runs is changed to 0.8 instead of 0.6 and the value for samples every two runs changes from 0.3 to 0.6, the process capability would be 1.43 when the total number of measurements is 130. Obviously, this result is better than that of Figure 2.19 from both economic and performance points of view.

Dynamic sampling can be an optimal solution for the data set with step disturbances regardless of their magnitude. Especially when there is data with large step disturbances (thread 4), the effect of dynamic sampling would be amplified compared to data with small step disturbance (thread 1).

2.4.4 Industrial Example 4 : Thin Film Deposition Process

The last industrial data set is a chemical vapor deposition (CVD) process with many threads from the DMOS6 wafer fab in Texas Instruments Inc. The goal is also deposition thickness control by adjusting deposition time.

A linear process model is used by the run-to-run controller

$$y_k / 4 = rate_k \times u_k + b_k + c_k \tag{2.34}$$

where y_k is the thickness, u_k is the deposition time, $rate_k$ is slope which is assumed to be constant, b_k is constant and c_k is an offset which needs to be updated run to run. There are two different products and six different machines, so a total of twelve threads are used for these simulations.

c_k is updated by EWMA estimator.

$$c_{k+1} = c_k + \lambda \times (y_k / 4 - rate_k \times u_k - b_k - c_k) \tag{2.35}$$

And deposition time can be calculated using the plant inversion control law with target.

$$u_{k+1} = (TGT / 4 - b_{k+1} - c_{k+1}) / rate_k \tag{2.36}$$

where TGT is target of the CD.

The simulation result of each thread is shown from Figure 2.21 to Figure 2.23 and Table 2.1 and 2.2. Note that all simulation results are normalized and actual scales are not reported because this is Texas Instrument's proprietary data set. Also note that it is auto-scaled in Figure 2.21 to 2.23 so every graph has different scale. For the comparison, the simulation results of uniform sampling with measuring every three runs are also shown in Table 2.1 and 2.2. Table 2.1 shows control performance in terms of process capability values and Table 2.2 has the number of measurement for each sampling.

C_{pk}	Dynamic sampling	Uniform sampling (every 3 runs)
Thread1	0.8144	0.8144
Thread2	0.1781	0.1381
Thread3	0.4190	0.4796
Thread4	0.3554	0.3407
Thread5	0.2670	0.2670
Thread6	0.2792	0.2693
Thread7	1.0328	1.0750
Thread8	0.2214	0.0938
Thread9	0.2683	0.2238
Thread10	0.6280	0.5688
Thread11	0.5008	0.5008
Thread12	0.1409	0.0350

Table 2.1 *Cpk* of dynamic sampling vs. uniform sampling (every 3 runs)

# of measurement	Dynamic sampling	Uniform sampling (every 3 runs)
Thread1	12	12
Thread2	30	26
Thread3	11	10
Thread4	31	28
Thread5	13	13
Thread6	27	25
Thread7	11	10
Thread8	22	18
Thread9	16	13
Thread10	27	17
Thread11	9	9
Thread12	17	10

Table 2.2 # of measurement of dynamic sampling vs. uniform sampling (every 3 runs)

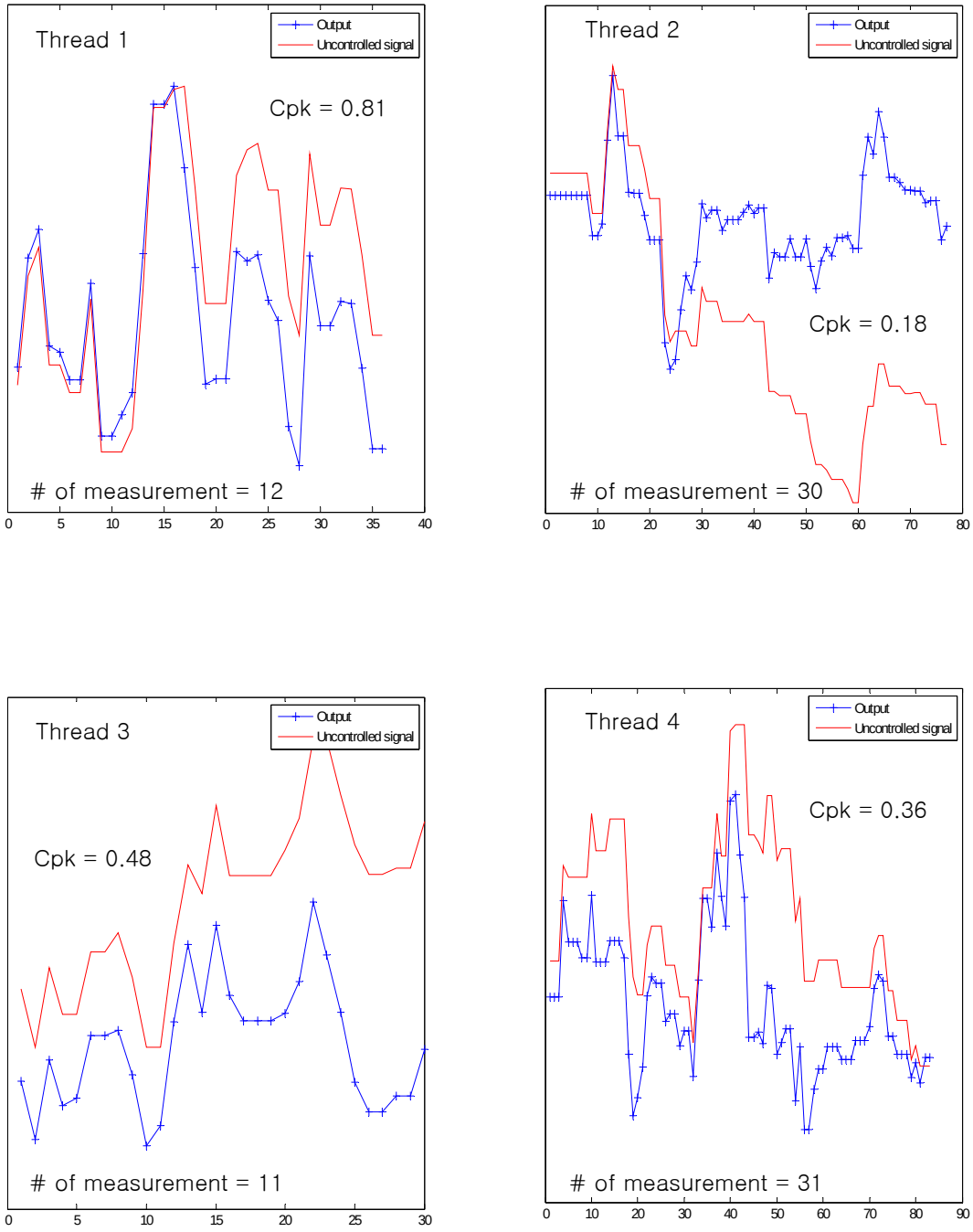


Figure 2.21 Simulation results of dynamic sampling in CVD process (Thread 1 ~ 4)

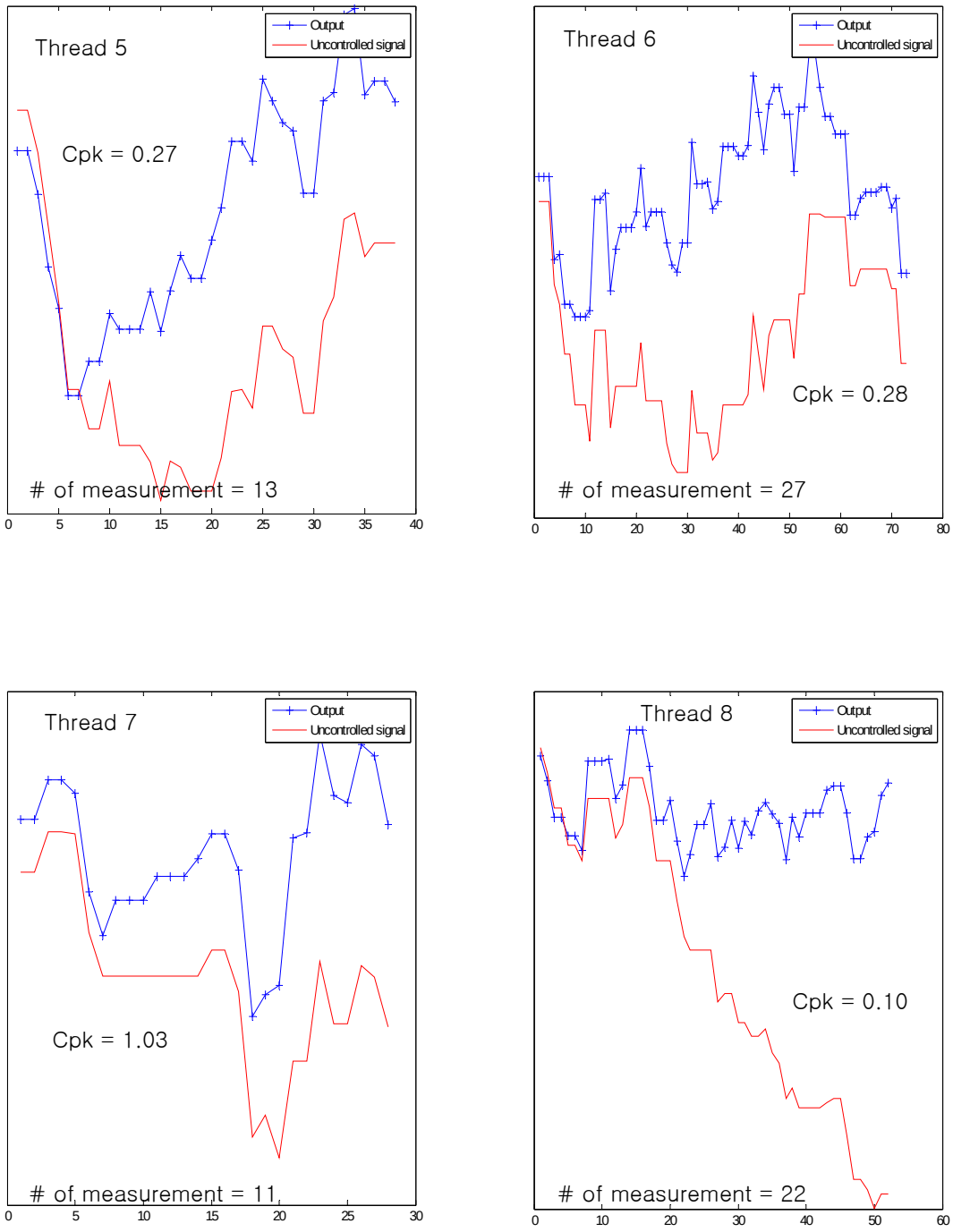


Figure 2.22 Simulation results of dynamic sampling in CVD process (Thread 5 ~ 8)

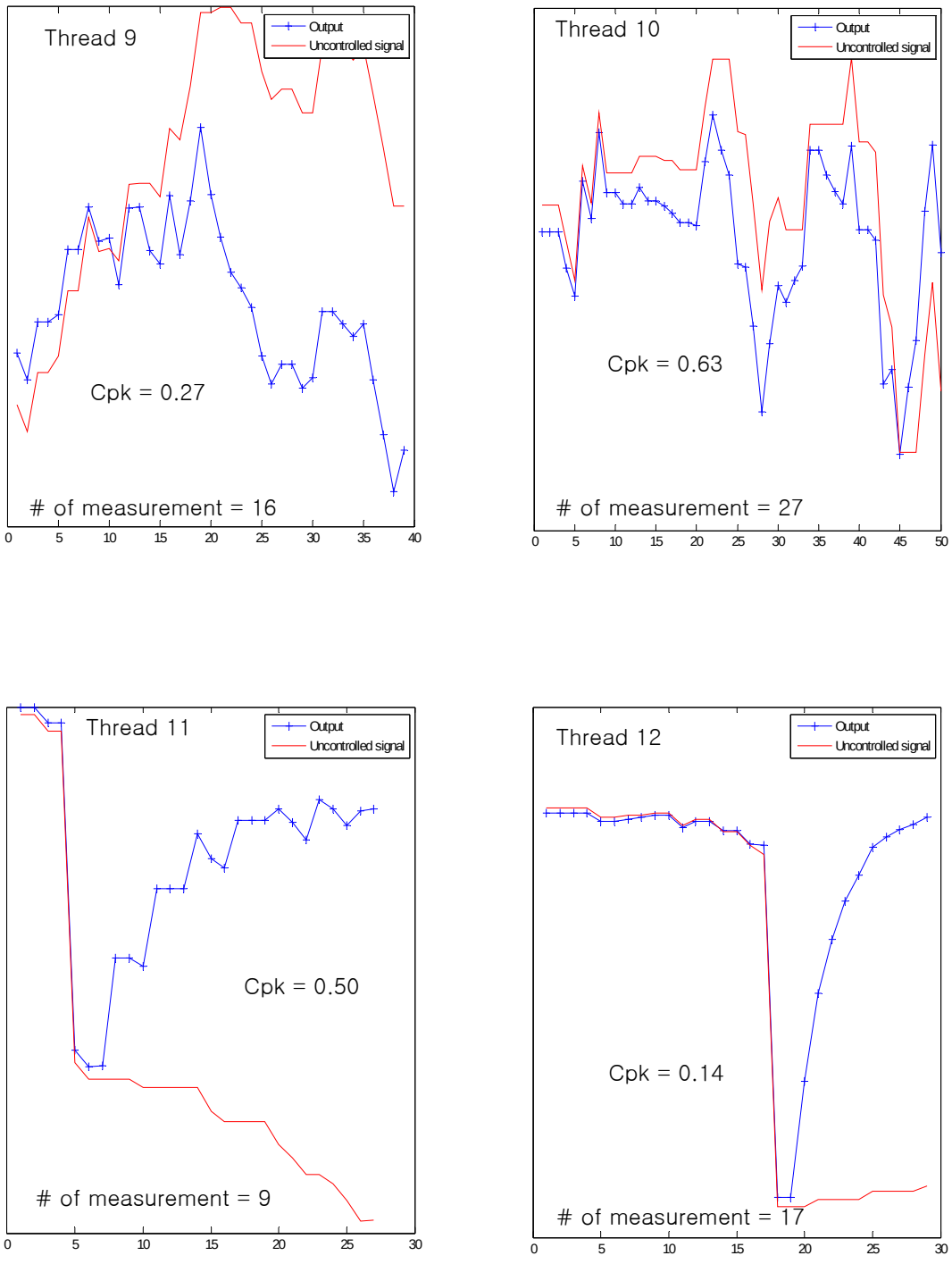


Figure 2.23 Simulation results of dynamic sampling in CVD process (Thread 9 ~ 12)

The number of measurements is the same for dynamic sampling and uniform sampling in threads 1 and 5. This means the process in those threads is stable so they do not trigger dynamic sampling. Step disturbances occur in threads 2, 8, 9, 10, and 12 so dynamic sampling can contribute to obtain higher *Cpk* results compared with those of uniform sampling.

2.5 SUMMARY AND CONCLUSION

Process cost reduction is always an issue once a technology's yield has matured, and it can be attacked by metrology reduction or elimination. Thus, optimal metrology sampling contributes to reduced process cost while maintaining high product quality. In other word, the net profit could be maximized.

Dynamic sampling is a method of changing the sampling frequency based on prior observations. The sampling rate is adjusted as the process progresses and increases when the process has moved away from target. Dynamic sampling is expected to be applicable to any processes online since it is flexible and adjustable to process dynamics. Thus, it would have better economical performance than a uniform rate of sampling (higher profit) which uses fixed sampling frequency. But several process parameters should be pre-determined for the best performance.

A baseline sampling rate can be determined by offline sampling rate selection using process profit function with an historical data set. We developed a process profit function and a new objective function for off-line sampling plan for a unit process based on the metrics of cost of ownership (CoO). By using process profit function as an economic objective function, an optimal sampling frequency can be found.

Also, the magnitude of error to trigger sampling rate changing could be selected from the ideas of Bayesian disturbance detection using the prior probability and the

process standard deviation. A statistical process control (SPC) chart with open-loop data gives suitable information to estimate the parameters.

Since a semiconductor manufacturing environment is very dynamic, even if the process is stable, it changes over time due to a variety of reasons. Therefore, online updating should be considered in order not to degrade the dynamic sampling performance. Thus, simulations are focused on the online dynamic sampling algorithm rather than getting information from an historical data set. Even if appropriate settings of several parameters are required to make the algorithm perform well, the online dynamic sampling algorithm is quite robust so we can obtain an excellent result with reasonable selections of those parameters.

A simple test signal with step disturbances and an industrial data set of photolithography and thin film deposition process from the DMOS6 wafer fab in Texas Instruments Inc. are simulated. According to the simulation results, the performance of dynamic sampling depends on the characteristic of the process. When the process is very stable so that every run is within the control limits, dynamic sampling has no effect. But it does not degrade the control performance and thus the online dynamic sampling algorithm would not give negative effects to a stable process. Dynamic sampling can be an optimal solution for the data set with step disturbances regardless of their magnitude. Especially when there is data with large step disturbances, the effect of dynamic sampling would be amplified compared to data with small step disturbance.

CHAPTER 3

Optimal Sampling for Multiple Product and Process Environment

3.1 BACKGROUND AND MOTIVATION

In semiconductor manufacturing it is becoming more common to make many different products by employing a variety of process sequences utilizing the same pieces of equipment, mainly because of the high capital costs associated with the tools and the limited capacity of the facility [57][58][59]. Because of the equipment and in situ metrology limitations, only a subset of the measurements important to a process is generally made after each run during production. By judicious selections of process measurements, the fab-wide sampling strategies maximize the amount of information that is shared across different batches by capturing tool and product characteristics in common parameters.

The development of an optimal measurement sampling strategy under multiple product and process environments ("high mix manufacturing" [60][61][62][63]) is challenging since the quality of control must be balanced with the cost of measurement. Frequent measurement facilitates every product reaching its target but it increases the cost. Thus, it is necessary to optimize the sampling plan in order to quickly identify the sources of prediction errors and decrease the metrology cost and cycle time.

The first step of this work is to clearly define the problem and the control environment while identifying current best practices using data. This means that realistic operating conditions such as multiple simultaneous products, measurement costs, and imperfect data are taken into account. Advanced state estimation techniques are utilized in order to quantify these concerns and provide an alternative to methods based

on sets of operating rules. An algorithm is created which treats sampling plans as operating decisions so that their effect on other variables of interest can be studied. Also an objective function is developed that combines the cost of sampling with performance of the controller in terms of yield and cycle times. A new controller formulation will be proposed that includes concerns about system identification and parameter uncertainty in its objective function and constraints.

The ultimate goal of this research is to better understand how to effectively generate and use the data available in a high-mix semiconductor manufacturing facility for process control. Understanding exactly what information is needed for effective process control allows better planning and integration with overall business needs and strategies.

3.2 CONTEXT MODEL

A key to making a successful control algorithm for a high-mix environment is proper selection of the plant or system models. In order to compensate for the different equipment and product combinations, one must able to expect how a run would achieve given the identified characteristics of the equipment and batch concerned. One of the better known techniques for achieving this task is segregated control. The main idea under this methodology is that while it is not constantly possible to directly measure the variables that make run perform differently; it is possible to classify runs simultaneously into groups, which contain similarly behaving runs. For example, the performance of a given run may depend on the size of the features on the product, the current state of the selected processing equipment, and the effects of previous processing. These effects can be hard to quantify, but if the process is relatively stationary, the current run is likely

to perform in a similar way to previous runs of the same product on the same tool with similar prior processing steps.

The key to success with segregated control is the appropriate identification of the source of variation in the process. It is significant to define the bins so that they receive all sufficient amounts of data in order to identify the source of variation properly. If the group is defined by too many source of variation, there will not be enough data in each bin to have an effective controller. Conversely if enough sources of variations are not identified, control performance would be degraded since the runs in the bin are not similar enough to each other.

As semiconductor devices continue to shrink, manufacturing requirement and tolerance become stricter. From a control point of view, this means that more sources of variation should be properly identified so that the controller can compensate for them. Unfortunately, because the number of runs does not change, less data is obtainable to the individual control loops in a segregated controller. Even though there are only two options for every source of variation, the number of states that should be identified increases exponentially.

The goal of this section is to analyze the sources of process variation to reach a design with a smaller number of states. Preferably, there would be only one state for each source of variation. The way to accomplish this is to have a single controller that knows how to interpret all the process data. The control action for a particular run depends on as much information as can be learned about it from all the other runs that have occurred.

In this section, we extend a simple run-to-run controller for a current single process to the case of multiple products and equipments using the concept of processing context. The term processing context is describing the combination of equipment,

product, and other factors that identifies the processing environment for a particular run. The context should have enough information to guarantee that different lots or wafers run under the same context would have similar results under similar processing conditions. A significant effort may be required to define a system that has all contexts under consideration [57][59]. Then, a single model can be obtained that covers every case within the selected system boundary. O. Vanli et al. proposed a model selection approach that uses ideas from statistical linear models and stepwise regression to identify the context variables that contribute most to the auto-correlation and to the offsets in the data [64].

A good example of a system that shows this behavior is the chemical mechanical planarization (CMP) of inter-layer dielectric (ILD) layers. Due to differences in pattern density and processing history, each layer / product combinations planarizes at a different rate. In addition, as each product is qualified to run on several equipments, there are also systematic variations caused by differences between equipments. Furthermore, the CMP process is characterized by the use of a slurry. On a given equipment, different slurries can be used to obtain different processing characteristics. Thus, the control problem is to determine the optimal settings for each product / layer / equipment combination.

The control objective is to keep each run at the target, regardless of product and equipment combinations. The simplest approach is to presume that the one set of states applies to all processing contexts. The disadvantage of this method is a situation with several contexts involved is that the outputs associated with each process can be markedly different from each other. When this happens, each knob to new contexts appears as a step disturbance to the controller, as it has no understanding of why the outputs would have changed so much.

The alternative method is the segregated control (threaded approach) described above, where runs with similar processing contexts are grouped so that they share parameter estimates. Each group of runs is called a control thread. In this method, there is no requirement to identify product and equipment biases individually from each other. Each combination basically has its own output estimate and updates this estimated based only on measurements from runs under the grouped contexts. This method has serious difficulty when a disturbance to one equipment, for example, has to be identified by every context where that equipment is used. This can be overwhelming in a large system because of the large number of runs that would ignore their targets while the different contexts were updating their parameter estimates. Ideally, that information should be instantly shared between all contexts when the disturbance affects.

There are cases where biases are recognized to be caused by different parts of the processing context (non-threaded approach). The information among different contexts can be shared in this approach [65]. Recently, several different methods based on this non-threaded approach have been proposed for high-mix manufacturing environment [66][67][68]. A key example is that equipment to equipment variation is repeatable regardless of the product being run. Likewise, product to product variation is consistent even when the products are run on different equipments.

Therefore, while many factors have an effect on the outputs of a run, the processing equipment and the product are assumed to be the main sources of output variations that have measurable effects on the run [69][70]. A hypothetical etching or polishing process is considered where there are several tools and several products. They can be modeled in any product / equipment combination in a multiplicative structure.

In a state space model format, consider the following system and measurement model:

$$x_{k+1} = Ax_k + Bu_k + Gw_k \tag{3.1}$$

$$y_k = Cx_k + v_k \tag{3.2}$$

where x_k is process state, u_k is the process input, y_k is measurement output, and w_k and v_k are random process noise and measurement noise with zero mean and covariance matrix Q and R, respectively.

If there is the system model with three equipments and three products, the process model for a particular tool and product can be written

$$y = x_{t_i} + x_{p_j} + u \tag{3.3}$$

where y is the output, x_{ti} and x_{pj} are the equipment and product context which have their own bias, and u is the chosen process input. The target is a particular value of y. For the simplicity, the target is assumed to be zero. For a system with three equipments and three products, a state vector is constructed as

$$\boldsymbol{x} = \begin{bmatrix} x_a & x_{t_1} & x_{t_2} & x_{t_3} & x_{p_1} & x_{p_2} & x_{p_3} \end{bmatrix}^T \tag{3.4}$$

where x_a is the adjustment state, and all equipments and products are tracked with individual elements. The adjustment state is used to account for the fact that the system has no "memory" of the processing times chosen on prior runs. This special state captures the effect of the process input on a particular run in a way that only affects the current run.

The state transition matrix $\boldsymbol{A}$ in equation 3.1 is a square matrix of size 3 + 3 + 1 by 3 + 3 + 1 and is given by

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.5}$$

which shows that all the equipment and product parameters are nominally stationary. The process input matrix $\boldsymbol{B}$ is assumed to affect only x_a, so it is of size (3 + 3 + 1) by 1 given by

$$\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \tag{3.6}$$

The complete output matrix $\boldsymbol{C}$ for the process is given by

$$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \tag{3.7}$$

The observability matrix $\boldsymbol{O}$ can be used to determine whether the system is observable or not [71][72]. For a time invariant linear system, a typical observability test can be performed by computing the rank of the matrix.

$$\boldsymbol{O} = [\boldsymbol{C}^{\mathrm{T}} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}} \quad (\boldsymbol{A}^{\mathrm{T}})^2\boldsymbol{C}^{\mathrm{T}} \quad \cdots] \tag{3.8}$$

Because the system in equation (3.8) is rank deficient by one, so the system is not observable for its current form [57]. This is because all runs are confounded by both a product bias and an equipment bias. Additional constraints are required in order to make the system observable. One option of adding supplementary measurements to the system is to use qualification runs using test wafers.

3.3 KALMAN FILTER AND QUALIFICATION RUNS

During the qualification runs, regular tests are run on the equipment that can incorporate the processing of specially prepared test lots or wafers. Because these runs are identical across the group of equipments, it is assumed that no product bias applies to these qualification runs. This means that the equipment bias can be measured directly during the procedure of the test on particular equipment.

When such runs are not an option, it can be also possible to simply choose a reference equipment or product which has a known bias. The main drawback of this method is that it may be hard to identify a reference equipment or product in a high-mix manufacturing environment where new products are added, old products are retired, and equipment needs to be replaced or shut down for maintenance events.

In the example above, qualification runs would be added as the following three equations:

$$y_1 = x_{t_1} \tag{3.9}$$

$$y_2 = x_{t_2} \tag{3.10}$$

$$y_3 = x_{t_3} \tag{3.11}$$

If the qualification runs are added to the example three equipments and three products system, then the new output matrix C is given

$$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.12}$$

which results in an observable system (full rank).

In a discrete-time system where only a subset of the all possible measurements are taken at each time step, it is required to update state whenever new measurements arrive, obtaining as much information as possible from that new measurements into the state estimates. In general, prediction errors which are deviations between the measured and predicted outputs make changes to the state estimates.

Each product measurement is assumed to be affected by more than one state. Conceptually, such a problem appears impossible to solve, because a single known measurement has to be used to update more than one unknown state. However, there is

additional information available. If the estimator already has estimates for all the states as well as a measure of the certainty of those estimates, then a statistical analysis can be performed to determine which states are most likely to be the cause of an observed error in the predicted measurement. It is necessary to optimally estimate the true values of the system states given only partial information.

State and variance estimates are updated at two distinct times: at the processing time and when measurement information arrives. Any time the process runs, the states and uncertainty in them may change, regardless of whether any measurements are taken during that run.

The update at the processing time is a predictive update using the known model of the system including estimates of the noise disturbance. In this case, the model error is assumed to have zero mean and known covariance. Both the current state and variance estimates are passed through the state transition matrix to arrive at new estimates. Performing this update at the processing step ensures that any predictions made using the state estimates use the best known values, even if measurements have not yet been taken.

Any time the process is run, the state estimates and error covariance are updated in the Kalman filter [73][74][75][76] using the following equations:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k \tag{3.13}$$

$$P_{k+1} = AP_kA^T + GQG^T \tag{3.14}$$

When a measurement is taken, the Kalman gain, the state estimates and error covariance are updated:

$$K_{k+1} = P_kC^T(CP_kC^T + R)^{-1} \tag{3.15}$$

$$P_{k+1} = (I - K_{k+1}C)P_k \tag{3.16}$$

$$\hat{x}_{k+1} = \hat{x}_k + K_{k+1}(y_{k+1} - C\hat{x}_k) \tag{3.17}$$

3.4 SIMULATION RESULTS

In this section, online dynamic sampling algorithm from chapter 2 is extended for high-mix manufacturing environments. Unlike threaded control using chapter 2, non-threaded state estimations are used with Kalman filter and qualification runs as described in a previous section. By using simulation, the effects of qualification runs and dynamic sampling will be verified.

Test data signals are created by emulating the characteristics of actual manufacturing data. As explained in section 3.2, two processing contexts are primary sources of variation. They are processing equipments and products. Three equipments and three products are used in this simulation. Thus, there are a total of six processing contexts. Each context has its own bias and characteristics such as time series property with a different magnitude of noise [77][78][79]. Figure 3.1 shows the created data signal with 1000 runs.

Equipment biases are set to have more variation since they drift as the process runs. After maintenance events, they are recovered to clean states. However, those changed states can act as a step disturbance since the biases might be different from previous states. Therefore a qualification run will be added right after maintenance events in order to reset new equipment states by direct measurements. For simplicity all equipments have maintenance events and qualification runs are made every 100 runs. There are several step disturbances in product contexts but no drift. The runs in which step disturbances occur are randomly selected.

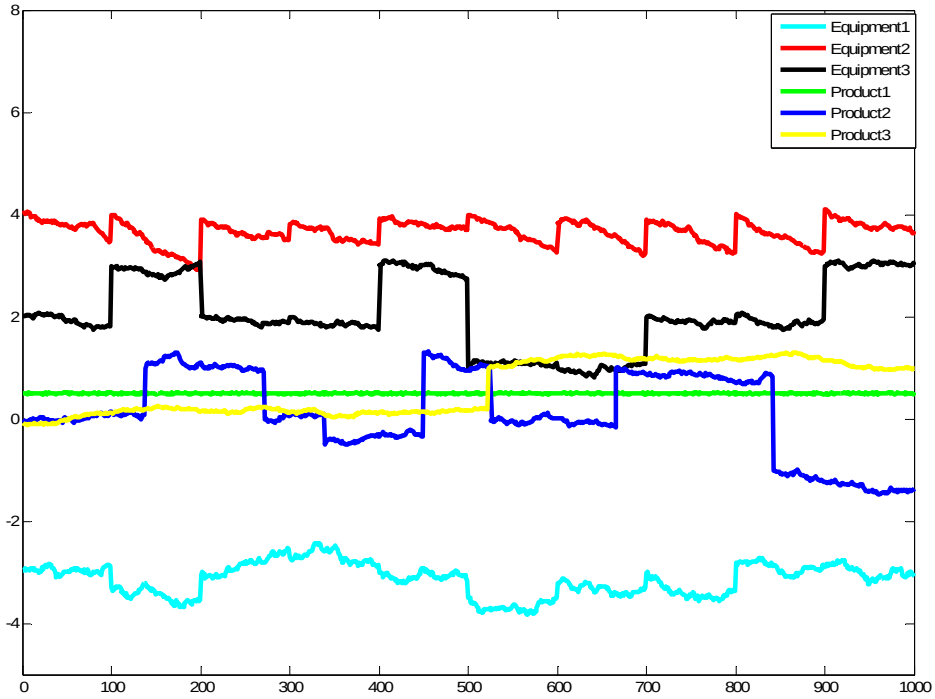


Figure 3.1 Created test data signal with three equipments and three products

The process model is exactly same model as in section 3.2. The first simulation is for the effect of qualification runs to estimate individual states as well as combined states. Figure 3.2 is the simulation result of 10% qualification runs (every 100 runs) and no qualification run in the case of 100% sampling, which means every wafer is measured. The simulation with 10% qualification is a slightly better than no qualification for overall control performance. However, there are many differences between those two simulations for estimating individual context states. From Figures 3.3 to 3.8 show the result of estimating each state of all six processing contexts.

With uniform sampling strategy, an additional analysis of the effect of qualification runs is performed. By varying sampling rate from one to four, the control performance results are obtained with *Cpk* values for the cases having 10% qualification runs and no qualification. Table 3.1 shows the simulation results.

C_{pk}	**10% qualification runs**	**No qualification run**
Every run	2.5940	2.4140
Every 2 runs	1.8706	1.7084
Every 3 runs	1.4748	1.4018
Every 4 runs	1.1323	1.0736

Table 3.1 Simulation results of uniform sampling – 10% qualification runs vs. No qualification

The performance of 10% qualification runs is better than that of no qualification run for all cases. However, the performance difference between those two cases becomes smaller as sampling rate decreases. That is because there are three more unknown product contexts even though equipment biases are identified with qualification runs.

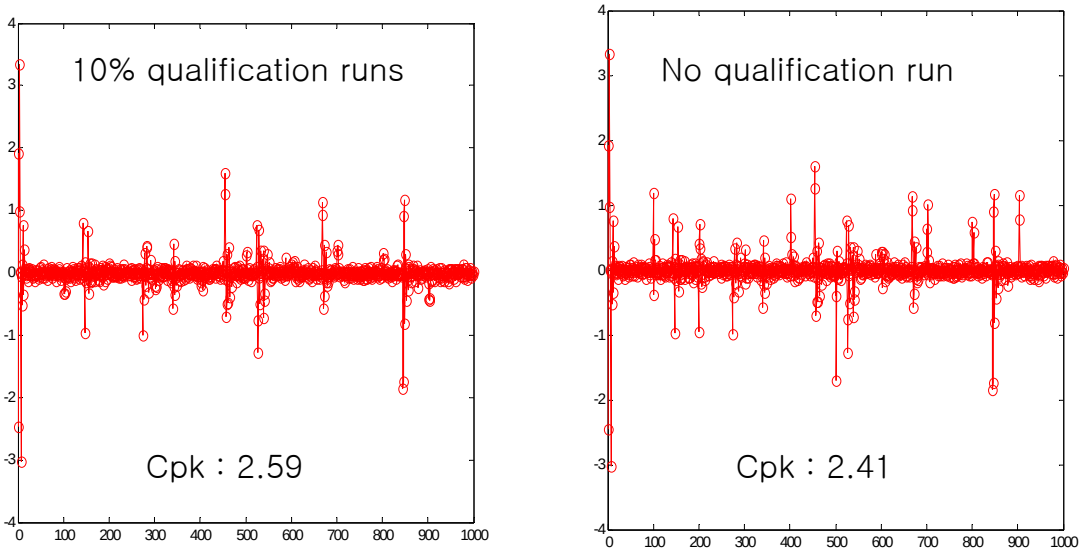
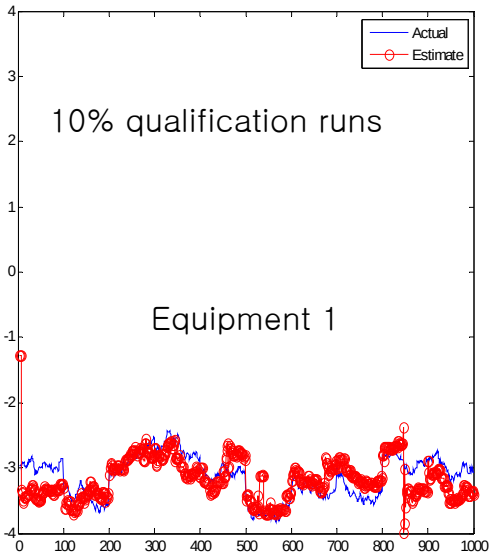


Figure 3.2 Control performance – 10% qualification runs vs. No qualification



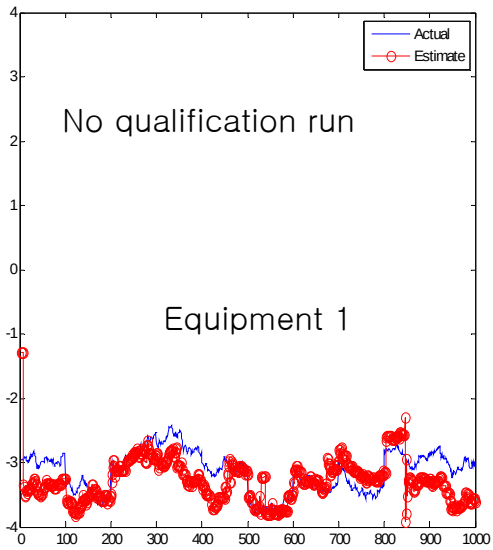
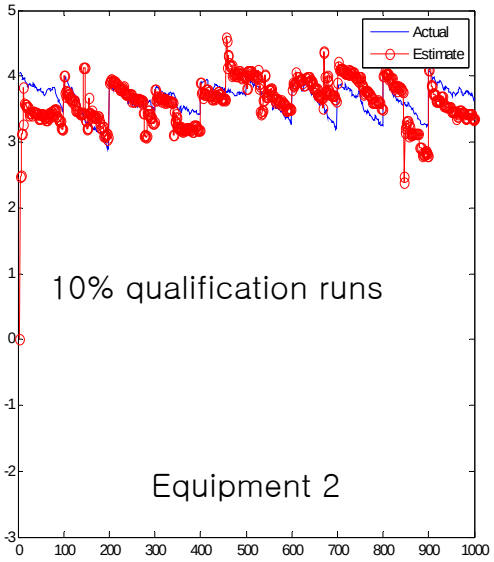


Figure 3.3 Estimation of individual state (equipment 1) – 10% qualification runs vs. No qualification



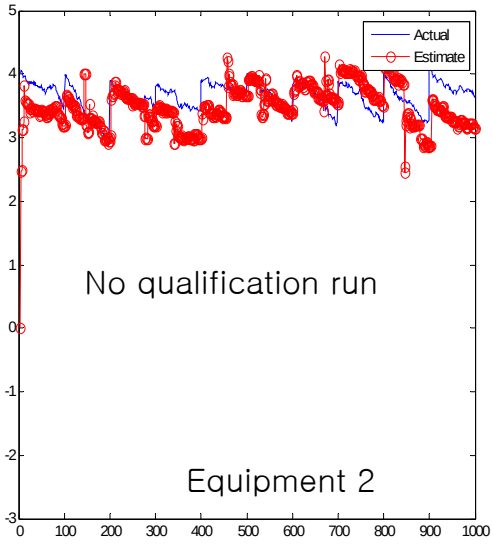


Figure 3.4 Estimation of individual state (equipment 2) – 10% qualification runs vs. No qualification

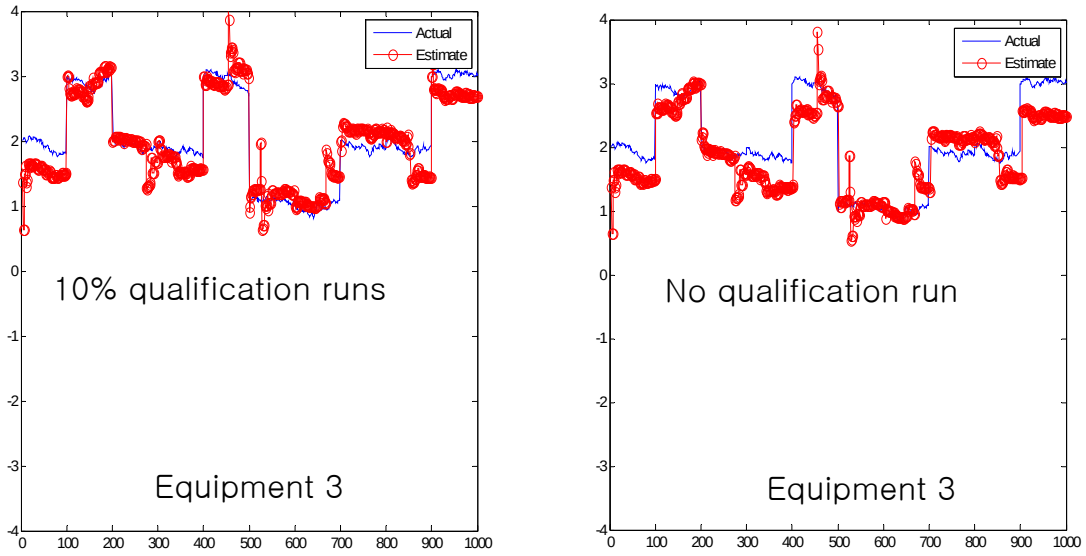


Figure 3.5 Estimation of individual state (equipment 3) – 10% qualification runs vs. No qualification

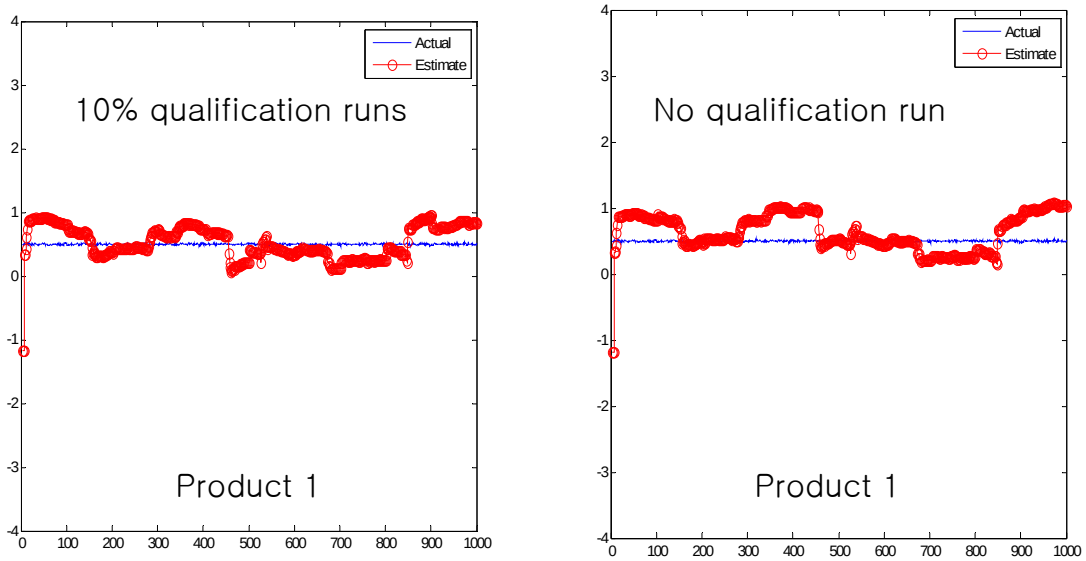


Figure 3.6 Estimation of individual state (product 1) – 10% qualification runs vs. No qualification

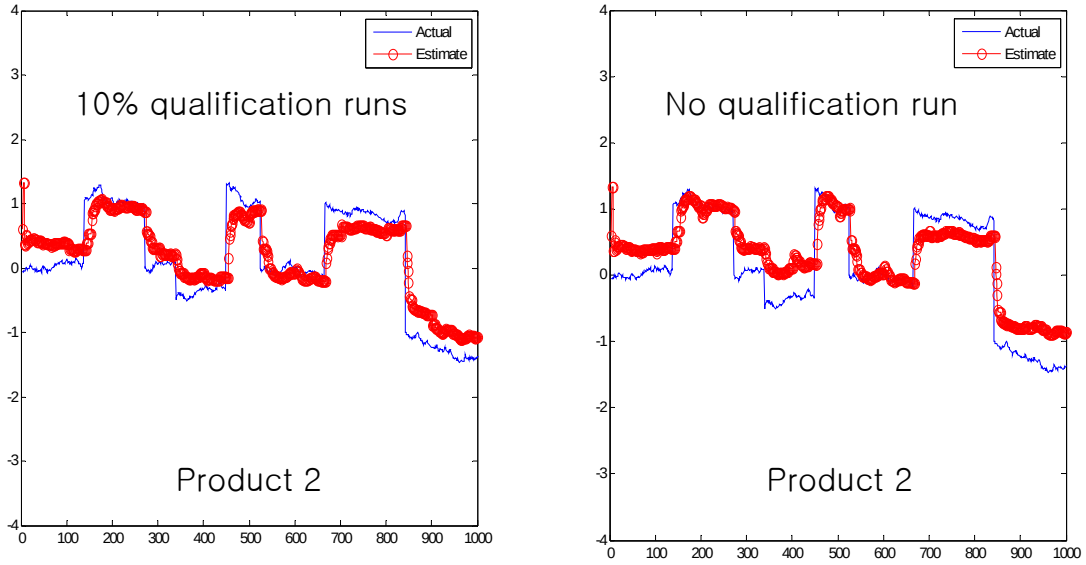


Figure 3.7 Estimation of individual state (product 2) – 10% qualification runs vs. No qualification

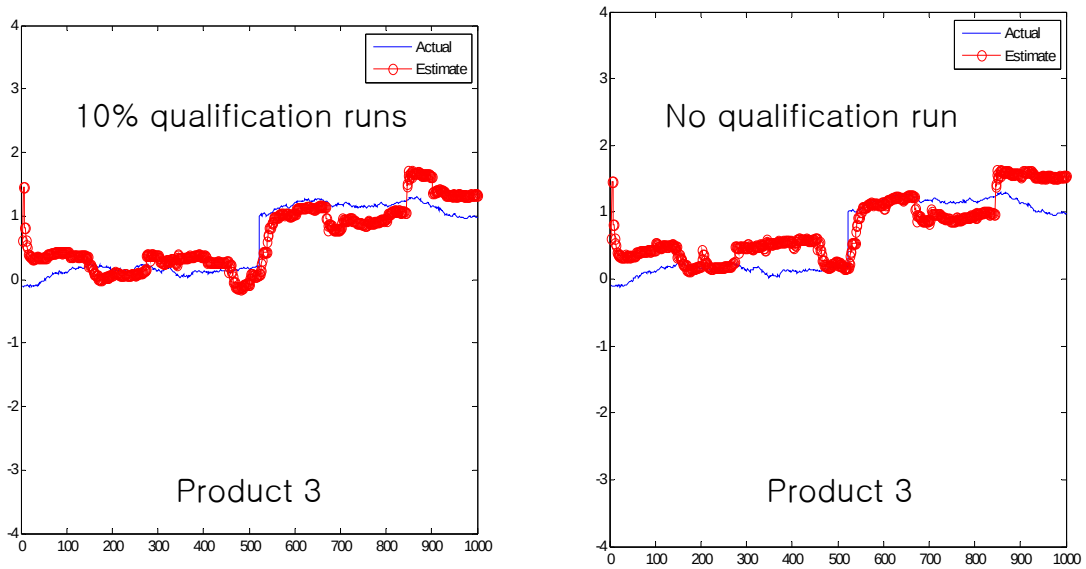


Figure 3.8 Estimation of individual state (product 3) – 10% qualification runs vs. No qualification

Therefore in order to improve control performance, the sampling strategy is still significant rather than just using qualification runs. The next simulation is the extension of online dynamic sampling algorithm to high mix manufacturing environments with qualification runs. Since the qualification runs can measure the equipment biases directly, dynamic sampling is more focused on identifying the product biases. Based on the assumption for non-threaded control, each product bias does not interact with the individual equipment bias. It means that each product bias is independent so maintained even though it runs with different equipment from the previous run. Thus, if step disturbances have occurred and change the product bias, the changed bias states of that product would be remain constant until the another step disturbance happens Therefore online dynamic sampling is implemented for each product separately regardless of the equipments they are running. The basic idea is to identify equipment states using the qualification runs and adjust sampling rate when the measurements detect the changes from product states in order to compensate the changes quickly.

For dynamic sampling algorithm, the prior probability, P_0, and the initial process standard deviation, σ, are assumed to be reasonable values. If the left hand side in equation (3.18) is greater than 0.6, dynamic sampling is turned on to measure every run, and if it is less than 0.6 but greater than 0.3, measurement are taken every two runs.

$$\frac{P_0}{P_0 + (1 - P_0)\exp[-\frac{1}{2}(\frac{x_1}{\sigma})^2]} \geq C \tag{3.18}$$

Otherwise, the sampling rate is kept at the baseline rate – measure every three runs. The standard deviation updates with five point-moving windows but excludes the points if dynamic sampling turns on to measure every run or every two runs. Those algorithm

settings are applied to all product contexts individually. The qualification runs are performed every 100 runs like previous simulations.

Figure 3.9 shows the simulation results of dynamic sampling and uniform sampling with measuring every three runs. Forty three more measurements improve control performance significantly. Table 3.2 has the result of dynamic sampling and uniform sampling with and without qualification runs.

Sampling	**Qualification runs**	**# of measurements**	***Cpk***
Dynamic	10%	383	2.1274
	No	385	1.8467
Uniform Every 3 runs	10%	340	1.4748
	No	334	1.4018

Table 3.2 Simulation result of dynamic sampling and uniform sampling with and without qualification runs

As shown in table 3.2, dynamic sampling improves control performance with no qualification runs compared to uniform sampling. However, the performance difference between 10% qualification runs and no qualification is much larger in dynamic sampling than uniform sampling. It means dynamic sampling has a synergetic effect to qualification runs for improving control performance. Unfortunately the applications of this online dynamic sampling algorithm together with qualification runs were not tested with industrial data sets in this chapter. The industrial data from Texas Instruments Inc. in other chapters might not be good examples since there are interactions between the equipments and products so a threaded approach is required. Therefore testing this method with appropriate data set is a recommendation for future work.

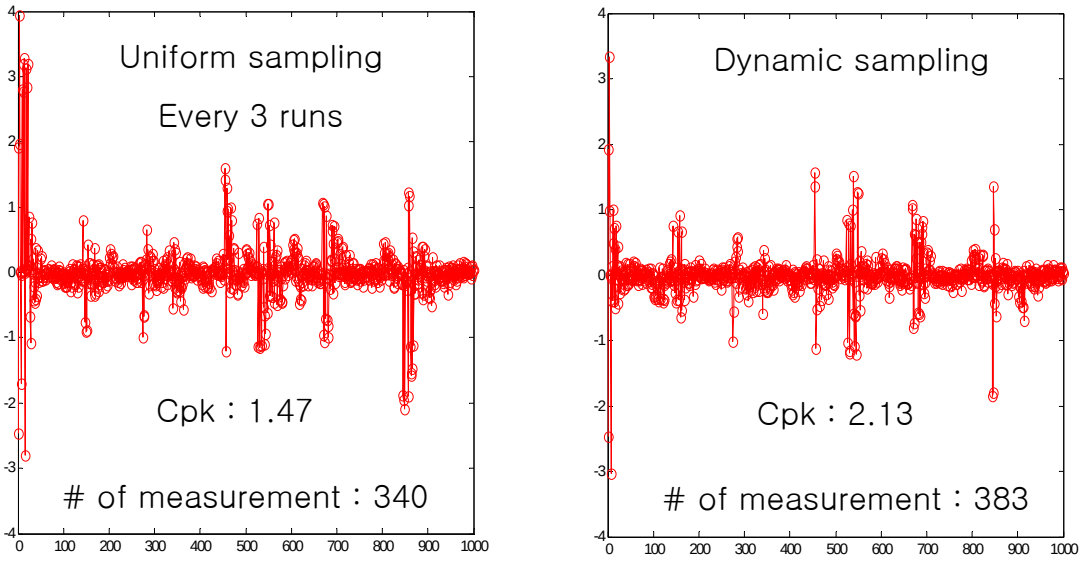


Figure 3.9 Simulation results of dynamic sampling vs. uniform sampling with measuring every three runs

3.5 SUMMARY AND CONCLUSION

In semiconductor manufacturing it is becoming more common to make many different products under multiple process environments ("high mix manufacturing") because of the high capital costs associated with the tools and the limited capacity of the facility. The development of an optimal measurement sampling strategy under high mix manufacturing is challenging since the quality of control must be balanced with the cost of measurement. Thus, it is required to optimize the sampling plan in order to quickly identify the sources of prediction errors and decrease the metrology cost and cycle time.

The term processing context describes the combination of equipment, product, and other factors that identify the processing environment for a particular run. While many factors have an effect on the outputs of a run, the processing equipment and the product are assumed to be the main sources of output variations that have measurable effects on the run.

The observability matrix can be used to determine whether the system is observable or not. If the system is not observable, additional constraints are required to make the system observable. One option is to add supplementary measurements to the system via qualification runs, where the equipment bias can be measured directly during the test procedure on particular equipment.

The online dynamic sampling algorithm from chapter 2 is extended for high-mix manufacturing environments with non-threaded state estimations using Kalman filtering and qualification runs. Test data signals with three equipments and three products are created by emulating the characteristics of actual manufacturing data. Each context has its own bias and dynamic characteristics with a different magnitude of noise.

According to the simulation results, dynamic sampling has a synergistic effect with qualification runs for improving control performance. The application of this

online dynamic sampling algorithm together with qualification runs was not tested with industrial data sets in this chapter. Therefore testing this method with appropriate manufacturing data set is recommended for future work.

CHAPTER 4

Advanced Feedback Control for Drift Cancellation

4.1 BACKGROUND AND MOTIVATION

Semiconductor manufacturing is characterized by a dynamic, varying environment. In this case, dynamic means that the process / product result is a function of time. It also signifies that the equipment variables change or that the target varies during the process step or between process steps. These changes act as disturbances and affect device yield and fab performance. Thus they should be compensated for by using an appropriate process control strategy.

Generally, there are two major disturbances in semiconductor manufacturing: gradual drift and abrupt shift (step disturbance). Each unit process has its own source of disturbance. For example, in a deposition process, the reactor walls may become fouled by deposition as many products are processed [1]. This slow drift in the reactor chamber state requires small changes to the batch recipe in order to make sure that the product outputs maintain on target. Eventually, the reactor chamber will be cleaned to remove the wall deposition, effectively causing a step disturbance to the process. Table 4.1 shows common sources of disturbances and drift for an etch process.

An exponentially weighted moving average (EWMA) filter is widely used to compensate for disturbances in semiconductor manufacturing. It shows good performance for a step disturbance but not as good for a gradual drift. Thus, an advanced feedback control method that treats drift as well as step disturbance is investigated in this chapter.

Cycle	Type	Cause	System
Maintenance Cycle-to-Cycle	Disturbance	Repairs, chamber cleans, PM, kit replacements, gas flow change, lamp change (IM)	Process or Metrology
Within a Maintenance Cycle	Drift	Gradual build-up on chamber, machine wear, sensor drift	Process or Metrology
Lot-to-Lot	Disturbance	Different incoming wafer state, due to current process (other lots processes run between lots of this recipe), control model mismatch, measurement model mismatch	Process
Within a Lot	Drift	1st wafer effect, warm-up, degassing, different steady state chamber condition	Process
Wafer-to-Wafer	Disturbance	Incoming wafer state, different pre-process chambers, litho module	Process
Within a Wafer, Within a Region	Drift	Changes in materials, local heating effects	Process
Within a Wafer, Region-to-Region	Disturbance	Different processing or material exposed	Process

Table 4.1 Common sources of disturbance and drift for an etch process

4.2 ADVANCED FEEDBACK CONTROL METHODOLOGY

4.2.1 Bayesian enhanced EWMA

Bayesian EWMA (B-EWMA) was first proposed by Wang and He [46] to improve state estimation performance with Bayesian detection. Bayesian EWMA enhances the control performance for a process with disturbances because detection and classification for disturbances precedes all control actions. Moreover, B-EWMA takes different actions to handle different types of disturbances. Observed disturbances may be classified as normal states, impulse disturbances, or step disturbances [46][80][81]. Step disturbances can be further classified as a sustained step or a short step, based on its duration. For instance, short-step disturbances can give continuous offsets with a duration of two, three, or four sample points. The detailed disturbance classification is shown in Figure 4.1. Wang and He have applied Bayesian EWMA to the off-line scenario in the chemical mechanical polishing (CMP) process and the plasma etch process. B-EWMA can be modified to apply to real-time on-line scenarios which occur in industrial application.

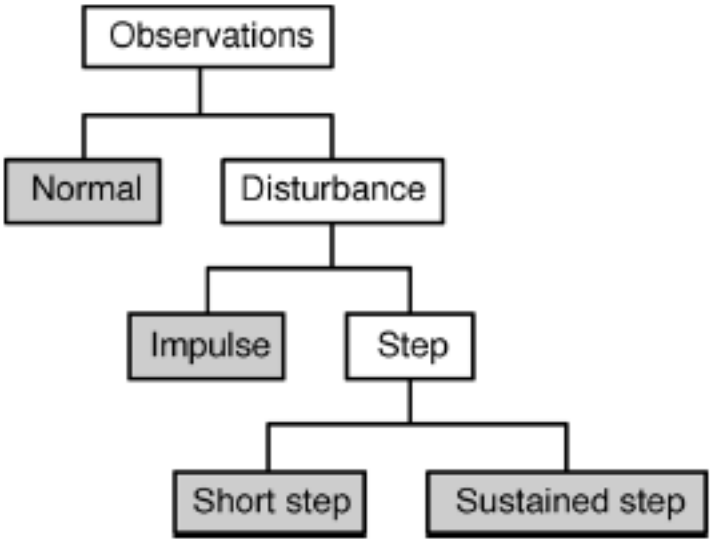


Figure 4.1 Classification of observed states

The Bayesian theorem [48][49][50] is the basic principle in Bayesian detection. Box and Tiao originally devised the Bayesian robust method for data that deviates moderately from the normal distribution model [82]. A normal distribution model can be used to describe a random process [83]. Thus, the Bayesian method is based on the assumption of random process and uses existing process information by assigning proper prior distributions.

Suppose we are interested in estimating the values of a set of parameters Θ For some data set $\boldsymbol{D}$ in some underlying model of the data. For any given model, one may write down an expression for the likelihood function $P(\boldsymbol{D} \mid \Theta)$ of obtaining the data vector $\boldsymbol{D}$ given a particular set of values for the parameter, Θ. In addition to the likelihood function, one may impose a prior distribution $P(\Theta)$ on the parameters, which represents our state of knowledge regarding the values of the parameters before analyzing the data $\boldsymbol{D}$. By Bayes' theorem, the posterior probability is then calculated as:

$$P(\Theta \mid D) = \frac{P(D \mid \Theta)P(\Theta)}{P(D)} \tag{4.1}$$

which gives the posterior distribution $P(\Theta \mid \boldsymbol{D})$ in terms of the likelihood, the prior and the evidence $P(\boldsymbol{D})$ [51]. Consequently, the posterior probability can be simply explained as the following:

$$\text{Posterior Probability} = \frac{\text{likelihood} \times \text{Prior}}{\text{evidence}} \tag{4.2}$$

In the estimation or prediction of the state of a process, the Bayesian method employs both the evidence contained in the observation signals and the accumulated prior probability of the process. Thus, the Bayesian estimation usually does not give a significant error in prediction [46][84].

Wang and He proposed predefined data windows for detecting various types of disturbances [46]. The term "pre-change data" denotes the samples prior to the pre-assumed disturbance onset and "post-change data" denotes the samples after and including the pre-assumed disturbance onset. A directly measured observation which has just gone beyond a certain threshold is defined as onset location. After any onset location is determined, pre- and post-change windows are built as shown in Figure 4.2. A pre-change window is a vector consisting of several past data points prior to the onset location. A post-change window consists of several data points, one at the onset location and some points right after the onset location.

The state space is divided into two subspaces, one for normal subspace Θ_N, and the other for disturbance subspace Θ_D. Then, the prior probability of disturbance $P(\Theta_D \mid X_k)$ is calculated from the post-change window vector X_k including the transient onset point at run *k*.

$$\begin{aligned} P(\Theta_D \mid X_k) &= \frac{P(X_k \mid \Theta_D)P(\Theta_D)}{P(D)} \\ &= \frac{P(X_k \mid \Theta_D)P(\Theta_D)}{P(\Theta_D)P(X_k \mid \Theta_D) + (1 - P(\Theta_D))P(X_k \mid \Theta_N)} \end{aligned} \tag{4.3}$$

In order to detect step disturbances, the posterior probability, $P(\Theta_D \mid X_k)$, derived from the post-change window, is required to be greater than a confidence level *C*. The prior probability, $P(\Theta_D)$, and the confidence level, C, can be obtained from historical data.

$$P(\Theta_D \mid x_1) \geq C \tag{4.4}$$

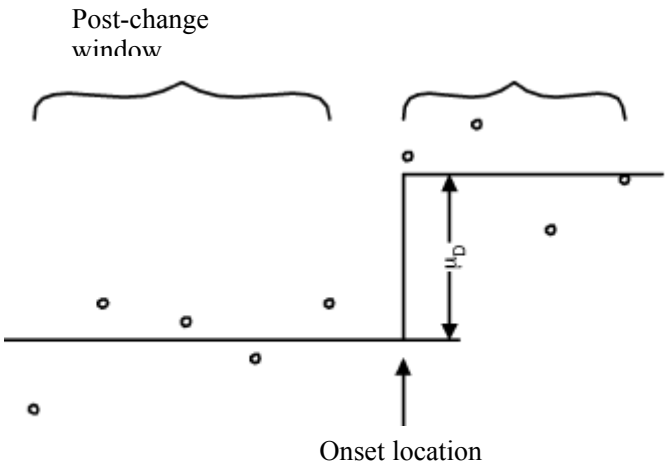


Figure 4.2 Onset location and pre-, post-change window

For a step disturbance, the posterior probability is generated by computing the joint posterior probability for each subset of the post-change window, X_k, where $X_k \equiv \{x_1, x_2, ..., x_k\}$ for k = 1, 2, ... , l_{w_2}. Assuming that the mean of samples in the pre-change window is zero, the step magnitude, μ_D, is calculated as the mean of X_k as in the following equation (4.5).

$$\mu_D = \frac{\sum_{i=1}^{k} x_i}{k} \tag{4.5}$$

If Gaussian distribution is assumed, the probability density function for normal and shifted process states are denoted by $N(0, \sigma^2)$ and $N(\mu_D, \sigma^2)$, respectively, where σ is the process standard deviation. The likelihood function of a step disturbance for a single observation, x_i, is calculated by the equation (4.6):

$$p(x_i \mid \Theta_D) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x_i - \mu_D)^2}{2\sigma^2}] \tag{4.6}$$

Assuming all samples in X_k are independent and identically distributed, the likelihood function for X_k is

$$p(X_k \mid \Theta_D) = \prod_{i=1}^{k} p(x_i \mid \Theta_D) \ . \tag{4.7}$$

By substituting equation (4.6) into (4.7), the function is derived as

$$p(X_k \mid \Theta_D) = \frac{1}{(\sqrt{2\pi\sigma^2})^k} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} (x_i - \mu_D)^2] \ . \tag{4.8}$$

Similarly, when the normal process data is assumed, the likelihood function is,

$$p(X_k \mid \Theta_D) = \frac{1}{(\sqrt{2\pi\sigma^2})^k} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{k} x_i^{\,2}] \ . \tag{4.9}$$

Combining all the equations mentioned above, the joint posterior probability, $P(\Theta_D \mid X_k)$, in equation (4.4), can be obtained as the following equation (4. 10):

$$p(\Theta_D \mid X_k) = \frac{P_0}{P_0 + (1-P_0)\exp[-\frac{(\sum_{i=1}^{k} x_i)^2}{2k\sigma^2}]} \tag{4.10}$$

If the disturbance observed is a step disturbance, the posterior probability of the first point in the post-change window, $P(\Theta_D \mid x_1)$ is

$$p(\Theta_D \mid x_1) = \frac{P_0}{P_0 + (1-P_0)\exp[-\frac{1}{2}(\frac{x_1}{\sigma})^2]} \ . \tag{4.11}$$

For online application of this method, only a step disturbance can be treated in the detection and classification process, not an impulse disturbance. In an online monitoring and controlling system, Bayesian methods are required to update the input for the next run based on the state conditioned by the last runs of historical data. If the last data point is beyond a threshold, a certain type of disturbance must have occurred. But, to classify the disturbance, at least a two run delay is inevitable. However, if the disturbance is an impulse disturbance of short duration, the delay makes a control action unnecessary.

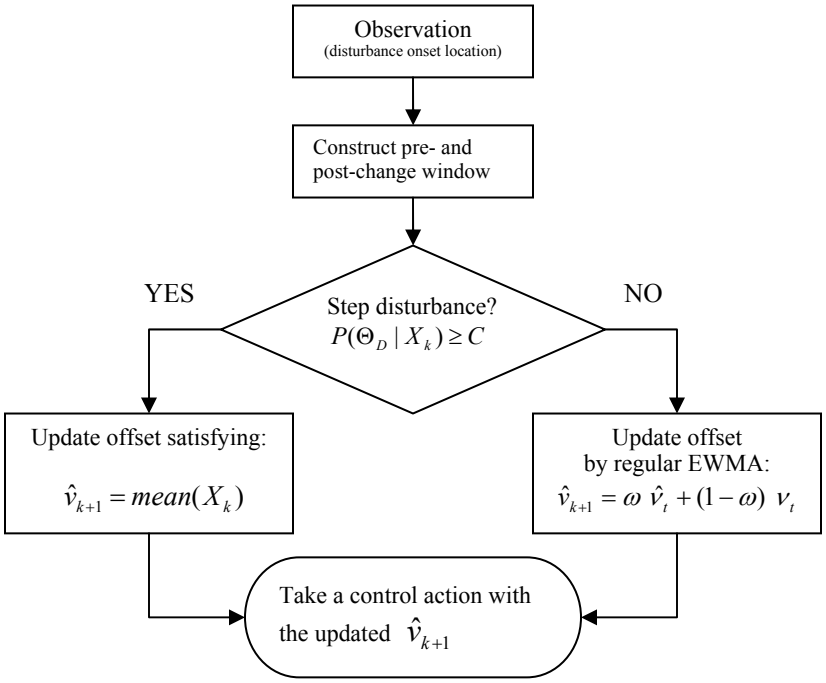


Figure 4.3 Flowchart of Bayesian EWMA

The flowchart of the Bayesian EWMA control method is presented in Figure 4.3. Due to the limitations of an on-line process, the B-EWMA is better for handling step disturbances. Since at least two data points in a post-change window are required to detect and classify disturbances, it is too late to take the control action for impulse disturbances in an online process. That is, it is impractical to compensate for impulse disturbances in real-time reaction control. Thus, in this chapter, the Bayesian method is targeted to control step disturbances in stationary processes, not impulse disturbances.

If the step disturbance is detected, the average of two outputs in a post-change window is recommended for simple and optimal control action to update the estimate for the next run of *k+1*. Otherwise, the regular EWMA method is implemented to control process.

4.2.2 Robust Drift Cancellation

Mussachio [85] first presented the robust drift cancellation method (RDC) as a variation of the Finite-impluse response (FIR) Wiener filter to achieve optimality in control. FIR filters are often preferred because FIR filters are intrinsically stable and they can easily realize various parameters [86][87]. In the RDC method, process variation is assumed to be entirely in the offset term just like the EWMA controller. The input sensitivities or tuning parameters are assumed to be constant and fixed for the entire process. The advantage of the RDC controller is to estimate the offset term, z_k, selecting the optimal tuning parameters based on historical data.

Just like the EWMA method, the RDC method assumes a nominal process model of the form of equation (4.12):

$$y_k = Au_k + z_k \tag{4.12}$$

where y_k is the process output, u_k is the process input, A is the gain between the manipulated input and the measured output, and z_k is an estimate term computed and updated.

The design of RDC starts with computing the residual signal, d_k, including both disturbances and drifts as shown in the following equation (4.13),

$$d_k = y_k - Au_k \tag{4.13}$$

Since the RDC uses the correlation matrix, matrix R and a tuning matrix, K can be obtained as presented in equations (4.14) and (4.15),

$$K > R = \begin{bmatrix} R_0 & R_1 & \cdots R_{n-1} \\ R_1 & R_0 & \cdots R_{n-2} \\ \vdots & \vdots & \ddots \ \vdots \\ R_{n-1} & R_{n-2} & \cdots R_0 \end{bmatrix} \tag{4.14}$$

$$R_l = \frac{\sum_{l=2}^{m} (d_k \cdot d_{k-l})}{m-1} \tag{4.15}$$

where m is the number of data record, and n is a window size $n \times n$ for the correlation matrix.

The size of n can be any number greater than 2. This is discussed later as one of the design issues in RDC, because performances are changed by the size of n. It has not been proposed how to choose the optimal value of n. According to the theory, however,

an n too small in size would not be sufficient to compensate; an n too large in size would be overly conservative [88]. The matrix K satisfying the equation above is called a measure of the worst case covariance of the residual. That is, it is a matrix upper bound of R.

The other parameter which needs to be found is a vector, C, satisfying the following conditions,

$$C \cdot K^{-1} \cdot C^T < L \cdot K^{-1} \cdot L^T \tag{4.16}$$

where

$$L = [R_1 \quad R_2 \quad \cdots R_n] \tag{4.17}$$

The row matrix C is also called a measure of the worst-case autocorrelation in the measured residuals. Then, the estimate input, u_k can be updated by equation (4.18).

$$Au_k = -(C)(K)^{-1}\begin{bmatrix} (Y_{k-1} - Au_{k-1}) \\ \vdots \\ (Y_{k-n-1} - Au_{k-n-1}) \end{bmatrix} = -(C)(K)^{-1}\begin{bmatrix} d_{k-1} \\ d_{k-2} \\ \vdots \\ d_{k-n} \end{bmatrix} \tag{4.18}$$

In general, the formulation of the RDC method results in a set of linear equations and has a closed-form solution which is stable and practical. It is robust to non-stationarity in the drift statistics. However, the main drawback of the RDC method is that it needs two sensitive coefficients to approximate a desired response and is theoretically expensive. Figure 4.4 describes the flow chart of RDC process, assuming two as the number of taps, n.

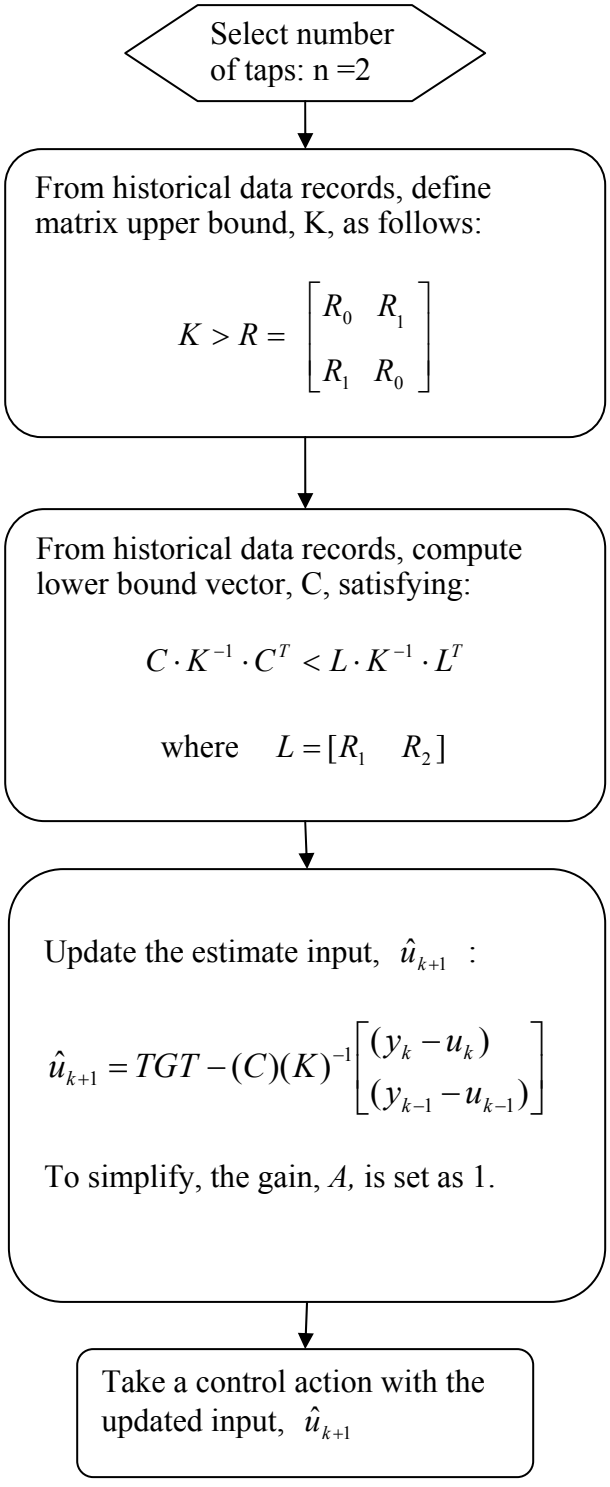


Figure 4.4 Robust drift cancellation (RDC) design procedure

4.2.3 Modified RDC with Bayesian EWMA

It is often the case in semiconductor manufacturing that tools drift too rapidly from one run to the next for a process disturbance to be considered locally constant [89]. Thus, in this chapter, two distinct but isolated problems, drifts and step disturbances, are simultaneously considered for robust control. By the Bayesian detection system, observed states are classified into two categories: one for normal states with drift, the other for step disturbances. Then, the classified drifts and step disturbances are controlled using different methods. Consequently, the modified robust drift cancellation (RDC) method takes actions to compensate for states having process drifts or normal noises, while the Bayesian EWMA (B-EWMA) method is performed to control step disturbances.

RDC is able to achieve optimality for a larger range of problems having offset drift as well as process drift. Since it is designed from historical data, its performance is entirely dependent on the past data record, which is used to calculate two significant parameters in the control. Therefore, RDC is much better for adjusting offset drifts than any other method, however, RDC has worse performance for normal step disturbances than EWMA. Thus, RDC requires modification to enhance its performance for step disturbance as well as states with drift.

Unlike the regular RDC, the modified RDC method is based on the RDC control law and the offset estimate updating by EWMA. First, the modified RDC observer assumes that the process output, y_k, is the linear combination of the process inputs, u_k, the process offset, v_k, and a measurement noise, e_k.

$$y_k = Au_k + v_k + e_k \tag{4.19}$$

where A is the gain between the manipulated input and the measured output.

To simplify the equation, it is also assumed that A is one and e_t is zero. Accordingly, the form is simplified as the following:

$$y_k = u_k + v_k \tag{4.20}$$

The input u_k is updated based on the control law of the RDC method as discussed in section 4.2.2. The residual signal, d_k, shown in equation (4.13) is replaced by the estimated offsets $\hat{v}_k$. Therefore, the input is computed by the FIR Wiener Filter, using the following matrices:

$$u_k = TGT - (C)(K)^{-1}\begin{bmatrix} \hat{v}_k \\ \hat{v}_{k-1} \end{bmatrix} \tag{4.21}$$

where the row matrix C is the worst-case average autocorrelation of the offsets, the matrix K is the worst-case covariance of the offsets, and TGT is the target.

In EWMA, the predicted offset, $\hat{v}_{k+1}$ is updated by the previous predicted estimate, $\hat{v}_k$, and the measured offset, v_k, as in the following equation (4.22):

$$\hat{v}_{k+1} = \omega\ \hat{v}_k + (1-\omega)\ v_k \tag{4.22}$$

Since the modified RDC method updates the input, u_k, prior to updating the offset, $\hat{v}_k$, the measured offset, v_k, is replaced by an error between the measured output and the predicted output in the equation (4.23):

$$\hat{v}_{k+1} = \omega\ \hat{v}_k + (1-\omega)\ (y_k - (u_k + \hat{v}_k)) \tag{4.23}$$

Alternately, when step disturbances are detected, the Bayesian EWMA controller is activated. The Bayesian detection procedure is already discussed in section 4.2.1. After step disturbances are observed by the Bayesian detection, the B-EWMA is implemented in the control action. The B-EWMA is based on regular EWMA [8]. The only difference between B-EWMA and EWMA is in the estimation of the offset, $\hat{v}_k$. While EWMA computes the minimum mean squared error for the estimate, B-EWMA simply takes the mean value of some past outputs in the post-change window in the following form in order to treat the step disturbance faster:

$$\hat{v}_{k+1} = mean(X_k). \tag{4.24}$$

Then, the regular EWMA control law is used to determine the input for the next run. The input is determined through a simple linear model:

$$u_{k+1} = TGT - \hat{v}_{k+1} \tag{4.25}$$

Figure 4.5 shows a flowchart of the modified RDC with Bayesian detection.

As discussed, in this controller, process drifts and step disturbances are controlled by two different methods: the modified RDC and Bayesian EWMA. These two methods are driven by two different ways to determine the input and to estimate the offset for the following run. This controller is unique in that it compensates for states with drift as well as step disturbance. However, the controller should be well-tuned because the performance of the controller is strongly dependent on tuning parameters, which should be properly defined prior to manufacturing application.

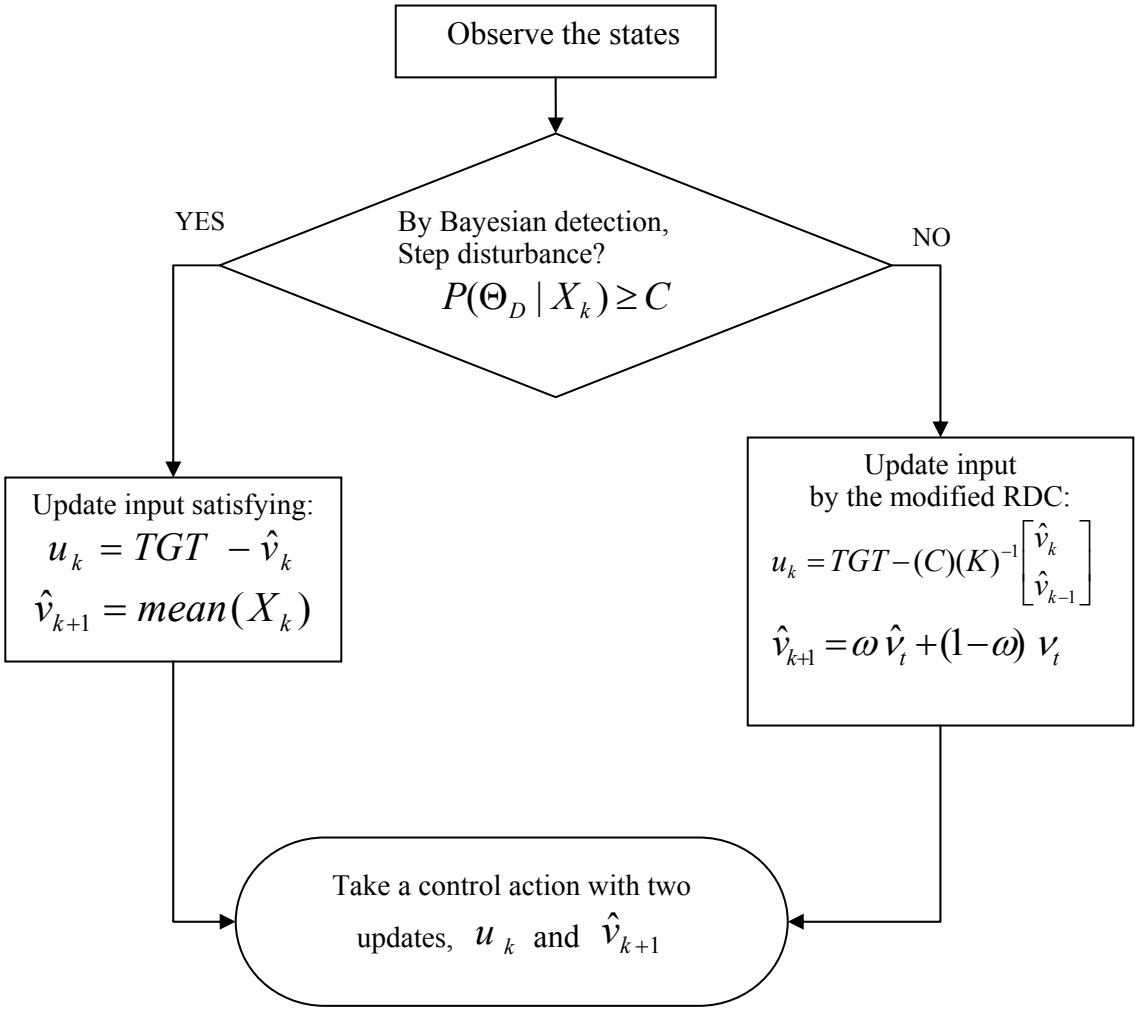


Figure 4.5 Flowchart of the modified RDC with Bayesian Detection

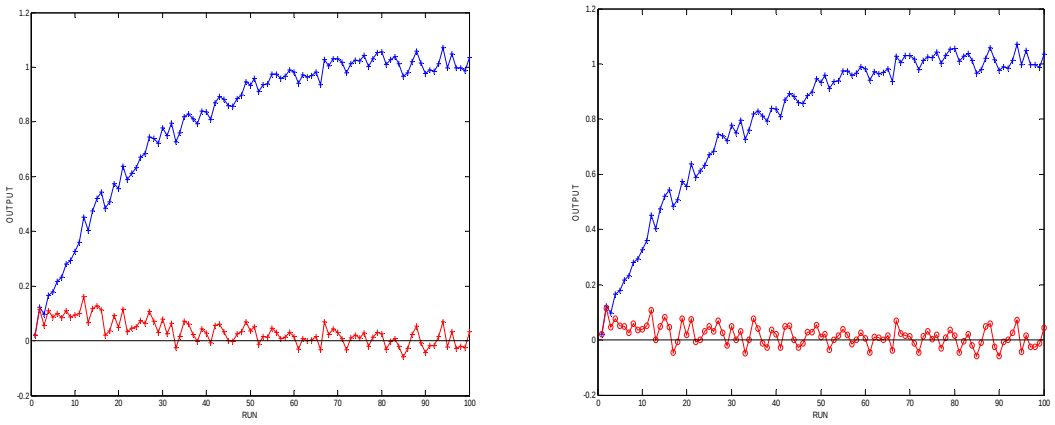
4.3 OPTIMAL TUNING ISSUES FOR MODIFIED RDC WITH BAYESIAN DETECTION

4.3.1 Tuning for RDC

The design of the modified RDC requires the computation of a matrix upper bound of $(C)(K)^{-1}$ in equation 4. 18. Computation of this upper bound is nontrivial, because one would like to find a least upper bound; an upper bound that is too large would lead to an overly conservative controller with a low value for the estimate input in equation 4.18 [90]. Fortunately, the set of matrices that satisfy the condition in equation 4.14 form a convex set, making this problem amenable to numerical optimization [91][92]. The multiplication of the worst-case autocorrelation vector, *C*, and the inverse of the worst case matrix bound, *K*, can be described as the simple vector form in Equation 4.26 by assuming the window size of two for the autocorrelation matrix.

$$u_k = TGT - (C)(K)^{-1}\begin{bmatrix}\hat{v}_k \\ \hat{v}_{k-1}\end{bmatrix} = [M_1 \quad M_2]\cdot\begin{bmatrix}\hat{v}_k \\ \hat{v}_{k-1}\end{bmatrix} \tag{4.26}$$

According to the simulation results in Figure 4.6, the performance of the modified RDC with Bayesian EWMA is similar to that of EWMA when the sum of M_1 and M_2 are 1. However, when the sum of M_1 and M_2 are greater than 1, such as [0.7 0.4] , the performance is improved no matter which *M* is greater. To verify the better performance of this controller, three sample datasets having different drifting slope in Figure 4.7 were applied to be controlled by using the upper bound, [0.7 0.4]. As shown in Table 4.2, *Cpk* values of the modified RDC are greater than those of EWMA in every case.



C_{pk}=1.1498

$(C)(K)^{-1} = [0.7 \quad 0.4]$

C_{pk}=1.3149

Figure 4.6 Simulation of EWMA vs. the modified RDC with Bayesian EWMA

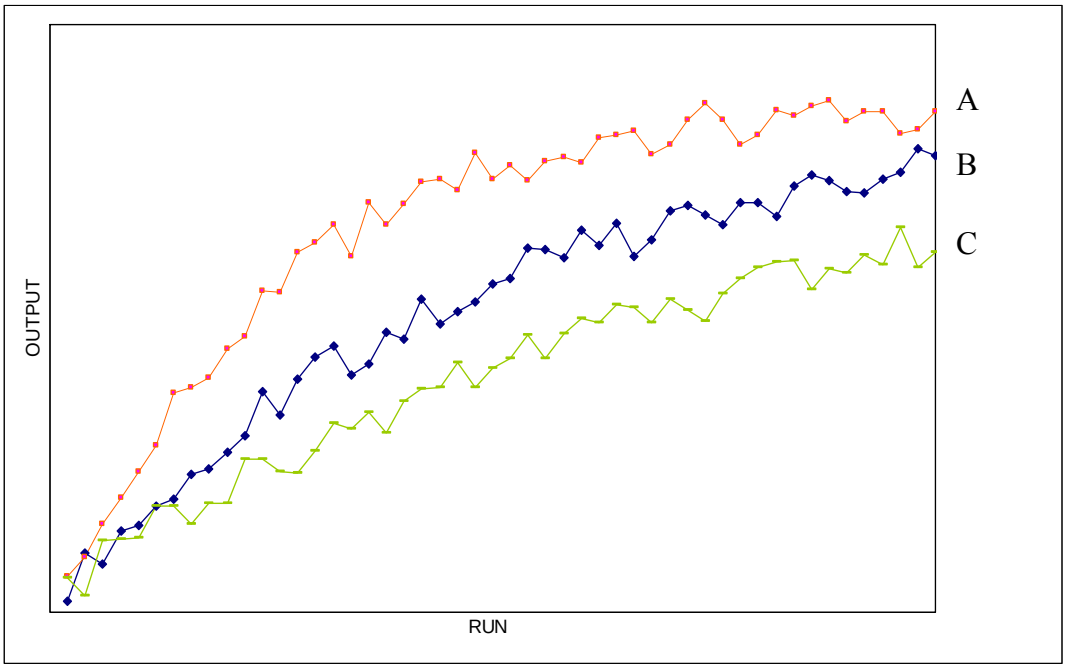


Figure 4.7 Uncontrolled drifting datasets for simulation

In addition, it is observed that the performance of modified RDC improves with the steeper slope of drifts as expected. That is because the sum of M's greater than 1 compensates faster. However, excessively high sums of M_1 and M_2 should be avoided because the control action would be too aggressive, causing the controller to overcompensate for offsets. Further, it is appropriate to set the value of M_1 greater than M_2, because it means more recent data will affect updating the offset.

C_{pk}	EWMA	Modified RDC with Bayesian Detection
A	1.2621	1.3149
B	1.3120	1.3474
C	1.0160	1.0848

Table 4.2 Simulation results of modified RDC with Bayesian detection with three different drifting states

4.3.2 Tuning for EWMA

In this method, process variation is assumed to be entirely in the offset term. The idea is to estimate the current offset term, $\hat{v}_k$, and select an input setting to compensate for the offset. The weighting factor, w, takes a value between zero and one [93]. When the weight, w, is chosen to be close to one, the estimate of $\hat{v}_{k+1}$ is, theoretically, updated very quickly by responding to process error (target - output) quickly. When the weight, w is zero, the prior process measurement has no effect on estimating the offset, $\hat{v}_{k+1}$. To demonstrate the importance of choosing the tuning parameter, w, simulations were made with three different tuning cases, $w = 0.1$, 0.5, and 0.8. It is clear from the simulation shown in Figure 4.8 that it is important to consider

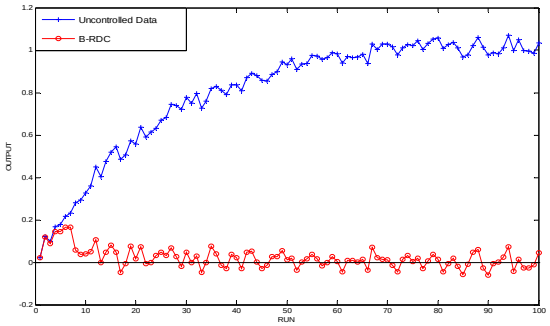
the tuning of w for a response for process variation. In this case, the larger the tuning weight is, the better the performance is to compensate for the states having drifts.

The optimal tuning needs to be chosen based on the desired properties of the data set. The small tuning value would be suitable for states with small deterministic drifts. In contrast, the states having significantly drifting offsets would be better controlled by using higher values of the weighting factor. Note that if a higher weighting factor is chosen, the system might have large oscillations because it is sensitive to noise. When it is hard to estimate the characteristics of states, it is typically safe to take the value of tuning factor, w, between 0.3 to 0.7.

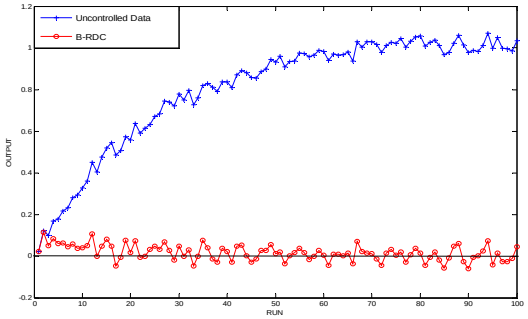
4.3.3 Tuning for Bayesian Detection

The confidence level, C, and the prior probability, P_0, in Figure 4.5 gauges the magnitude of step disturbance that will be detected. To illustrate this, the relationship between the prior probability, P_0, in equation 4.11 and the confidence level, C, is important. The joint posterior probability calculated using equation 4.11 is sensitive to step disturbances. By assuming the prior probability, P_0, for a step disturbance to occur is 0.1, C = 80% will detect a step disturbance with magnitude μ_D=2.667σ, while setting C = 95% will only detect a disturbance with magnitude μ_D=3.207σ.

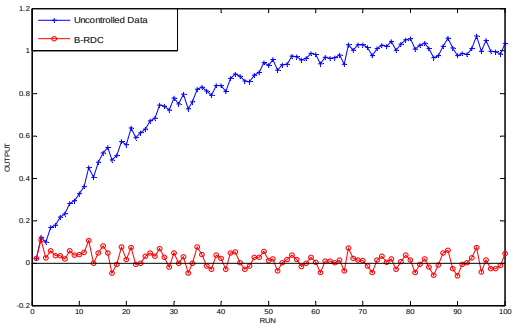
For example, there is a dataset which has joint post probabilities in the pre-change window like {0.501, 0.508, 0.516 0.503 0.520}, while {0.908, 0.978, 0.999} is used in the post-change window by selecting the prior probability, P_0 = 0.5.



w=0.1, C_{pk} = 1.1607



w=0.5, C_{pk} = 1.2914



w=0.8, C_{pk} = 1. 3301

Figure 4.8 Simulations of the modified RDC with Bayesian EWMA with w = 0.1, 0.5, and 0.8

If we choose C = 0.99, the step disturbance would be detected only at the third point in the post-change window, having $P(\Theta_D \mid x_1)$=0.99. Likewise, C = 0.95 would classify the step disturbance with the second point, having $P(\Theta_D \mid x_1)$=0.978; C = 0.90 would recognize the step disturbance from the first point with $P(\Theta_D \mid x_1)$ =0.908. Accordingly, it is easy to verify the criterion $P(\Theta_D \mid X_k) \geq C$ to detect step disturbances with a confidence level, C.

4.4 SIMULATION RESULTS

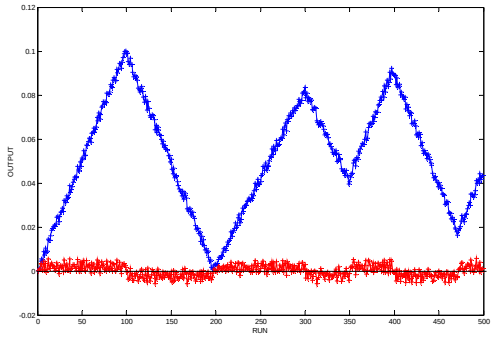
4.4.1 Simple Drift

To investigate the performance of the modified RDC with Bayesian detection method, a simple drifting dataset was applied to EWMA, RDC, the modified RDC, and the modified RDC with Bayesian detection method as shown in Figure 4.9. The weighting factor, w, is set as 0.5 for all the methods. The matrix upper bound, $(C)(K)^{-1}$ for RDC methods is taken as [0.7 0.4] as discussed in section 4.3.1.

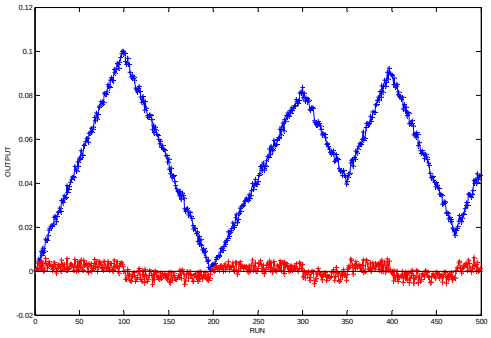
According to the results in Table 4.3, the modified RDC and modified RDC with Bayesian detection methods performed the best. Since the dataset only contains the drift states without step disturbances, Bayesian detection has no effect on this case. Thus, the performances of modified RDC and modified RDC with Bayesian detection methods are exactly the same. The EWMA performed slightly worse than the modified RDC; the regular RDC method was the worst.

	EWMA	RDC	Modified RDC	Modified RDC with Bayesian Detection
C_{pk}	1.0450	0.9674	1.1856	1.1856

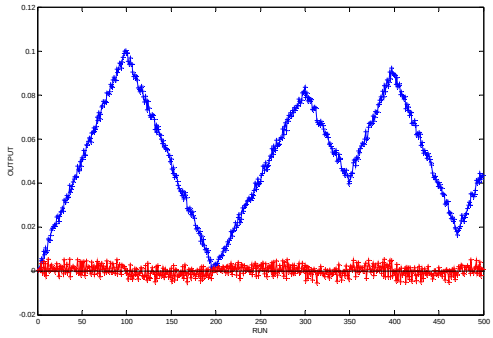
Table 4.3 C_{pk} values of simulation results with simple drift



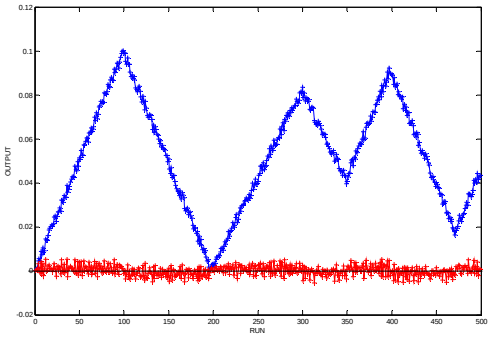
EWMA (w=0.5)



RDC ($(C)(K)^{-1}$ = [0.7 0.4])



modified RDC
($(C)(K)^{-1}$ = [0.7 0.4], w = 0.5)



modified RDC +Bayesian Detection
($(C)(K)^{-1}$ = [0.7 0.4], w = 0.5)

Figure 4.9 Simulations of EWMA, RDC, modified RDC, the modified RDC with Bayesian EWMA

4.4.2 Simple Step Disturbance

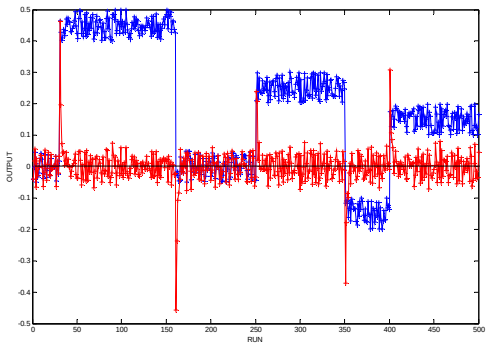
In this case, the dataset with simple step disturbance was used to compare the performance of four methods: EWMA, RDC, the modified RDC, and the modified RDC with Bayesian detection. As shown in Table 4.4, the modified RDC with Bayesian detection method is absolutely better than any other method. It is also observed, in Figure 4.10, that the control actions in the modified RDC with Bayesian detection method are faster than the other methods when it detects the step disturbance. Comparing the performance of EWMA and the modified RDC with Bayesian method in Figure 4.11, we can see that the modified RDC with Bayesian method compensates for step disturbances much faster than any other method. In accordance with the simulation results, the modified RDC with Bayesian detection method obviously improves to compensate for step disturbances.

	EWMA	RDC	Modified RDC	Modified RDC with Bayesian Detection
C_{pk}	1.2091	1.0412	1.1681	1.3059

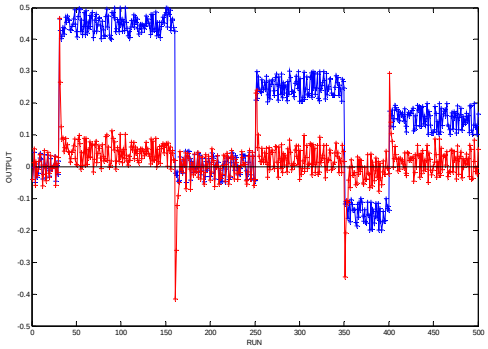
Table 4.4 C_{pk} values of simulation results in simple step disturbances

4.4.3 Industrial Example 1 : Photolithography Process

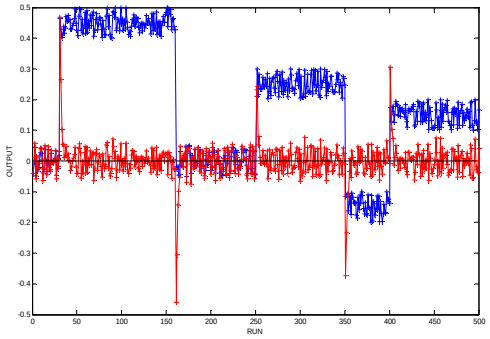
The first industrial example is a photolithography process for shallow trench isolation (STI) of the critical dimension (CD) controller from DMOS6 in Texas Instruments Inc. CD should be controlled to make mask patterns for following etch process with adjusting exposure time. The process model has a linear format as equation (4.27).



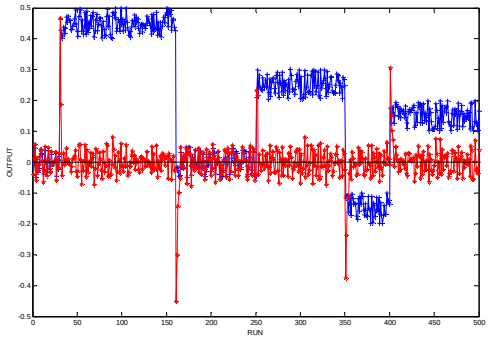
EWMA (w=0.5)



RDC ($(C)(K)^{-1} = [0.7 \quad 0.4]$)

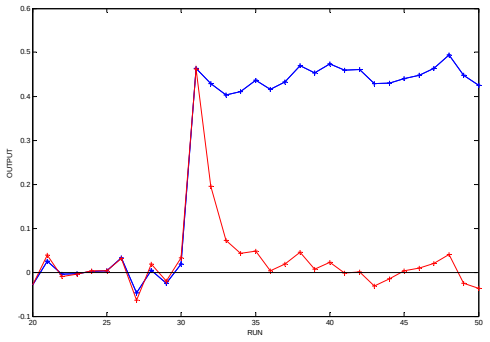


modified RDC
($(C)(K)^{-1} = [0.7 \quad 0.4]$, $w = 0.5$)



modified RDC +Bayesian Detection
($(C)(K)^{-1} = [0.7 \quad 0.4]$, $w = 0.5$)

Figure 4.10 Simulations of EWMA, RDC, modified RDC, the modified RDC with Bayesian EWMA



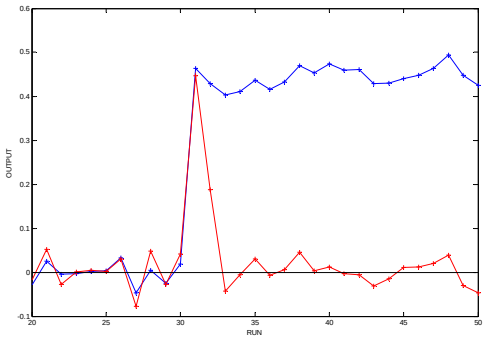


Figure 4.11 Comparison of EWMA, the modified RDC with Bayesian EWMA in Step Disturbance Control

$$y_k = s \times u_k + b_k + c_k \tag{4.27}$$

where y_k is the CD, u_k is the exposure time, s is the slope which is constant, b_k is the reticle offset, which is also constant with the values from the data set, and c_k is the scanner offset .

There are two different lithography tools and two different products in this data set. Thus, four processing threads are considered in the simulations. The simulation result of each thread is shown in Figure 4.12 and Table 4.5. For comparison, the simulation results of EWMA and modified RDC are also shown in Table 4.5. All simulation results are normalized and actual scales are not reported because this is Texas Instrument's proprietary data set.

C_{pk}	EWMA	Modified RDC	Modified RDC With Bayesian Detection
Thread1	1.7861	1.7699	1.7707
Thread2	2.2734	2.2644	2.2188
Thread3	1.5394	1.5114	1.4929
Thread4	1.8007	1.7228	1.9536

Table 4.5 C_{pk} values of simulation results in photolithography process

As shown in Figure 4.12 all threads have an amount of noise rather than drift. Thus, modified RDC does not affect the control performance of each thread. However, it shows a similar performance as EWMA for all cases. Bayesian detection has an excellent performance for step disturbances shown in thread 4 but shows a worse performance for the situations that have a large magnitude of noise (thread 3) or are already well-controlled (thread 1).

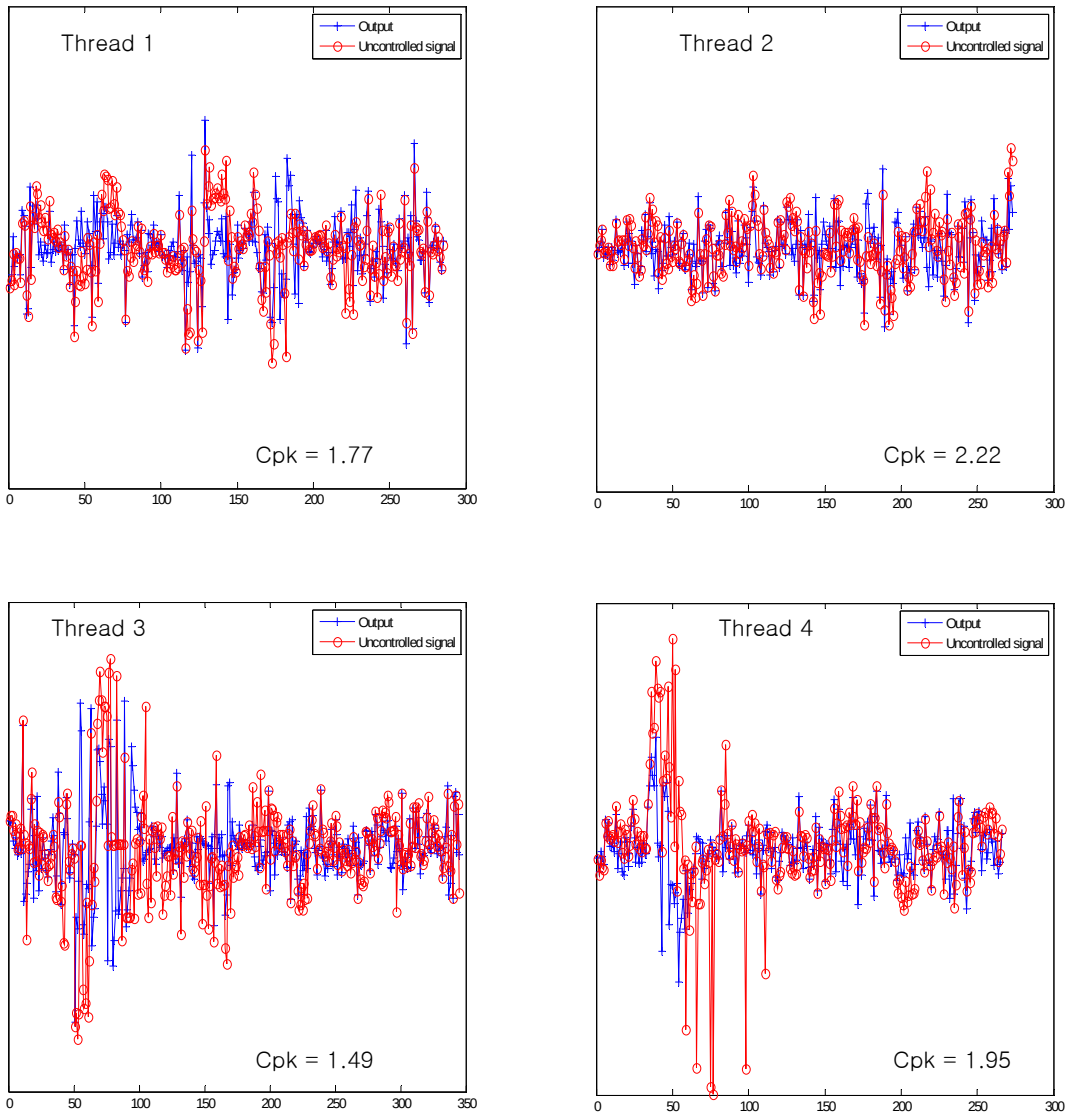


Figure 4.12 Simulation results in photolithography process

4.2.4 Industrial Example 2 : Thin Film Deposition Process

The next industrial data set is a chemical vapor deposition (CVD) process from the DMOS6 wafer fab in Texas Instruments Inc. The goal is deposition thickness control by adjusting deposition time.

A linear process model is used by the run-to-run controller

$$y_k / 4 = rate_k \times u_k + b_k + c_k \tag{4.28}$$

where y_k is the thickness, u_k is the deposition time, b_k is constant and c_k is an offset which needs to be updated run to run. There are two different products and six different machines so a total twelve threads are used for these simulations. The simulation result of each thread is shown in Figures 4.13 to 4.15 and Table 4.6.

C_{pk}	EWMA	Modified RDC (aggressive)	Modified RDC With Bayesian Detection (aggressive)	Modified RDC With Bayesian Detection (regular)
Thread1	0.9746	0.7984	0.8914	1.0371
Thread2	0.3459	0.3603	0.4186	0.4031
Thread3	0.6755	0.5271	0.5471	0.6428
Thread4	0.5250	0.3680	0.4404	0.5503
Thread5	0.6072	0.5817	0.6591	0.6869
Thread6	0.4745	0.3647	0.4318	0.5200
Thread7	1.3627	1.2597	1.2211	1.4071
Thread8	0.2833	0.4479	0.5411	0.3729
Thread9	0.5601	0.4383	0.6402	0.7215
Thread10	0.6623	0.4870	0.5248	0.7261
Thread11	1.0175	0.9297	0.9214	0.9331
Thread12	0.1578	0.1421	0.2222	0.2157

Table 4.6 C_{pk} values of simulation results in CVD process

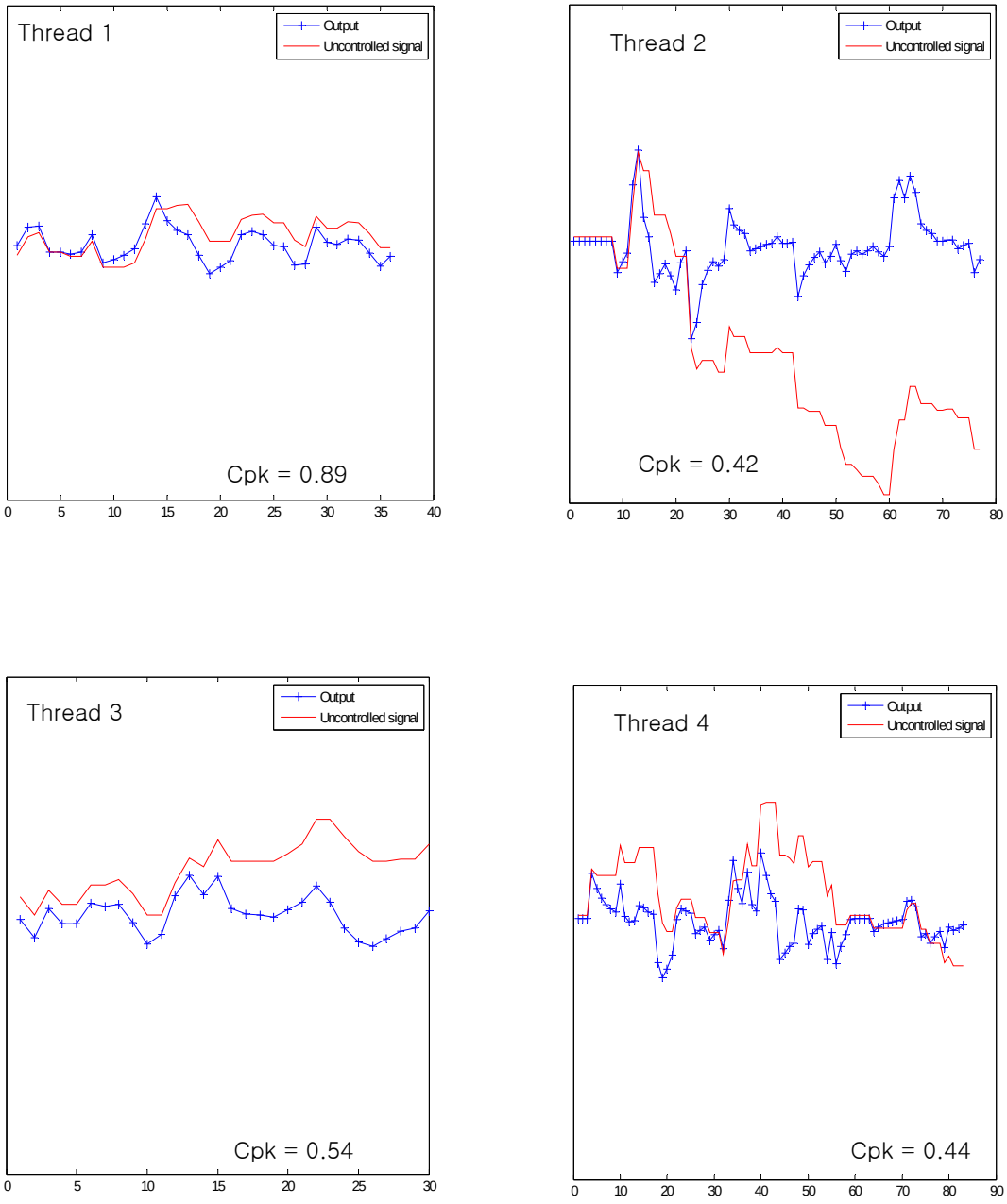


Figure 4.13 Simulation results in CVD process (Thread 1 ~ 4)

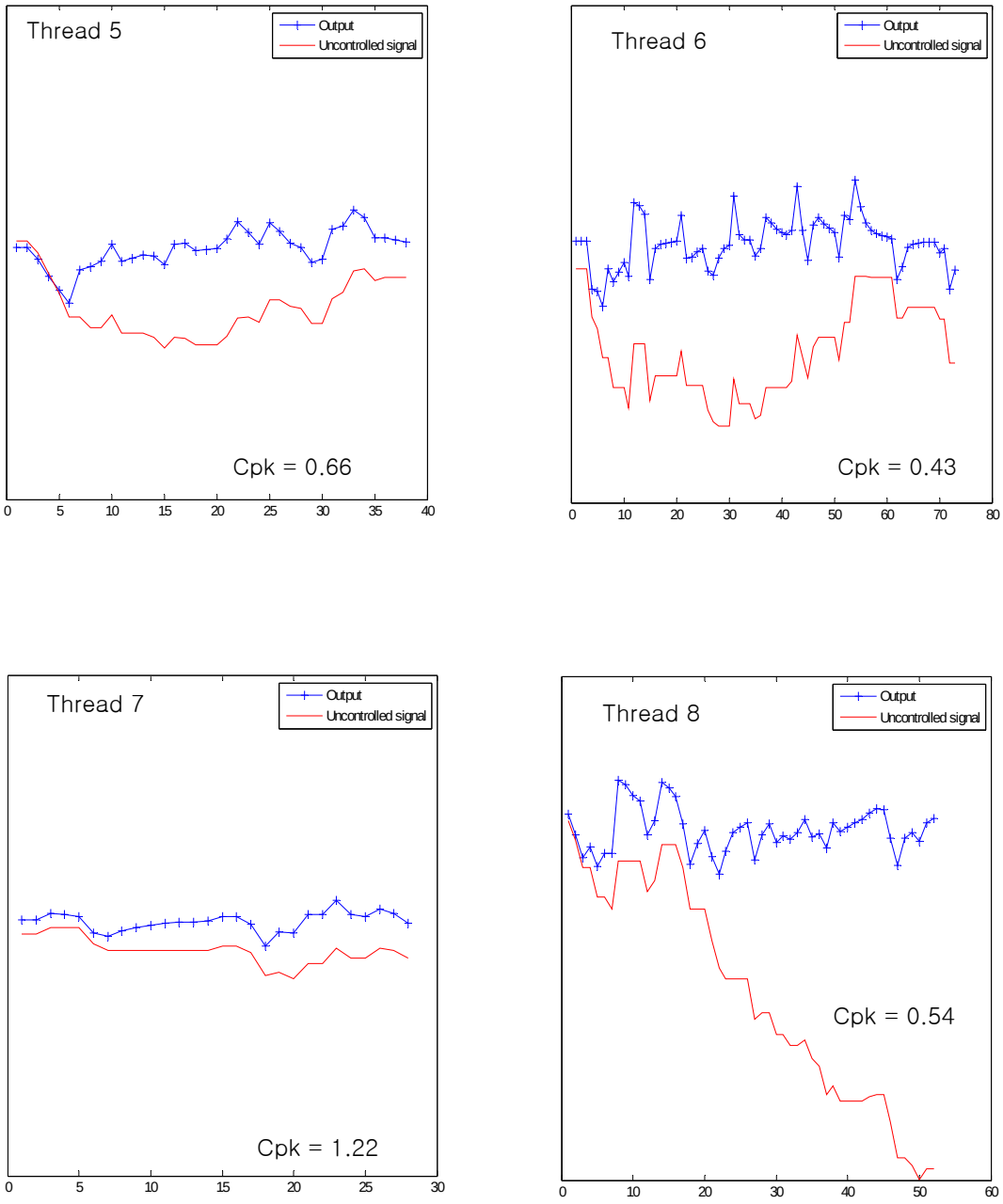


Figure 4.14 Simulation results in CVD process (Thread 5 ~ 8)

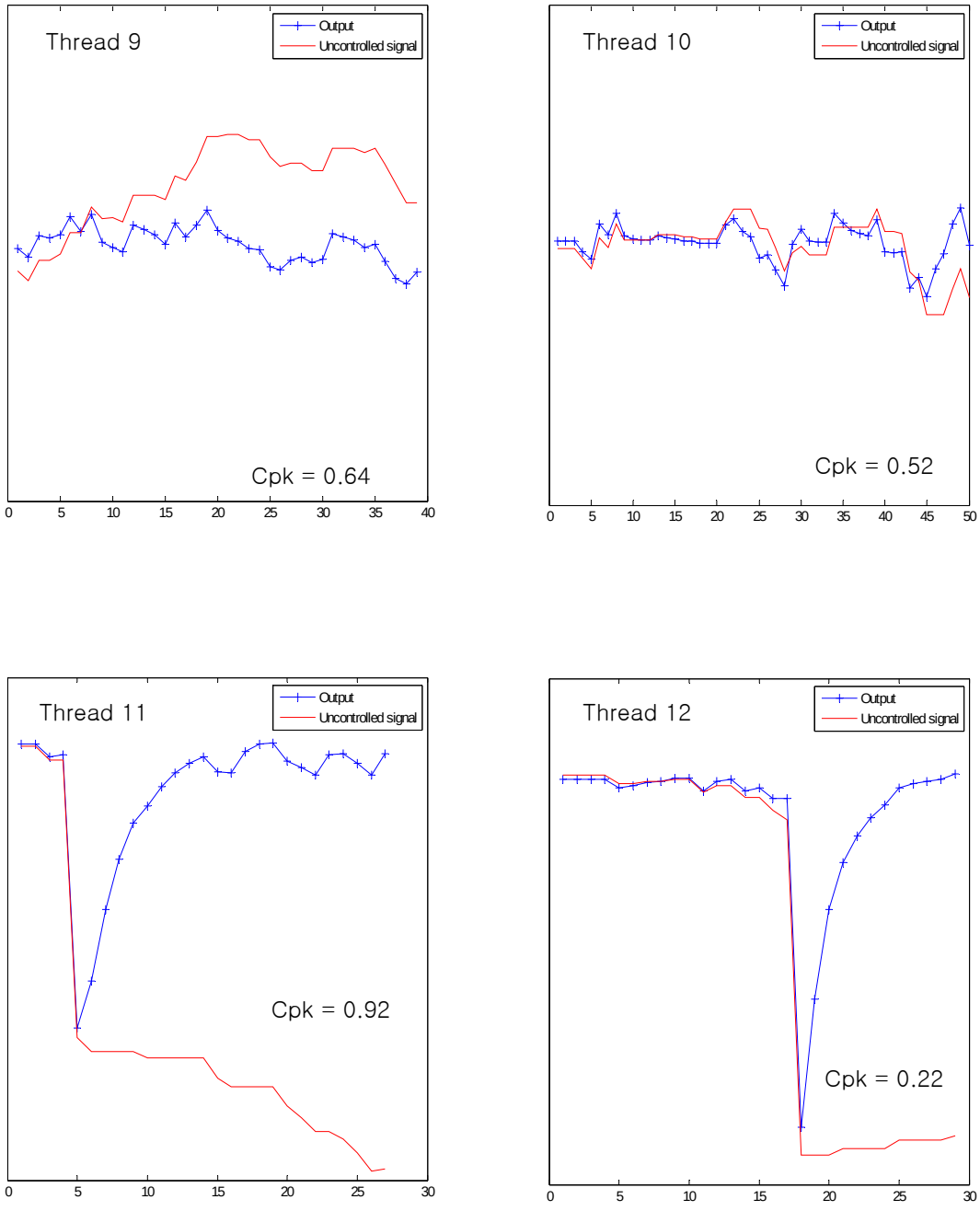


Figure 4.15 Simulation results in CVD process (Thread 9 ~ 12)

Note that the results shown in Figures 4.13 to 4.15 result from aggressive tuning of the modified RDC with Bayesian detection. Since the data set of CVD process has a small amount of noise, unlike that of the photolithography process in previous section, more aggressive tuning for drift compensation is applied. For comparison, the simulation results of EWMA and modified RDC are also shown in Table 4.6.

An obvious drift is shown in thread 2 and 8 so the modified RDC with aggressive tuning has a better performance than EWMA as expected. However, if there is no severe drift, aggressive tuning might overcompensate for process error. Thus in most cases, the modified RDC with aggressive tuning shows worse performance than regular tuning. To detect drift using statistics like the step disturbance detection with Bayesian approach will be very useful for drift compensation with an adjustment of tuning. Updating tuning parameters online according to process results is also a good option for better performance. Those two topics will be addressed in chapter 7 (recommendations for future work section).

Step disturbances occurred in most threads, so Bayesian detection can contribute to increase *Cpk* values. Note that Bayesian detection has no effect on thread 11 even though there is obviously a large step disturbance. That is because of the window size of pre-change windows described in section 4.2.1. In this simulation, the size is five points so Bayesian detection can be applied from the 6^{th} onward. However, a step disturbance occurs in the 5^{th} run in thread 11, so Bayesian detection cannot work for this case. If the size of pre-change windows is changed to four, the simulation result of control performance would be different. The process capability is changed from 0.9214 to 1.1603. That result is better than that of EWMA.

4.5 SUMMARY AND CONCLUSION

Semiconductor manufacturing is characterized by a dynamic, varying environment, where the equipment variables change or the target varies during the process step or between process steps. These changes act as disturbances and thus they should be compensated for using an appropriate process control strategy. There are two major disturbances in semiconductor manufacturing: gradual drift and abrupt shift (step disturbance).

The exponentially weighted moving average (EWMA) filter is widely used to compensate for disturbances in semiconductor manufacturing. It shows good performance for a step disturbance but does not perform as well for a gradual drift. Thus, an advanced feedback control method that treats drift as well as step disturbance is investigated in this chapter.

Two distinct but isolated problems, drifts and step disturbances, are considered for robust control at the same time by using the modified RDC with Bayesian detection. In the Bayesian detection system, observed states are classified into two categories: one for normal states with drift, the other for step disturbances. Then, drifts and step disturbances are controlled using different methods. The modified robust drift cancellation (RDC) method takes actions to compensate for states having process drifts or normal noises, while the Bayesian EWMA (B-EWMA) method is preferred to control step disturbances. However, the controller should be well-tuned because the performance of the controller is strongly dependent on tuning parameters, which should be properly defined prior to manufacturing application.

To investigate the performance of the modified RDC with Bayesian detection method, a simple drift, a simple step disturbance, and industrial examples of a photolithography process and a thin film deposition process from DMOS6 in Texas

Instruments Inc. was simulated. When an obvious drift is shown in process, the modified RDC with aggressive tuning has a better performance. When step disturbances occurred in process, Bayesian detection can contribute to increase *Cpk* values.

CHAPTER 5

Wafer-to-Wafer Feedback Control using Integrated Metrology

5.1 BACKGROUND AND MOTIVATION

The need for wafer-to-wafer (W2W) control of gate length has been accepted in the semiconductor industry, along with scatterometry for measurement, but different implementation approaches have been reported. W2W feedforward (FF) control can be implemented by using standalone scatterometry where each wafer is measured on a standalone tool and the measurements are then used to adjust the etch recipe for each wafer (W2W FF) [94]. Feedback for this method is generally lot-to-lot (L2L) with a variable number of lots delay. Using integrated scatterometry on the etch tool enables the measurement of each wafer (W2W FF) as with the IBM example. Implementation started with 90 nm node, and continued at the 65 nm node SOI-based microprocessors in IBM's 300 mm manufacturing facility [14][15]. In addition with integrated scatterometry, feedback can be updated within the lot (W2W FB) or after the lot has completed processing [15][95]. In this chapter, we evaluate W2W FB in process control by comparing its control performance with L2L FB. An optimum feedback filter for W2W FB is investigated by simulation. Also we evaluate chamber matching in a W2W process control system and research new methods to reduce sensitivity to disturbances for robust control.

5.2 REFERENCE DATA SET

Production data for modeling the incoming process disturbances and etch chamber drift was taken over a period of time that was statically significant, covering multiple maintenance events, multiple tools and chambers. Step disturbances at the first wafer of some lots were observed and included in the simulation model. Different patterns related to process disturbances and drift over a preventive maintenance cycle were observed and are important to the feedback modeling. All data were normalized to show the variation from target. Figure 5.1 shows the measurement structure used in feedforward and feedback. To be consistent with IBM papers, this paper refers to wafers measured before etch as Develop Inspect (DI) and wafers measured after etch as Stack Etch with Soft Mask remaining (SESM) [14][15][96].

With integrated scatterometry on etch systems SWA measurements are available in addition to CD. DI and SWA variability were observed for L2L and W2W in the data set. IBM and TEL found DI SWA affects SESM CD variation since previous CD-SEM based control model does not consider incoming SWA. (Figure 5.2) Therefore multiple inputs (DI CD + DI SWA) from integrated scatterometry were added to the control model to improve SESM CD control [96]. (Figure 5.3)

After observing SWA disturbance and corresponding CD shift, the IBM and TEL process team set out to develop a strategy to correct these CD shifts. Multivariate analysis of production data was performed to develop a new model. The new model takes a multiple input and single output (MISO) model format rather than current single input single output (SISO) format for better prediction of SESM CD. Figure 5.4 shows the prediction capability of the new MISO model vs. the current SISO model. Figure 5.5 also shows sensitivity analysis of SWA for both models.

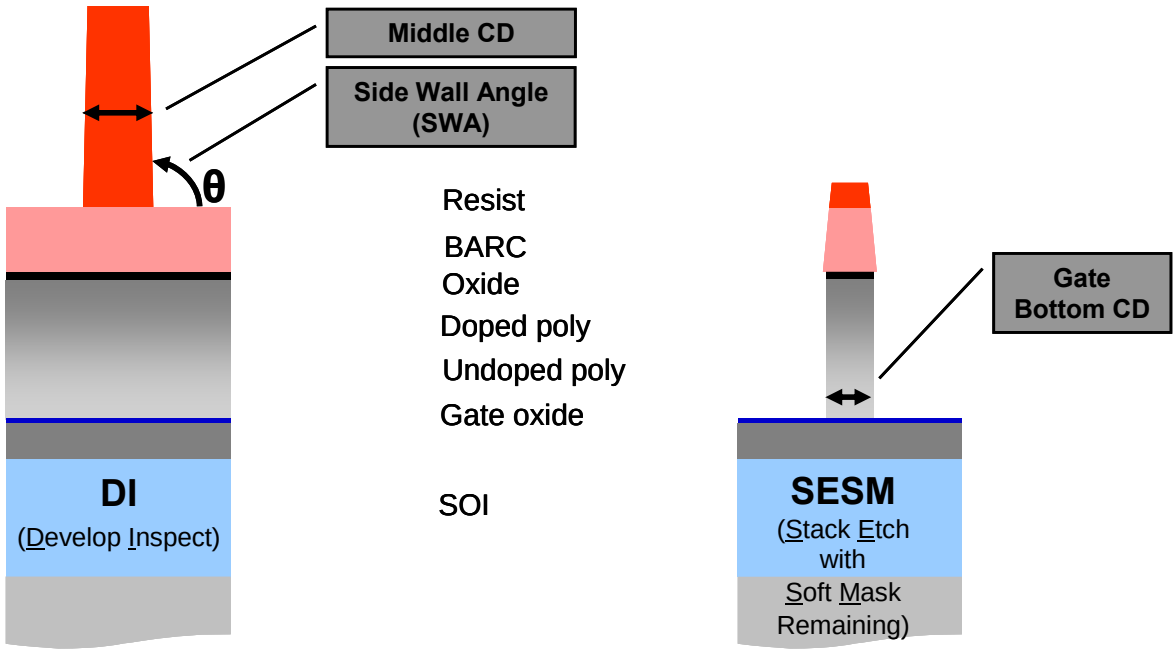


Figure 5.1 Measurement structure used in production

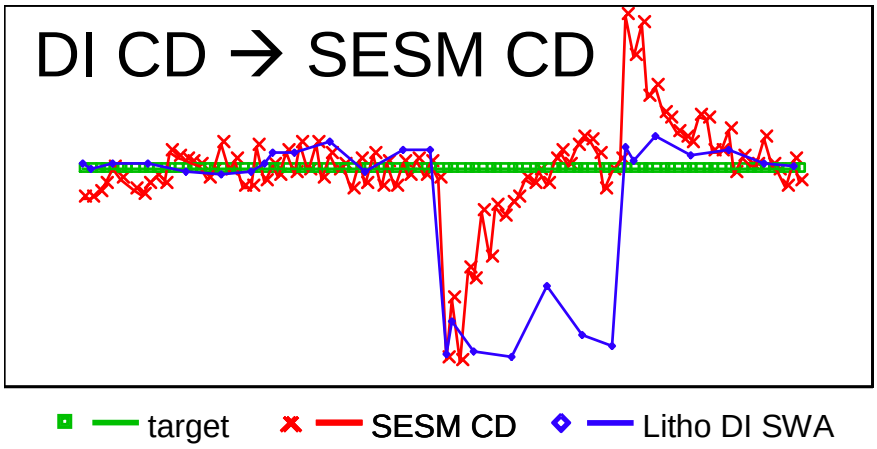


Figure 5.2 SESM CD variation with an impact of DI SWA. (DI CD intentionally not shown for simplicity)

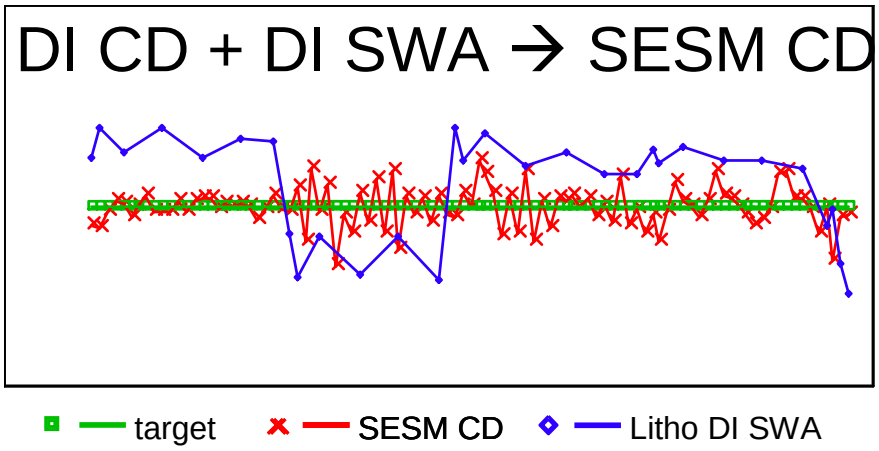


Figure 5.3 SESM CD variation of new control model with multiple inputs. (DI CD intentionally not shown for simplicity)

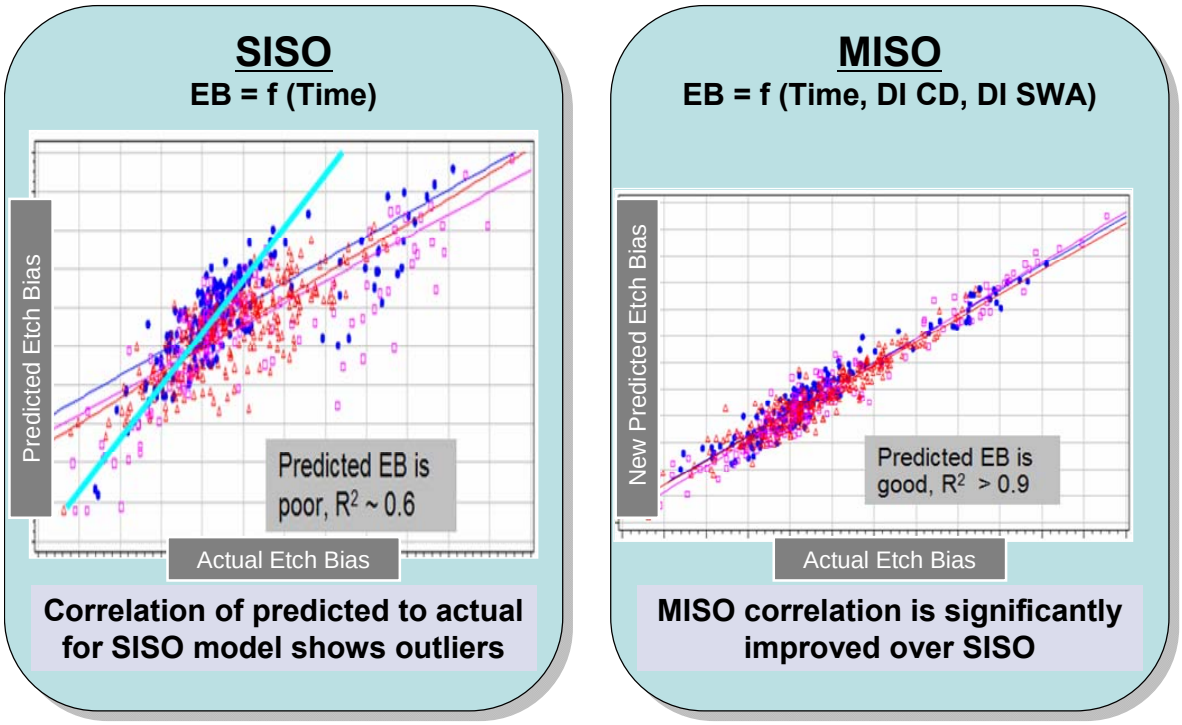


Figure 5.4 Etch Bias Model Comparison – SISO vs. MISO

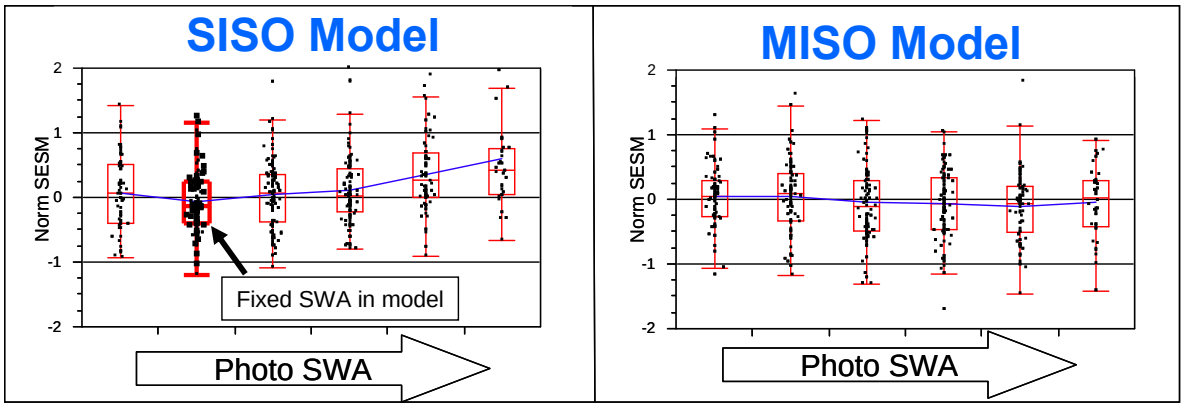


Figure 5.5 Sensitivity analysis of SWA – SISO vs. MISO

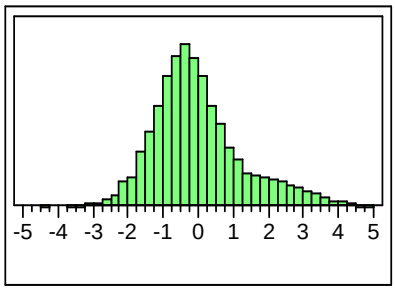


Figure 5.6 Normalized DI CD distribution shows a 3σ variation of 3.87nm

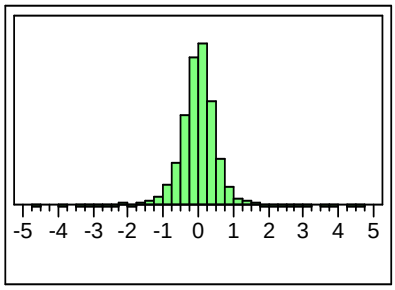


Figure 5.7 Normalized SESM CD distribution shows a 3σ variation of 1.74 nm

The data set used for modeling the incoming data with covering multiple lithography maintenance events and multiple scanners and tracks is shown in Figure 5.6. When used for feedforward control, the model assumes 100% of the wafers were used for adjustment of incoming W2W variation.

The data set used for modeling the etch process using W2W FB control with a two wafer delay is shown in Figure 5.7.

5.3 SIMULATION MODEL

The simulation model is built using MATLAB software. This model is designed to execute the same sequence as the poly gate etch process used in production. The data set for simulation is 65 nm node Silicon on Insulator (SOI) based microprocessors in IBM's 300 mm manufacturing with multiple maintenance events, multiple etchers, chambers and integrated measurement units. First, it collects incoming data from the data set. It then feeds forward to the controller unit to calculate the recipe settings. The manufacturing model formula and control law in production are used with the target value. Next the etching process is run using this recipe in the simulated process unit. Disturbance and noise are added in this step. A disturbance and noise are modeled based on feedback update by considering output error results from the production data set based on model predictions. There are common sources of variation for the etch process in semiconductor manufacturing as shown in Table 5.1. The disturbance is mostly related to plasma chamber conditions [1][97]. After the end of the process the output value is measured, an error (target minus output) is calculated, and these values are saved in the database with lot ID and wafer ID. Finally, these saved values move to the observer unit to update an offset using an exponentially weighted moving average (EWMA) filter [1][5][98]. Generally, the process model has a linear form:

$$y_k = Bu_{k/k-1} + c_{k/k-1} + e_k \tag{5.1}$$

where y_k is the output at batch k, B is the process gain, $u_{k/k-1}$ is the input at batch k calculated from information up through batch $k-1$, $c_{k/k-1}$ is the estimate for the offset, and e_k is unknown process noise entering the system. The offset is updated recursively by an observer of the form:

$$c_{k/k-1} = \lambda(y_{k-1} - Bu_{k-1/k-2}) + (1-\lambda)c_{k-1/k-2} \tag{5.2}$$

where λ is the exponential weighting factor, or tuning parameter, of the observer. The λ takes a value between 0 and 1 and is chosen based on the desired properties of the observer.

Cycle	Type	Cause	System
Maintenance Cycle-to-Cycle	Disturbance	Repairs, chamber cleans, PM, kit replacements, gas flow change, lamp change (IM)	Process or Metrology
Within a Maintenance Cycle	Drift	Gradual build-up on chamber, machine wear, sensor drift	Process or Metrology
Lot-to-Lot	Disturbance	Different incoming wafer state, due to current process (other lots processes run between lots of this recipe), control model mismatch, measurement model mismatch	Process
Within a Lot	Drift	1st wafer effect, warm-up, degassing, different steady state chamber condition	Process
Wafer-to-Wafer	Disturbance	Incoming wafer state, different pre-process chambers, litho module	Process
Within a Wafer, Within a Region	Drift	Changes in materials, local heating effects	Process
Within a Wafer, Region-to-Region	Disturbance	Different processing or material exposed	Process

Table 5.1 Common sources of variation

This offset value is delivered to the controller unit to be used for compensating the disturbance. In the case of L2L feedback, the mean within lot error value is used for updating an offset and this offset is used by the next lots until a new update comes out. A time delay is inserted into this model to simulate the offset update. This procedure continues until the final wafer is processed, and the various delays were modeled.

Three different measurement strategies are modeled as instances in this paper. The control system in Figure 5.8 shows how feedforward and feedback can be performed using standalone metrology. A multiple lot time delay, out of order measurements, and a mix of standalone metrology can be simulated in this environment [95][99]. For modeling simplicity only five lot delay in feedback was simulated. The results of using L2L feedback with standalone metrology is shown in Figure 5.9. If we consider out of order metrology, the result would be much worse than this.

The control system shown in Figure 5.10 represents an example where integrated scatterometry is used for feedforward (W2W FF) and feedback control, with L2L FB. This system has only one lot time delay and no out of order meteorology problems since post process results can be immediately measured using IM [14] Simulation result is shown in Figure 5.11.

Figure 5.12 represents a case where integrated scatterometry is used for feedforward (W2W FF) and feedback (W2W FB). Here the controller resides at the tool and the host communicates the target used for control and other control variables at run-time.

Figure 5.13 shows the results of W2W feedback with a two wafer delay. By comparing with previous L2L feedback results, it turns out that wafer-level feedback enables compensation of most unknown disturbances within lot.

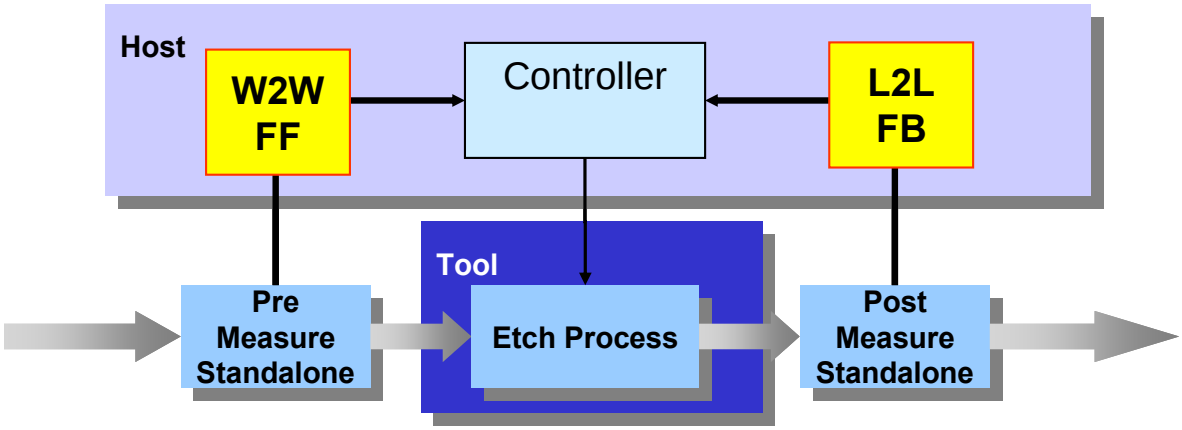


Figure 5.8 Simulated control system using standalone metrology for W2W FF and L2L FB

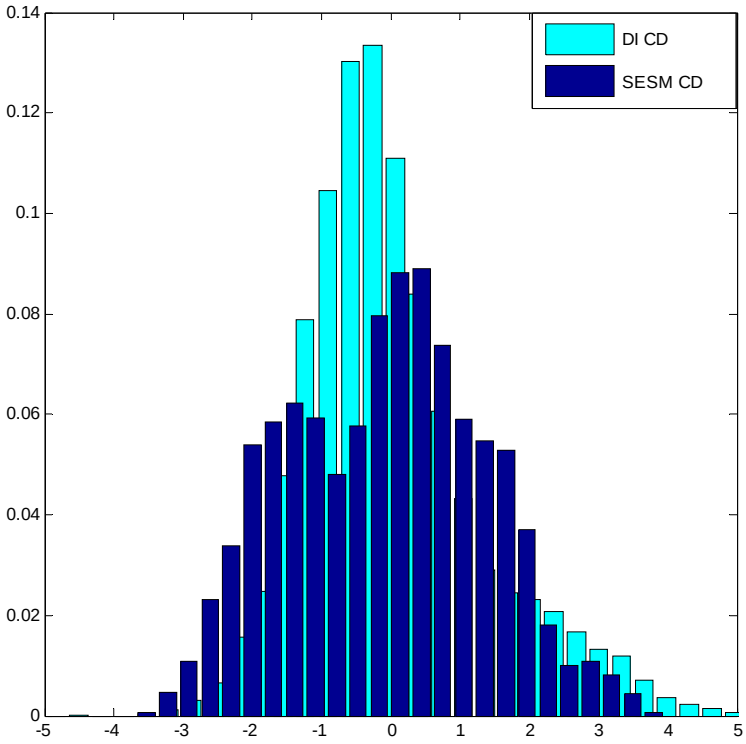


Figure 5.9 Simulated L2L FB of standalone with five lots delay ($3\sigma = 4.26$nm)

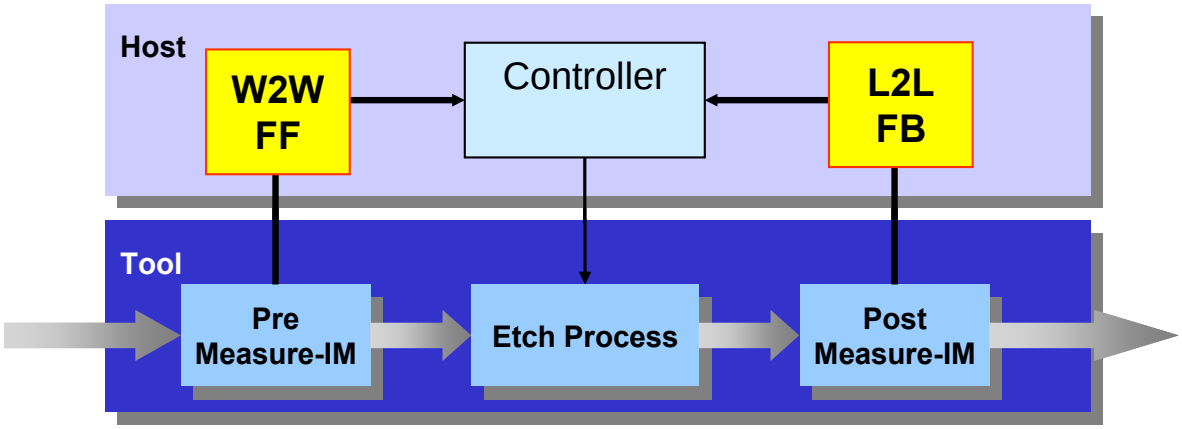


Figure 5.10 Simulated control system using integrated scatterometry for W2W FF and L2L FB

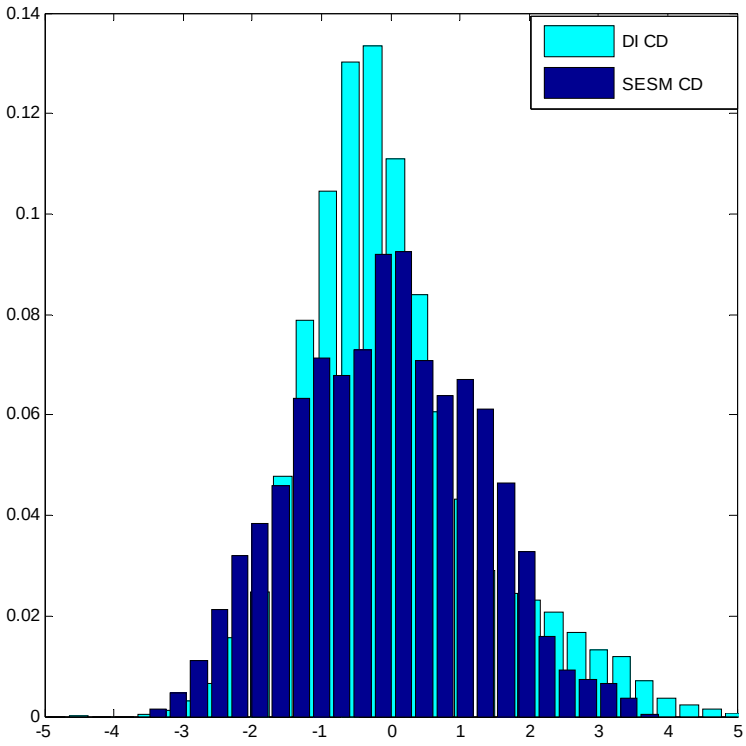


Figure 5.11 Simulated L2L FB of IM with one lot delay (3σ = 3.94 nm)

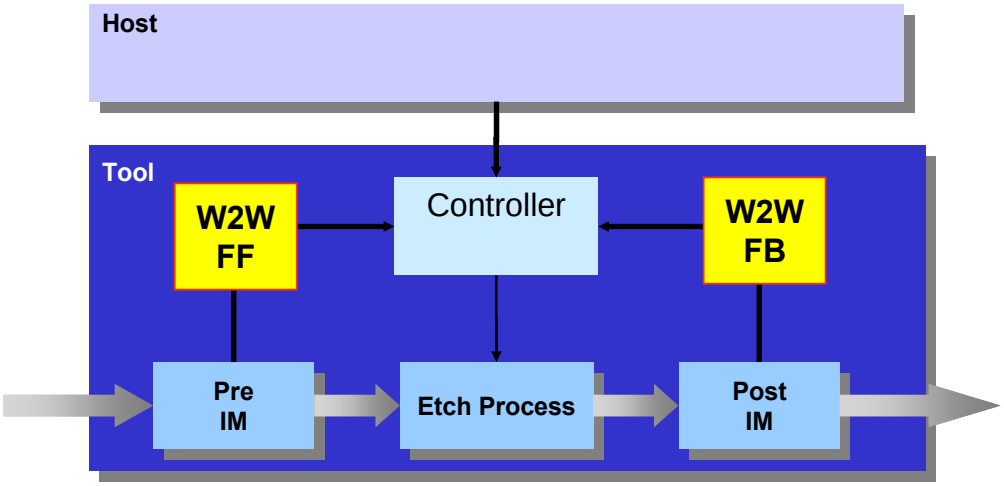


Figure 5.12 Control system used in production and simulation with integrated scatterometry residing on the tool for W2W FF and W2W FB

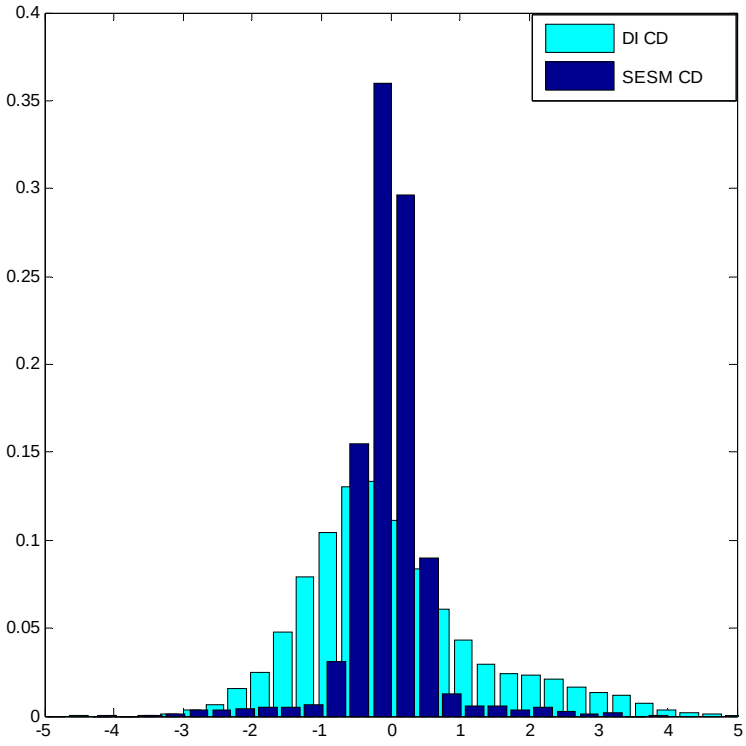


Figure 5.13 Simulated W2W FB using a two wafer delay shows a SESM CD 3σ variation of 1.74 nm

Therefore these simulation results can verify the advantage of wafer-to-wafer (W2W) feedback control as well as the utilization of integrated metrology (IM).

By comparing with L2L FB with the standalone case, W2W FB using integrated scatterometry shows a 59.2% improvement for control performance. (Figure 5.14)

5.4 FEEDBACK CONTROL PROPERTIES

In W2W FB, the feedback filter plays an important role to estimate a disturbance accurately. Although there are many state estimation filters, three methods – EWMA, Kalman filter [73], and strong tracking filter (STF) [100][101] are used as simulation examples.

To analyze the performance data consider the following system and measurement model:

$$x_{k+1} = Ax_k + Bu_k + Gw_k \tag{5.3}$$

$$y_k = Cx_k + v_k \tag{5.4}$$

where x_k is process state, u_k is the process input, y_k is measurement output, and w_k and v_k are random process noise and measurement noise with zero mean and covariance matrix Q and R, respectively.

Any time the process is run, the state estimates and error covariance are updated in Kalman filter using the following equations:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k \tag{5.5}$$

$$P_{k+1} = AP_kA^T + GQG^T \tag{5.6}$$

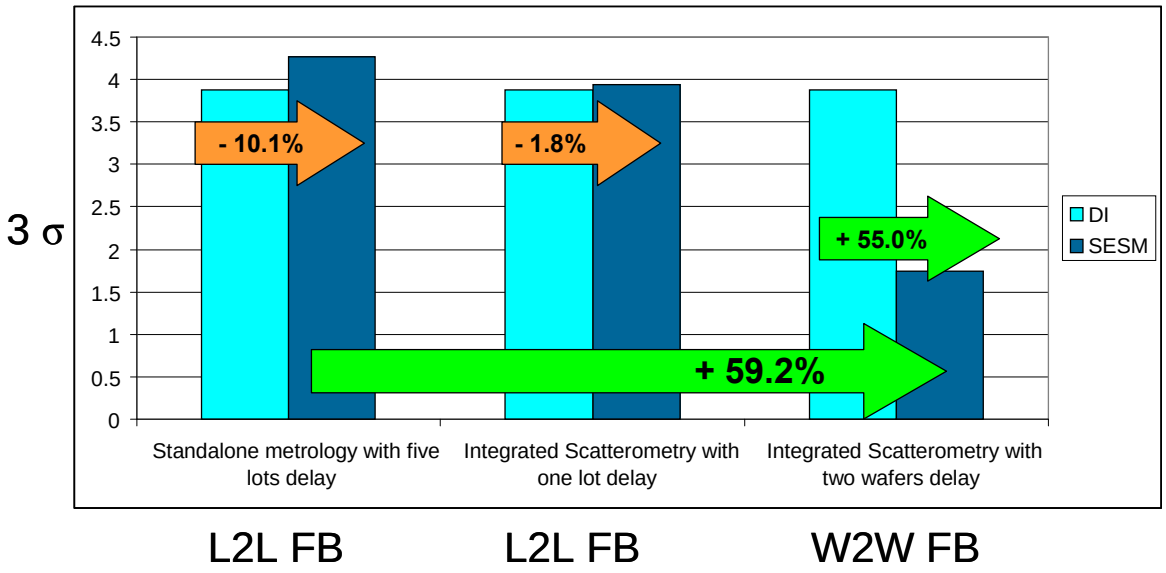


Figure 5.14 Comparison of the results

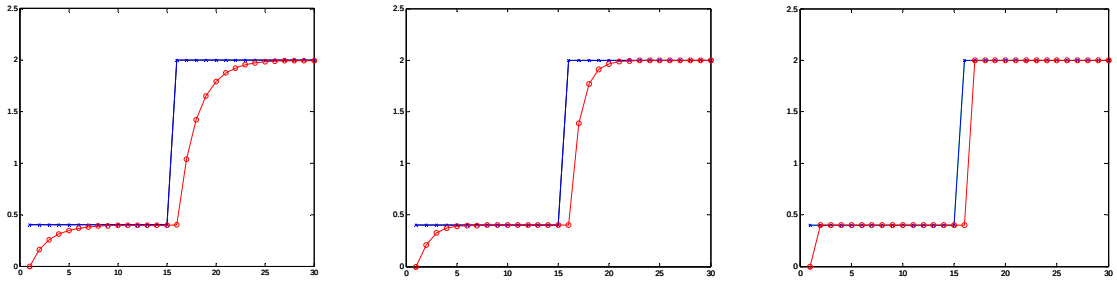


Figure 5.15 The estimation of the step disturbance without noise (a) EWMA (b) Kalman filter (c) STF

When a measurement is taken, the Kalman gain, the state estimates and error covariance are updated:

$$K_{k+1} = P_k C^T (C P_k C^T + R)^{-1} \tag{5.7}$$

$$P_{k+1} = (I - K_{k+1} C) P_k \tag{5.8}$$

$$\hat{x}_{k+1} = \hat{x}_k + K_{k+1}(y_{k+1} - C\hat{x}_k) \tag{5.9}$$

STF modifies error covariance update (5) by introducing suboptimal fading factors.

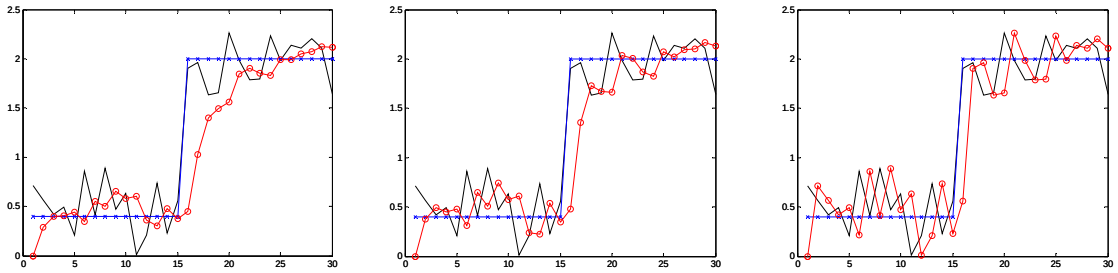
$$P_{k+1} = LMD(k+1) A P_k A^T + G Q G^T \tag{5.10}$$

$$LMD(k+1) = diag[\lambda_1(k+1), \cdots, \lambda_n(k+1)] \tag{5.11}$$

where $\lambda_n(k+1)$ are suboptimal fading factors each greater than or equal to unity. Barad, *et. al.* [100] and Zhou, *et. al.* [101] explained how to determine these suboptimal fading factors. A measurements update in STF is the same as the Kalman filter (5.7)-(5.9).

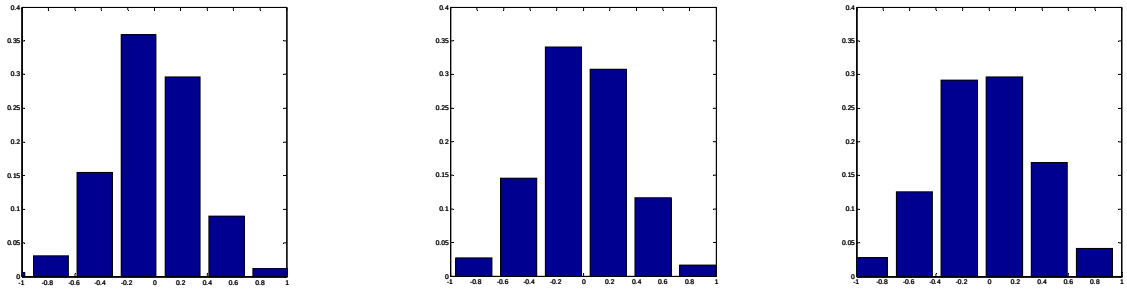
Figures 5.15 and 5.16 show the estimation of the step disturbance with and without noise, respectively, and of three different filter methods: EWMA, Kalman and Strong Tracking Filter (STF).

The strong tracking filter responds to a step disturbance much faster than EWMA and Kalman filters, but it is more sensitive to noise. Therefore if the magnitude of process or measurement noise is small enough to be ignored, then the STF would be the best filter.



(a) EWMA (b) Kalman Filter (c) STF

Figure 5.16 The estimation of the step disturbance with noise (a) EWMA (b) Kalman filter (c) STF



(a) EWMA (b) Kalman Filter (c) STF

Figure 5.17 Error distribution : (a) EWMA filter λ=0.4 (3σ = 1.74), (b) Kalman filter (3σ = 1.72) , (c) STF (3σ = 1.81)

However, most etch processes have a few nanometers of noise, which can affect control performance [1]. Thus robustness to noise is also a requirement of the estimator in real manufacturing. This means that STF is not the best filter in this simulated feedback control model. Figure 5.17 shows the error distribution of W2W feedback control by implementing EWMA, Kalman filter, and STF.

According to the above results, If the magnitude of process and measurement noise is small enough to be ignored, the STF is the best filter. But most etch processes have a few nanometers of noise, which can affect control performance. Kalman filter shows the best performance among these three filters. However, it has three tuning parameters while EWMA has only one. In a production environment with changing part numbers and products, it will require more setup and re-tuning. Therefore EWMA is the best state estimation filter in terms of robustness to noise and ease of tuning.

λ was varied to optimize MSE and the results are shown in Table 5.2.

λ	3σ	MSE
0.2	1.8919	0.3979
0.3	1.7890	0.3558
0.4	1.7429	0.3377
0.5	1.7245	0.3306
0.6	**1.7214**	**0.3295**
0.7	1.7292	0.3324
0.8	1.7462	0.3390
0.9	1.7721	0.3491

Table 5.2 The effect of λ on MSE in EWMA filter

The best value of the tuning parameter (λ) is 0.6. However, the 3 sigma value did not change much for λ between 0.4 to 0.8. This means that EWMA is robust and stable in this simulated model regardless of tuning parameter [98]. In other words, whatever value of λ is chosen from the range between 0.4 and 0.8, the system will provide good performance.

Another variable in question is the sensitivity to feedback delay when wafer-to-wafer FB is available since feedback delay between manufacturing and metrology occurs often in semiconductor manufacturing, leading possibly to closed-loop instabilities [102]. Table 5.3 shows the simulated results ranging from no delay to a six wafer delay.

delay	3σ	MSE
no delay	1.3494	0.2025
1	1.5591	0.2703
2	1.7429	0.3377
3	1.9013	0.4019
4	2.0458	0.4652
5	2.1816	0.5290
6	2.3042	0.5901

Table 5.3 Sensitivity of wafer delay using W2W feedback and EWMA filtering with λ=0.4

As the delay increases, the control performance becomes worse, thus feedback delays need to be accounted for in the process. Integrated metrology is the only viable solution to reduce feedback delay in wafer-level control. According to the results in

Table 5.3, with a few send-ahead wafers the benefit of integrated metrology can be maximized.

5.5 SIMULATION RESULTS

Using 12,000 wafer sample data set measured pre and post profile in 65nm poly gate etch process of IBM production, more simulations were carried out. This high volume data set has 3.2nm 3σ incoming variation with multiple scanners, tracks, etchers, chambers, integrated measurement units, and maintenance events such as chamber cleans and IM lamp changes.

Figure 5.18 shows the CD distribution by chamber. There are multiple tools and chambers in this production but since the number of tool and chamber used is proprietary, they are not reported in this chapter.

5.5.1 Control Performance of W2W Feedback

Two cases of simulation are performed. They are standalone metrology system with L2L FB with five lot delay and integrated metrology with W2W FB with two wafer delay. Figure 5.19 and Table 5.4 shows the results of simulation.

Data type	3σ	
	Reference data set	Simulated data set
DI	3.207 nm	
Standalone L2L	N/A	2.234 nm
IM W2W	1.500nm	1.500 nm

Table 5.4 Control performance of standalone L2L vs. IM W2W

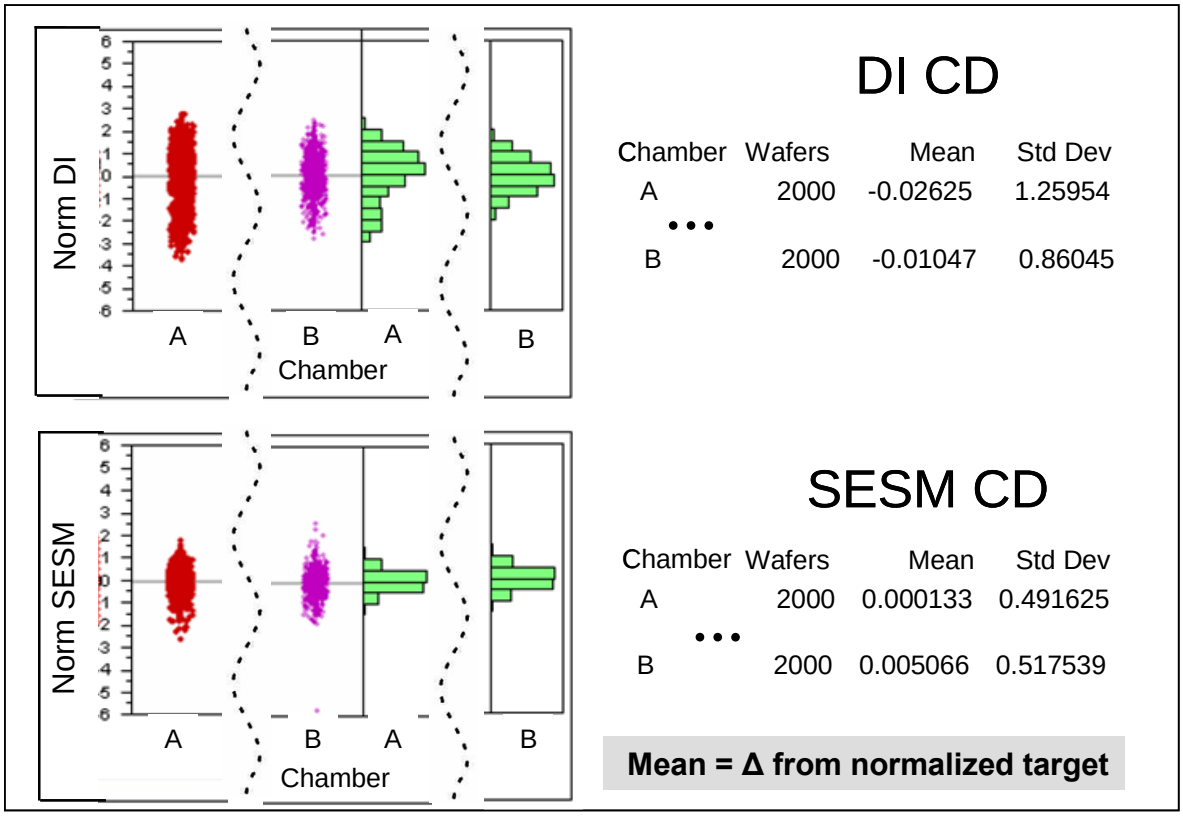


Figure 5.18 CD distribution by chamber

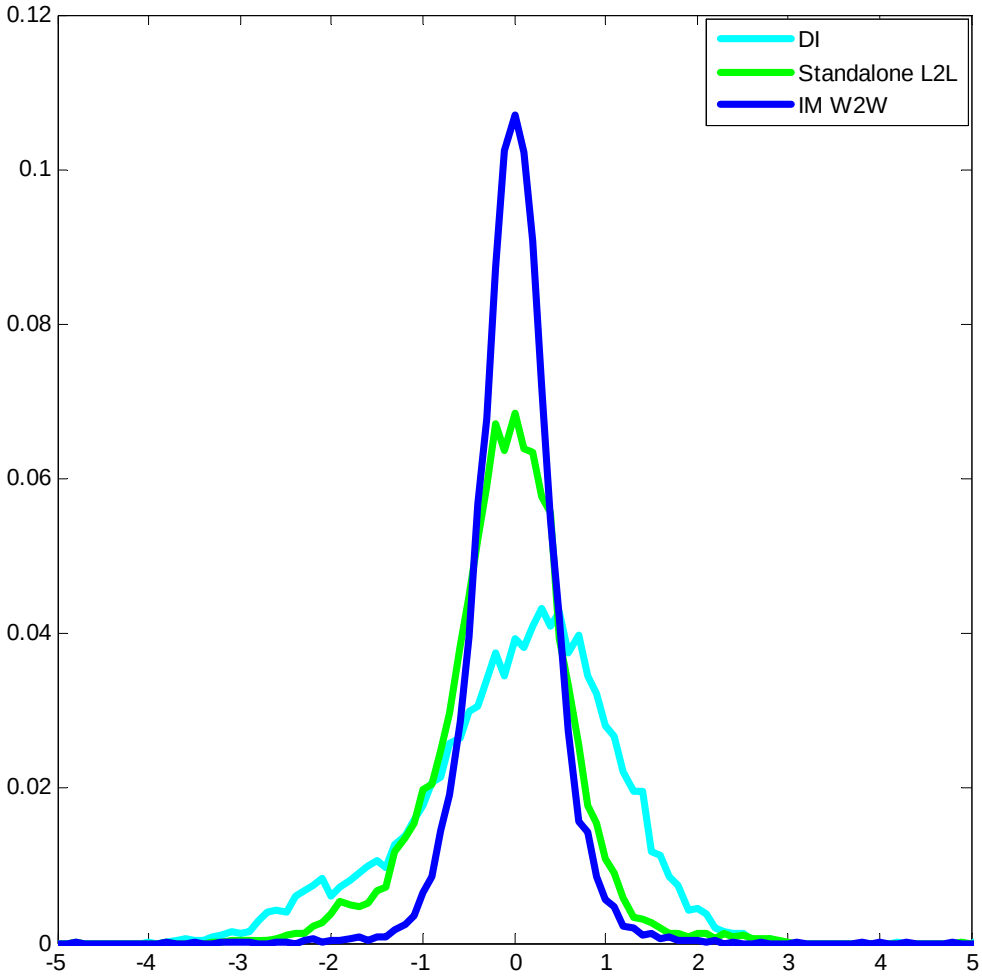


Figure 5.19 Simulated control performance of standalone L2L vs. IM W2W

As mentioned in Section 5.3, IM W2W FB has much better performance than standalone metrology system with L2L APC. Both systems show tighter distribution than incoming (DI) distribution, which means both controllers compensate for incoming variation. However, because IM W2W eliminates incoming variation as well as process variation such as drift and disturbance effectively, it can achieve a much tighter distribution. Also according to Table 5.4, the simulation results are well matched with reference CD distribution in production.

5.5.2 Chamber Matching

For the chamber matching effect of both systems, the simulated CD distribution is arranged by chamber. Figure 5.20 and Table 5.5 shows chamber matching results for both standalone L2L and IM W2W. In Table 5.5 max Δ means maximum difference between chamber averages and σ means the standard deviation of chamber averages. Therefore IM W2W shows 10 times improvement for chamber matching in comparison with standalone L2L.

Data	max Δ	σ
Standalone L2L	0.2363	0.09842
IM W2W	0.0203	0.00919

Table 5.5 Chamber matching – standalone L2L vs. IM W2W

As shown in this simulation, R2R control with IM W2W positively affects chamber matching. Even though it is not the ultimate solution for chamber matching, it provides a good alternative.

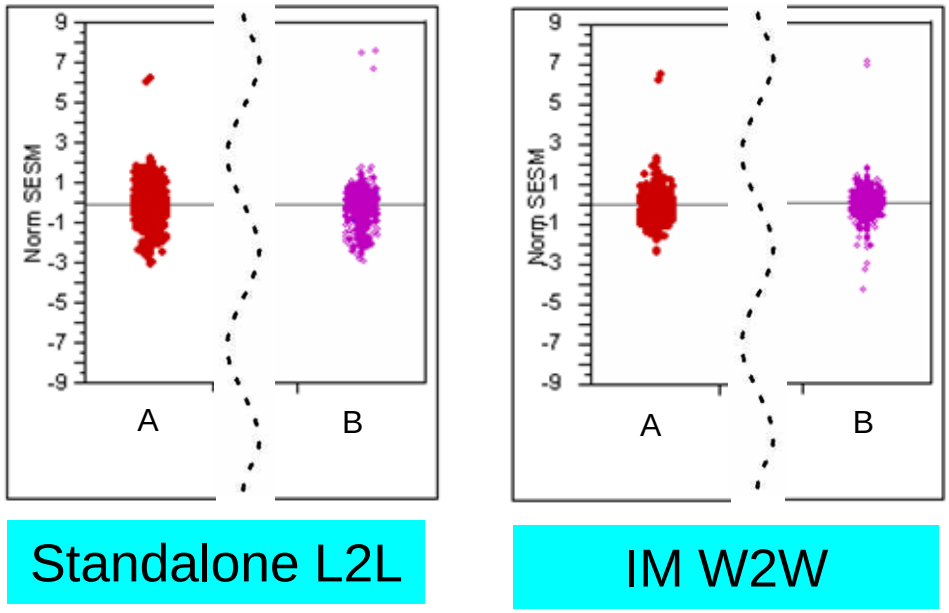


Figure 5.20 Chamber matching – standalone L2L vs. IM W2W

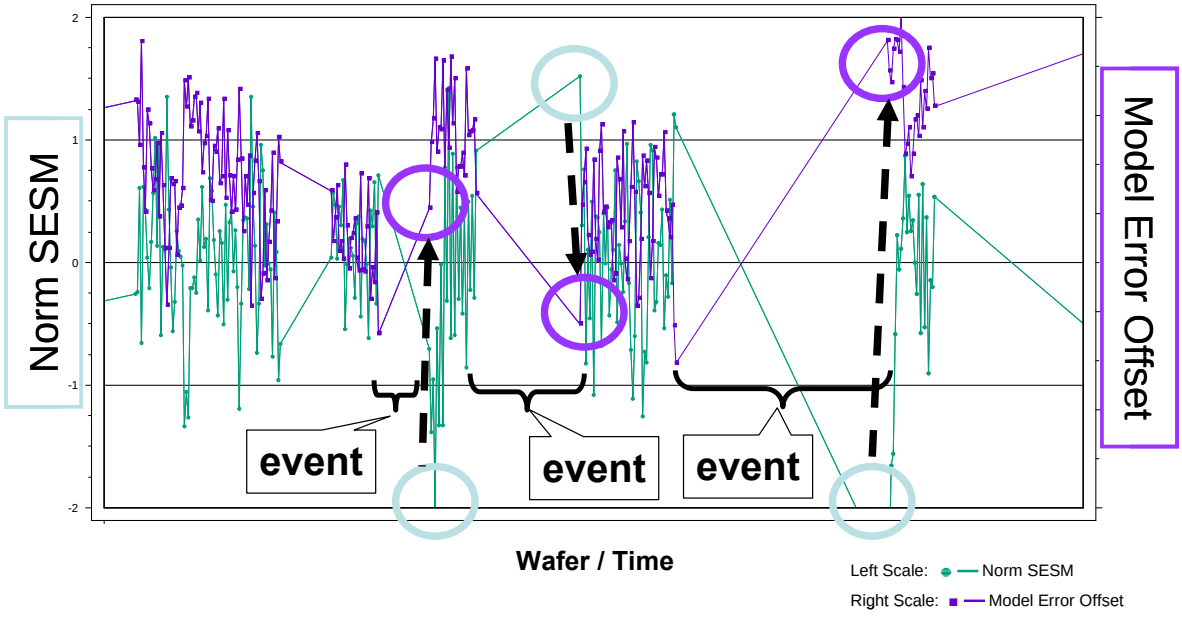


Figure 5.21 Observation from data set

5.5.3 Advanced Feedback Control Approach

Figure 5.21 is a snapshot of one chamber's data by wafer / time. We define an "event" as the gap between production due to issues such as maintenance events and product changes. In Figure 5.21 a large disturbance is observed at the first wafer after such events. In order to reduce errors that can be detected based on known events, two advanced methods are proposed in this section – rule-based model update tuning and send-ahead wafer running. The reference data set is same as previous simulation and there are 21 events observed in the data set.

Instead of a fixed tuning parameter, rule-based model update tuning is applied by varying λ of the EWMA filter. A higher λ value is used for tuning for the first few wafers and a lower λ value is used for tuning the last wafers. As shown in Figure 5.22, it results in slightly better performance. The 3 sigma value of rule-based model update tuning is 1.472nm and that of fixed tuning is 1.500 nm.

In send-ahead wafer running method, two send-ahead wafers are run after events to reset the model offsets. So they use a λ of 1.0 and zero delay for feedback in simulations. Since there are 21 events in the reference data set, a total of 42 send-ahead wafers are used. The result has 1.440 nm of 3 sigma variation.

If we combine send-ahead wafers and rule-based model update tuning, much better results are obtained. It shows 1.404 nm of 3 sigma variation (Figure 5.24). This modification of APC could have a quite large benefit for control performance in high volume manufacturing.

Figure 5.25 shows the improvement of control performance for each APC approach. As shown in the chart, APC should contribute to reducing 3 sigma variations in order to satisfy tight control spec in production.

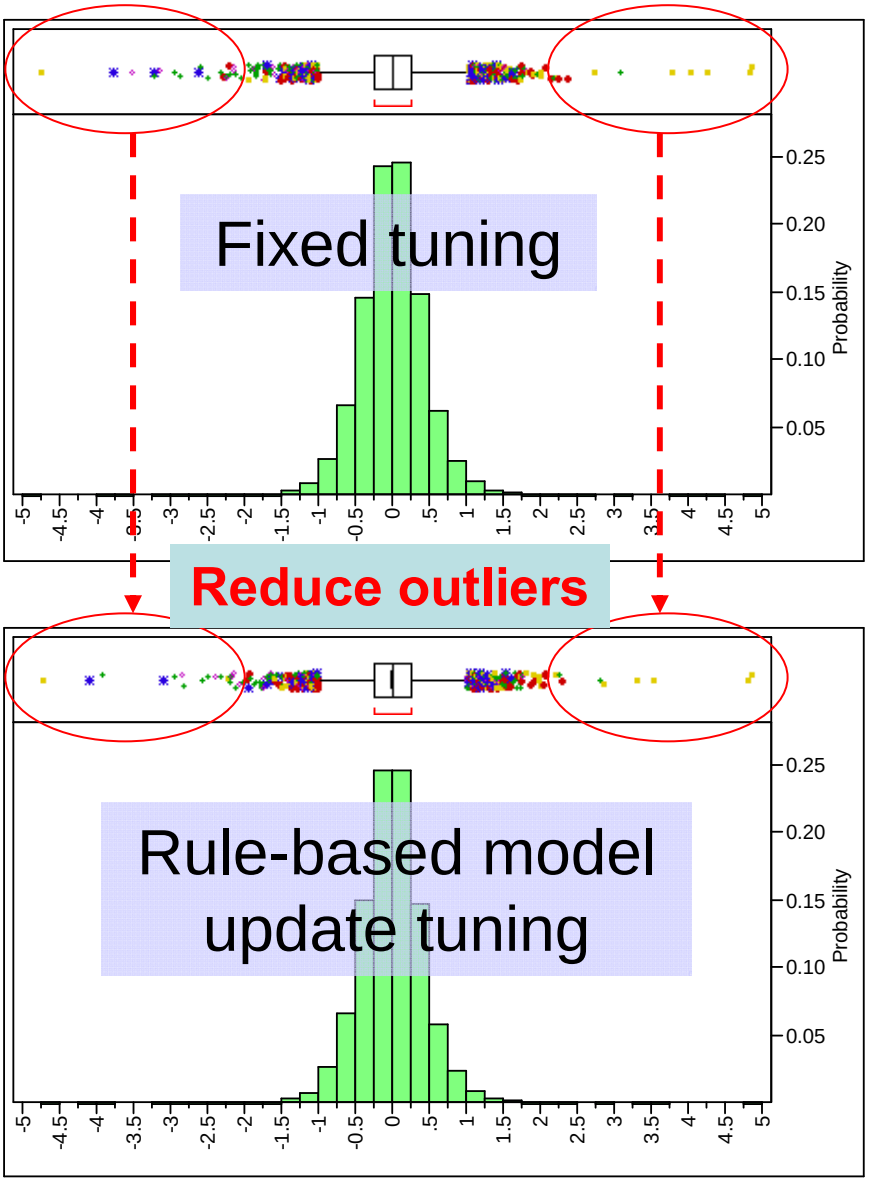


Figure 5.22 Rule-based model update tuning

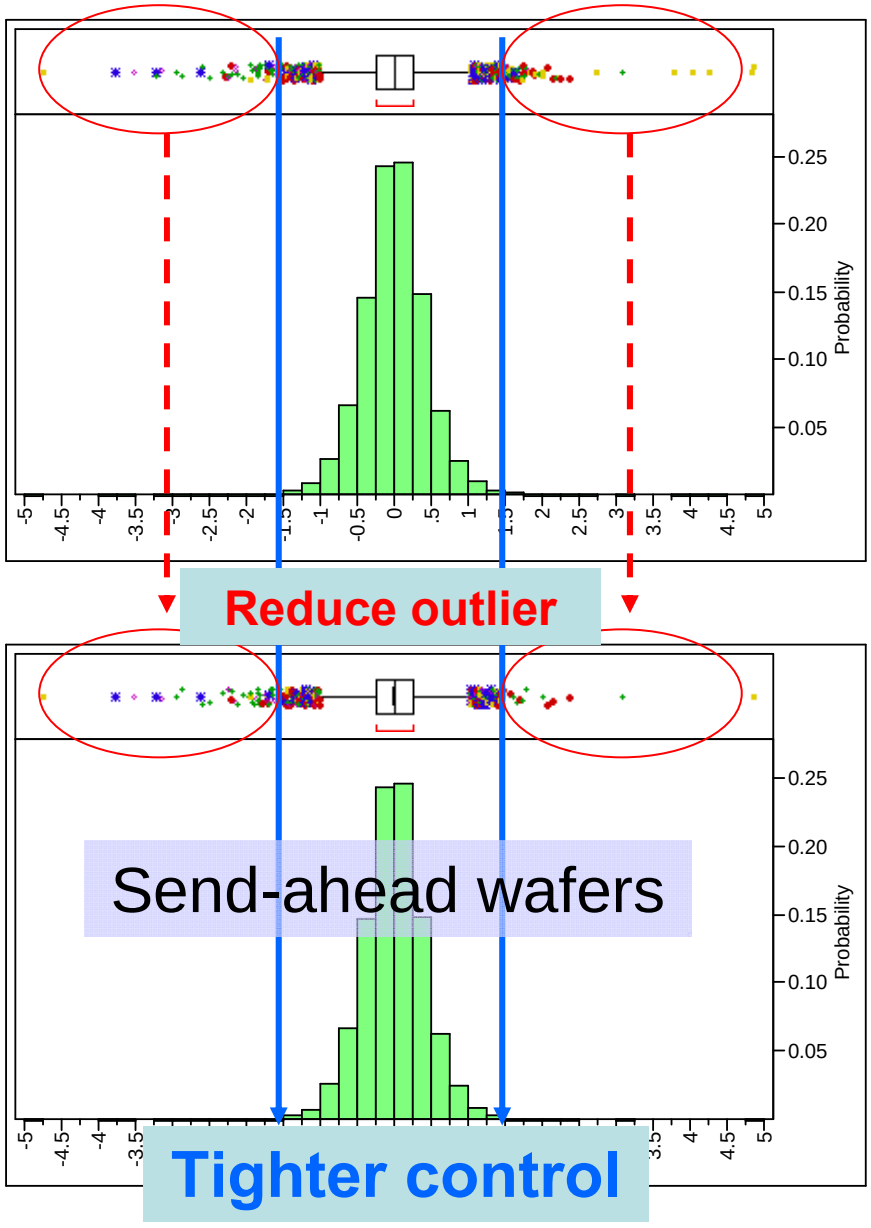


Figure 5.23 Send-ahead wafer running

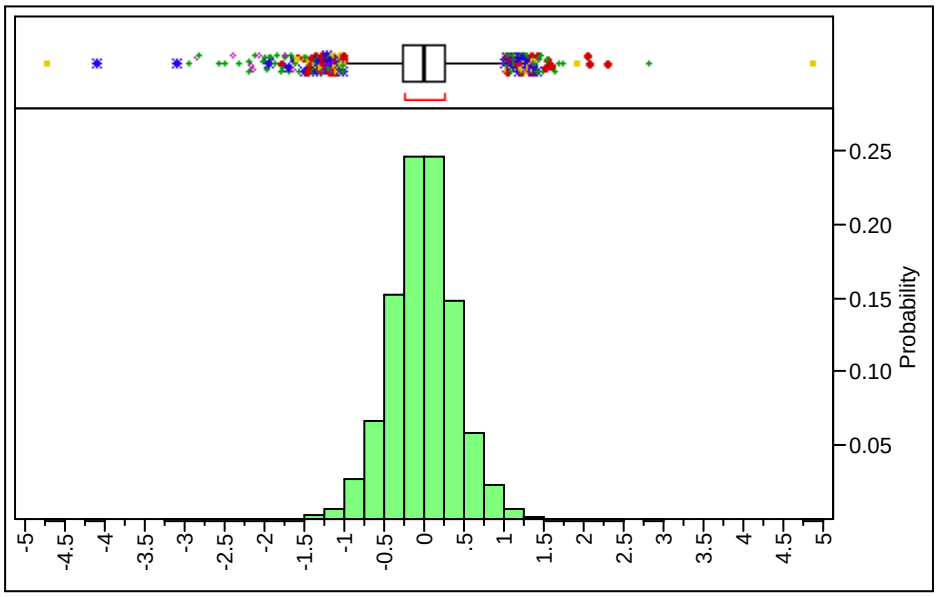


Figure 5.24 Event based send-ahead wafers with rule-based model update tuning

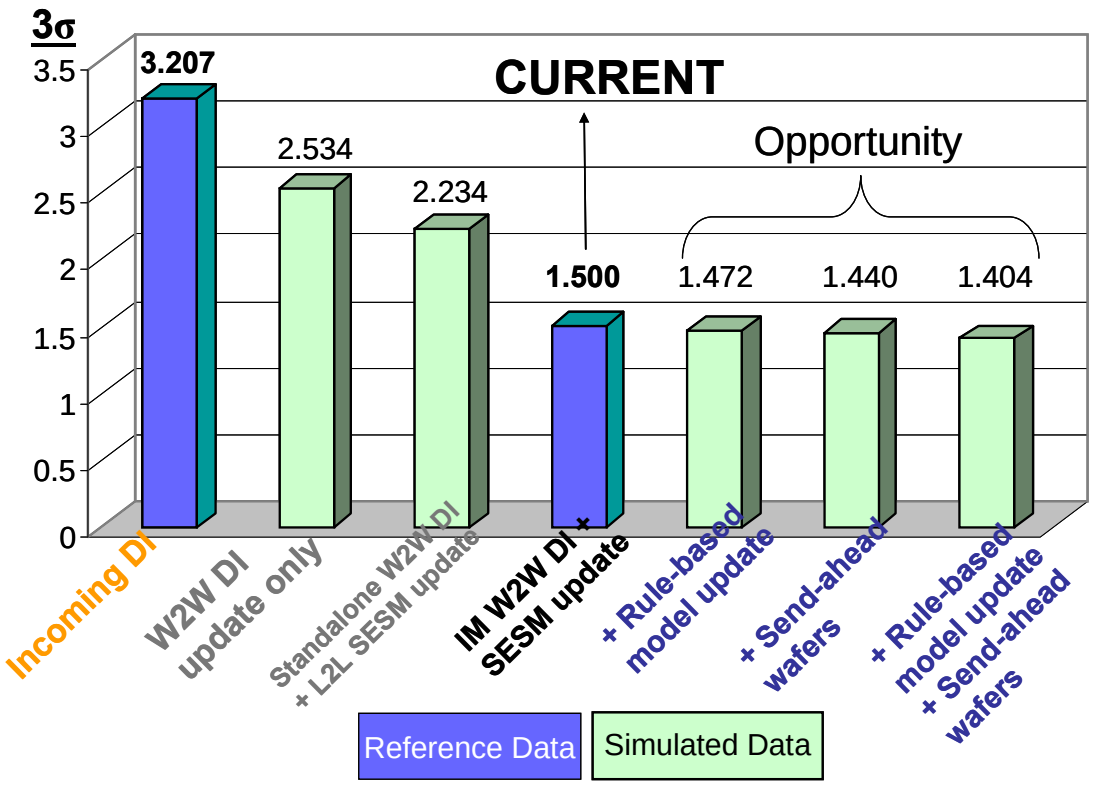


Figure 5.25 W2W CD simulated projection

5.6 SUMMARY AND CONCLUSION

It is generally accepted that W2W feedforward can compensate for incoming variation from a previous tool, such as a litho cluster. With wafer-to-wafer control the previous system can manage more variation, which is compensated for by making a wafer-to-wafer recipe adjustment.

The results shown in this chapter demonstrate the value of using integrated scatterometry with W2W FB control to correct for step disturbances (L2L) and for variation of an etch chamber over time within a lot. It also helps to refine and tune the control and measurement models for compensating for process model error quickly.

With the simulation model the sensitivity to different feedback filters was explored. EWMA was found to have excellent robustness and utility. The optimum EWMA settings for W2W FB were recommended and the impact of wafer delay in a W2W FB control system was also investigated.

W2W FB with integrated scatterometry shows better control performance as well as chamber matching than standalone L2L FB by simulation using high volume manufacturing data set. Advanced model update methods reduce 3σ of SESM to 1.404 nm and the limitation of 3σ in this data set is 1.332 nm (zero delay) for fixed tuning model update.

For better profile control, the focus needs to be on new process control models with more variables. For better physical CD control, we need to continue refining model parameters of control model and add additional films and profile measurements. Also for electrical CD, we can add poly SWA control, which will be discussed in the next chapter.

One key achievement is chamber matching effect in poly gate etch using IM W2W APC. This opens the door for other APC applications that require robust manufacturing such as spacer, STI, contact, trench, etc.

CHAPTER 6

Multiple Input Multiple Output Optimization and Control

6.1 BACKGROUND AND MOTIVATION

Advanced process control (APC) in semiconductor device manufacturing is widely used when the process window requires tight control and implementation costs and support can be justified. Examining process step variability and interaction with the circuit design performance is the first step in determining the value of adding process control. Robust process control systems require accurate measurements and process tool models that can work across process equipment modules over time. A strong correlation of physical measurements to electrical device characteristics means that the physical control will drive an improvement in device performance. Most semiconductor process control applications operate at a lot level where wafers are sampled between process steps and measured on standalone metrology. As the geometries shrink and the performance and chip densities continue to increase, critical steps need finer levels of control and monitoring, such as wafer-to-wafer and with-in-wafer. In addition to more measurements, profile information is necessary: therefore, scatterometry is the preferred measurement for etch process control. Scatterometry provides faster measurements, while improving accuracy and repeatability of measurements.

Additionally, scatterometry allows integrated metrology (IM) since the unit can be small enough to physically locate on a process tool. Integrated metrology measurements can be made in parallel with processing and have the ability to make litho and etch measurements without the complexities and delays of standalone metrology. Most processes can be controlled using a single measurement, such as CD or thickness

and adjusting a single process recipe parameter, such as time. Single input single output control (SISO) is the basic process control and can reduce the wafer-to-wafer (W2W) variation significantly over lot-to-lot (L2L) [14][15]. Figure 6.1 shows the complexities of the gate pattern sequence and the many critical measurements needed for monitoring and control at both the litho and etch steps. For example, gate CD control has evolved to include multiple input parameters, such as CD and sidewall angle (SWA) of the resist, as shown CR_{PR} and θ_R respectively [96].

Based on the 2007 ITRS projections resist gate CD and pitch will continue to shrink. If we look at the line spacing CD_P calculated as a function of pitch and resist gate CD_{PR} in Figure 6.2, the area will reduce by a factor of 2 in the next four years. As geometries shrink, the process windows get smaller and variations such as SWA, film thicknesses, and CD that were once independent of each other are now showing an interaction that requires multivariate process control methods and models. Therefore, developing advanced model-based multivariable control technology for plasma etching is highly desirable. Steady state and dynamic mathematical models of an etching process should be developed during process development. In previous generations this modeling included resist etch curves. Unlike purely empirical models, fundamental mathematical models are based on the physics and chemistry of the etching process and could provide valuable insight into the complex interactions among the process variables that directly affect the quality of the final product. For gate, contact, and trench processes, there is more than one measured variable (CD, SWA, depth etc), referred to as controlled variables (CV) so that wafer uniformity and line density sensitivity can be maintained.

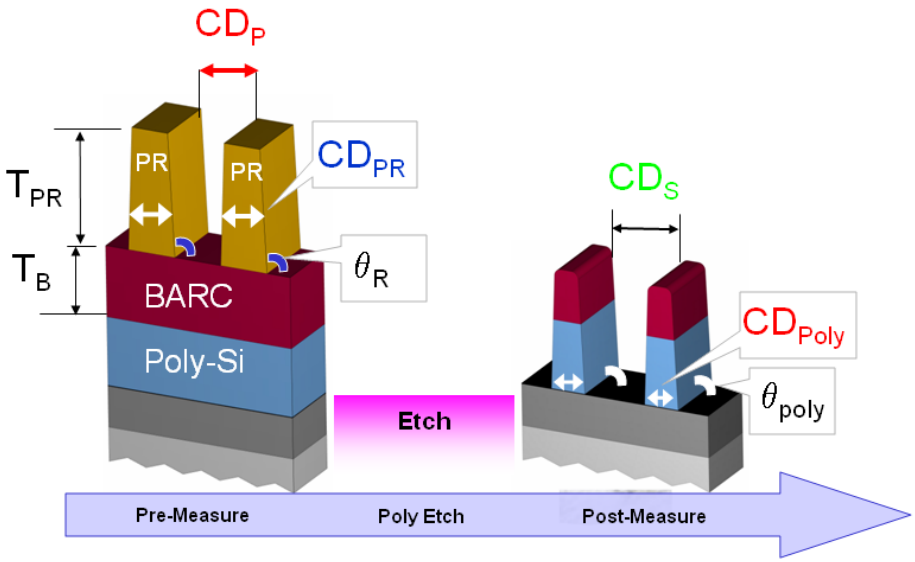


Figure 6.1 Litho-Etch sequence and structures

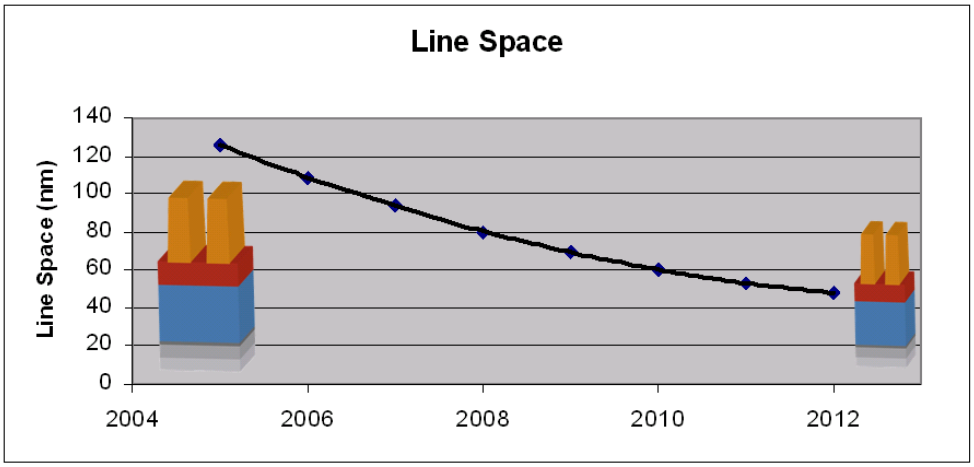


Figure 6.2 2007 IRTS calculated gate litho line space reduction

Controlling etch profile and linewidth requires adjusting more than one recipe parameter (step time, gas flow, pressure etc.), referred to as manipulated variables (MV) for both center and edge profile control. To address profile and uniformity control the system requires a multi-input multi-output (MIMO) approach that recognizes the interactions among CVs and MVs.

6.2 MEASUREMENT FOR CONTROL

Scatterometry is an optical method in which the 0th order reflected light from a regular array is collected and analyzed. The measured spectra are compared to an existing group of theoretical spectra, called a library. The theoretical spectra are typically generated prior to measurement using known information about the sample, including the film optical constants, the period of the features, and the profile characteristics. Parameters describing the sample are varied, and a theoretical spectrum is generated for each set of parameter values. As shown in Figures 6.3 and 6.4, a measured spectrum can then be matched to the set of theoretical spectra to determine what parameter values best describe that spectrum [103].

Before using scatterometry for process control, the scatterometry measurements should be correlated against respected reference metrology systems and physical electrical data. During the early implementation of IM-based gate CD control, the use of scatterometry was extensively studied, including the methods for comparisons [104], [103]. Often this reference metrology is the CD-SEM, since these systems are generally used for in-sector dispositioning and previously for L2L control. Each measurement approach has benefits and challenges [103]. For process control a scatterometry-based IM system has the advantage of measuring repeated patterns over a larger area such as a scribe line grating structure or a memory array pattered area.

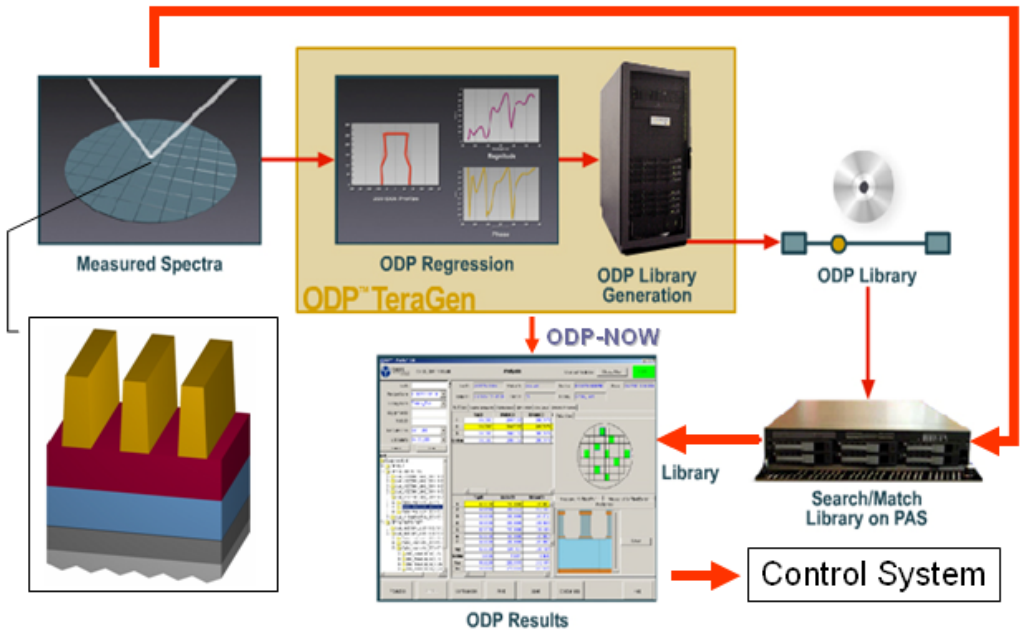


Figure 6.3 The sequence of scatterometry measurements

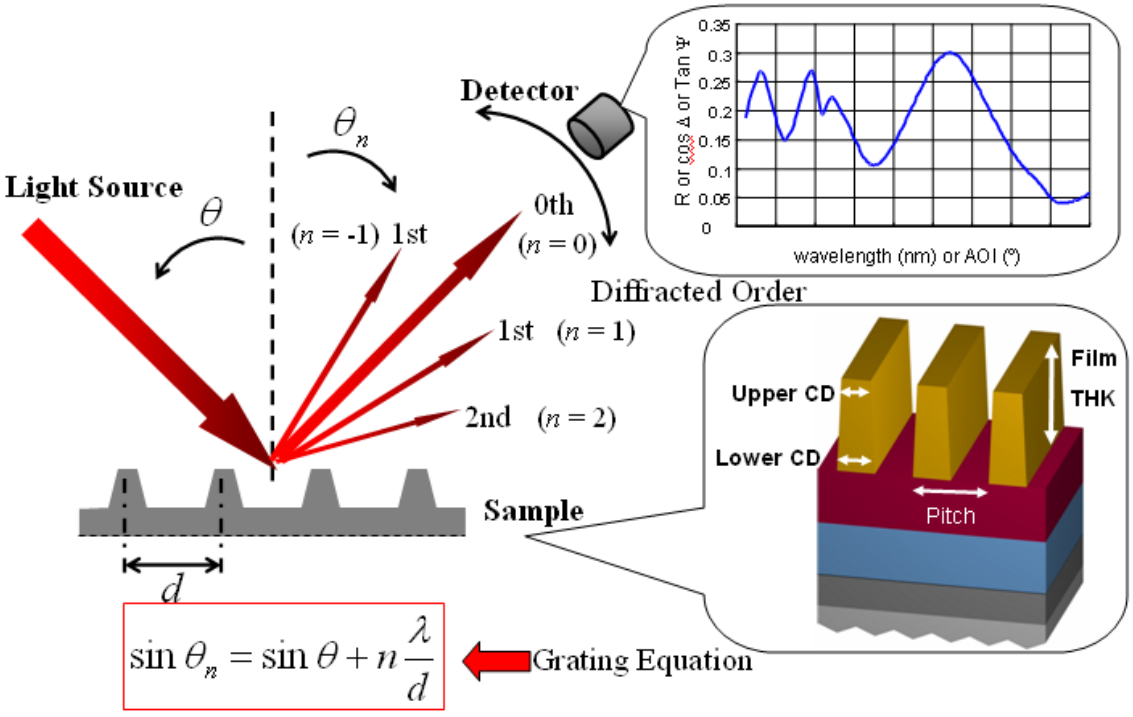


Figure 6.4 Scatterometry fundamental physics

This is advantageous since it allows for an effective averaging of the measurement to minimize line edge roughness or pattern defects skewing the measurement. In the case of CD-SEM only a single line or physical gate across a die, such as used in the example of Figure 6.4, is measured. This is potentially better for process development where more information about a specific structure is required. However, for process control accurate wafer-level measurements cannot be archived without averaging multiple CD-SEM measurements at each wafer site.

Before using scatterometry for process control, the scatterometry measurements should be correlated against respected reference metrology systems and physical electrical data. During the early implementation of IM-based gate CD control, the use of scatterometry was extensively studied, including the methods for comparisons [103], [104]. Often this reference metrology is the CD-SEM, since these systems are generally used for in-sector dispositioning and previously for L2L control. Each measurement approach has benefits and challenges [103]. For process control a scatterometry-based IM system has the advantage of measuring repeated patterns over a larger area such as a scribe line grating structure or a memory array pattered area. This is advantageous since it allows for an effective averaging of the measurement to minimize line edge roughness or pattern defects skewing the measurement. In the case of CD-SEM only a single line or physical gate across a die, such as used in the example of figure 4, is measured. This is potentially better for process development where more information about a specific structure is required. However, for process control accurate wafer-level measurements cannot be archived without averaging multiple CD-SEM measurements at each wafer site.

Profile has traditionally been measured in a laboratory using cross section (XSEM) or atomic force microscope (AFM). In the case of XSEM, in order to image a

profile, the wafer must be broken, thereby eliminating it from use as a real-time process control input and the time to take a measurement is often measured in hours. AFM is not destructive, however, it is susceptible to under-measuring high aspect ratio structures. Additionally, these systems have fairly long measurement times and the cost of periodic tip changes precludes dense sampling. Thus, SWA measurements can be more difficult to validate using traditional measurement methods because they are destructive and slow, making it difficult to take a large number of measurements [103]. Given this, scatterometry has the advantage that it can measure profiles, including retrograde SWA's (greater than 90 degrees) in a few seconds. However, little information has been published related to SWA accuracy of scatterometry to date.

Recently it has been shown that the addition of SWA measurements to the control model can improve CD control, both measured by a reference CD-SEM and by electrical measurements [10], [96], [105]. In Figure 6.5, the left graph shows the result of post CD control using only pre-CD while right graph shows the control performance of post-CD using a model with pre-CD as well as pre-SWA. These results indirectly show SWA can be measured accurately and more importantly, the value of measuring SWA.

6.3 PROFILE CONTROL IN LITHOGRAPHY

Any work to develop a closed loop etch process controller must start with an analysis of the incoming data. Sources of litho profile variation and current control schemes are first explored below in order to understand the areas of greatest need. Three metrics define the shape of a patterned resist feature: thickness, nominal CD, and side wall angle.

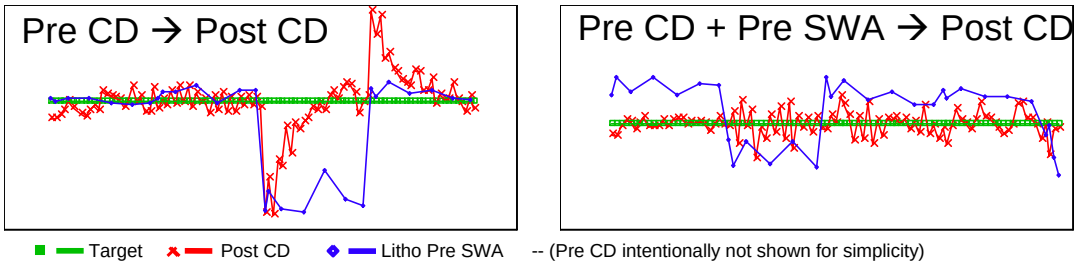


Figure 6.5 Control performance for post CD using pre CD only (left) and litho pre CD and SWA (right)

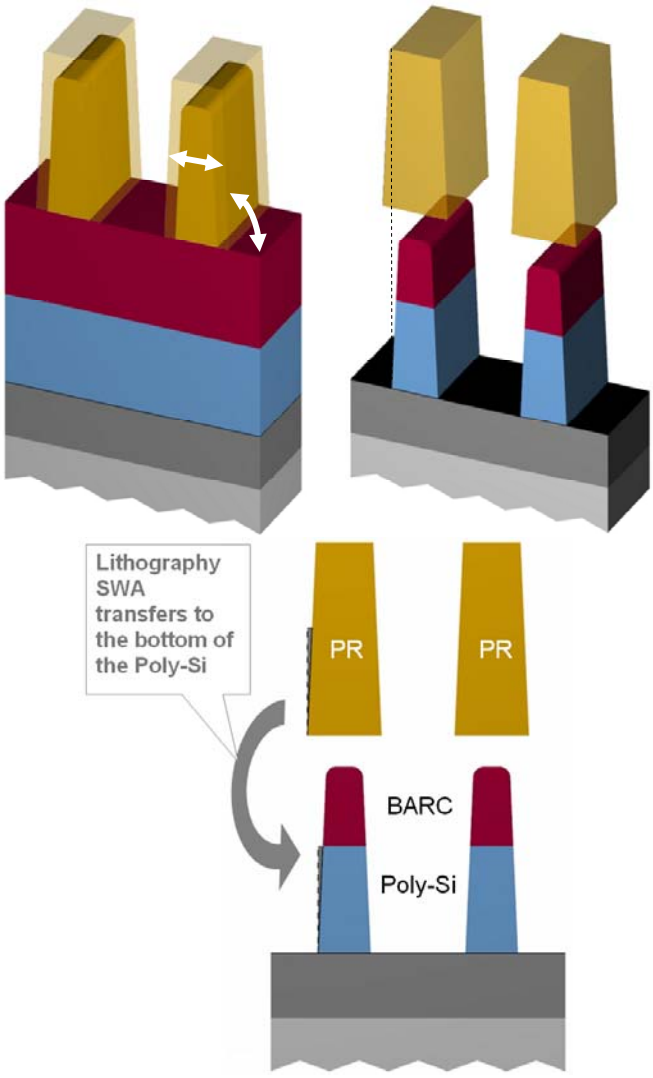


Figure 6.6 Litho SWA transfer to polysilicon SWA

Resist height is designed to be thick enough to provide etch resistance needed but thin enough to provide good imaging quality. CD swing curve effects further narrow the resist thickness operating control limits. The patterned feature height results from unpatterned resist thickness minus the erosion from developer dark loss and low level exposure light intensity. Coater spin speed adjustments are used to correct for wafer mean resist thicknesses variations. While many manufacturing fabs use unpatterned test wafers to monitor thickness, fabs with integrated scatterometry metrology are moving towards monitoring patterned resist height and BARC thickness on the product itself. The increased wafer sampling improves the thickness control capability for both coating layers and the reduction in scheduled downtime required to run monitors results in increased lithocluster utilization. Future trends in resist thickness do not indicate any need for increased levels of control beyond coater lot level adjustments. Overall, film thickness control does not pose significant challenges in the near future as sufficient, cost effective monitoring tools are available.

Post-litho develop inspect CD (DICD) control is meant to ensure post etch final inspect CD (FICD) control, and is typically monitored locally, across field, and globally. While local and across field CD variations are dominated by mask and exposure effects, and can sometimes be a significant portion of the overall gate CD uniformity budget, etch processes do not contribute to nor correct the existing litho non-uniformities. Global CD variation however, defined as across-wafer or with-in-wafer (WIW), wafer-to-wafer (W2W), or lot-to-lot (L2L) uniformity, stems from scanner, track, and etch tools. Periodic tool optimization to create uniform dose maps and Post Exposure Bake (PEB) temperature profiles are performed to ensure low across-wafer DICD variations. Additionally, advanced fabs have incorporated field by field dose adjustment maps to correct for post etch CD uniformity. Increasingly, fabs are now using multi-zone PEB

offset adjustments to optimize WIW and W2W level control of CD uniformity. Global CD uniformity on the most advanced lithocluster cells with CLEAN TRACK™ coater/developer have been demonstrated at <1.5nm, 3s on test monitors [106]. To maintain long term CD stability, exposure dose update systems have been implemented to correct lot-level CD variations [107]. As mentioned above, sites with integrated scatterometry metrology benefit from increased sampling and reduced CD monitor wafers. In addition, there is significant cycle time savings gained by reducing standalone CD-SEM sampling, helping to increase the overall throughput of photolithography step [108].

However, fabs cannot monitor CD independently of SWA. It is well known that Litho SWA can affect FICD despite having the correct DICD values. SWA variation can stem from multiple sources; currently, the most well known effect is exposure tool focus variation. Since SWA is not well monitored by top down CD-SEMs, process engineers seek to insulate themselves from focus variations by centering their process in the region of the Bossung curve where CDs are flat through a range of focus values. Scatterometry has an advantage by enabling profile control. Work is being done to separate the effects of dose and focus from litho profile measurements, with the ultimate goal being on product monitoring of equipment parameters [109]. While today's processes have adequate depth of focus (DOF) to withstand typical equipment and process variations, future requirements will be smaller. Projected shrinking process DOF and increased wafer warpage combine to exacerbate future SWA variations [20].

We examined the impact of litho profile on etch patterning and investigated what level of today's litho profile variation can be corrected through new etch APC techniques. If current process variations can be handled with new etch process control techniques, today's current system of lithography focus monitors can remain intact.

6.4 ANALYSIS OF ETCH PROCESS

For any semiconductor manufacturing process, it is important to understand the physical mechanisms that are at work. In the case of gate linewidth control, we should understand how the litho CD and SWA profile transfers to the final polysilicon CD and SWA, and how the etch process settings and chamber state interacts to form a final profile. The effect of various process parameters such as gas flows, (reactor) pressure, (wafer) temperature and (electrode) powers on polysilicon CD and SWA were studied.

A dual frequency, capacitively coupled plasma reactor with a top frequency of 60 MHz and a bottom frequency of 13.56 MHz was used in the present study. For resist etch, CF_4/O_2 plasma was used, along with HBr/O_2 plasma for polysilicon etching.

6.4.1 Impact of Resist Etch and Polysilicon Etch on SWA

The process interactions that take place in the etch process chamber allow for control of the CD and profile of the gate line by varying time, pressure, power, temperature, and gas flows. This control is critical since it has been observed that the SWA of the resist has significant effect on SWA of the polysilicon gate line. Figure 6.6 shows how a tapered resist litho profile is transferred to the post-etch polysilicon profile. Physically this is because during the resist etch process, once the edge of the foot of the resist and the bottom of the space between the gate lines clear etching of polysilicon starts. As etching of the polysilicon starts, the CD is set by the bottom of the resist. As etching continues the resist foot will clear exposing more polysilicon and the litho SWA will be transferred into the poly-silicon.

Figure 6.7 shows the interactions taking place during the resist etch step, and highlights the dissociation of CF4 into F and CFx, O2 into atomic O, and chemical reactions of resist with F, CFx and O.

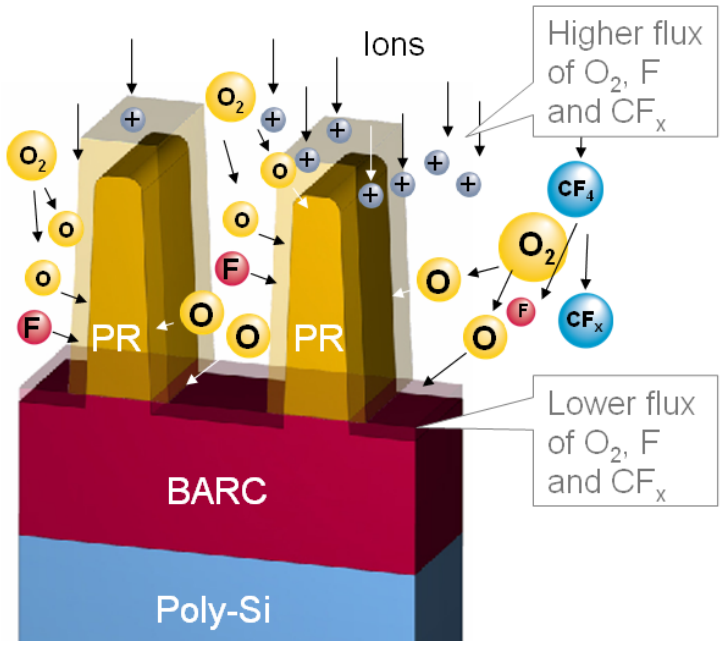


Figure 6.7 Photo resist etch process

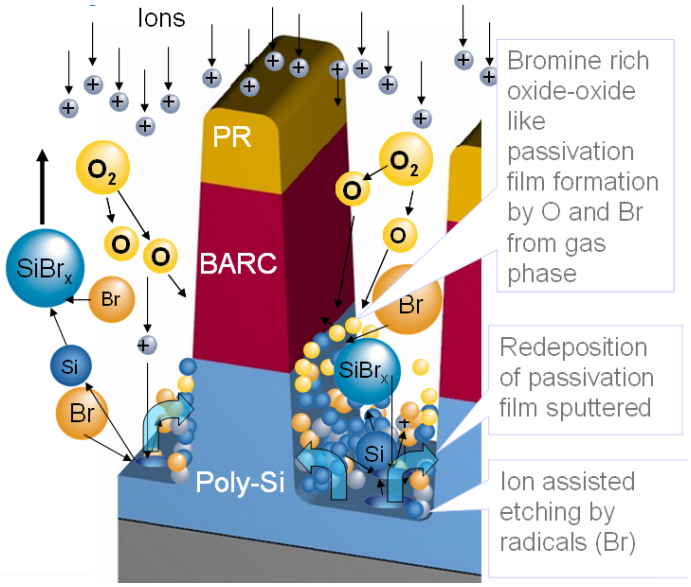


Figure 6.8 Poly-Si etch process

CF2 radicals are responsible for polymer deposition on the resist surface, this causes anisotropic (ion-assisted) etching. Atomic O and F isotropically etch resist. Since these neutral species have a broad angular distribution, the etch rate resulting from atomic O and F flux is higher at the top of the resist line than the bottom. Additionally, the top of the feature receives more neutral atoms compared to the bottom of the feature, causing tapering of resist lines. This study found that the etch rate and tapering of sidewall increases with the O2 flow rate. Additionally, if the O2 flow rate is increased with CF4 flow rate kept constant, the atomic O concentration increases, resulting in an increase of O flux to the wafer. The reaction rate of atomic O and F with the resist increases with higher temperature, which directly leads a higher etch rate.

If we look at increasing the chamber pressure, we will see several effects on the etch process. First, O2 and CF4 concentration in the gas phase increases while the electron temperature of the plasma decreases. Dissociation of O2 and CF4 into fragments depends on the electron temperature and their concentrations. While the concentration of F, CF2 and O increases, ion flux to the wafer decreases with pressure [110][111]. Therefore, etch rate increases with pressure in the resist etch step, since there is less passivation and more oxygen etching the resist line.

A second effect of increasing chamber pressure is that the mean free path and diffusion of atomic O and F to the bottom of the feature will decrease. This will decrease the flux from the top to the bottom of gate line. Since there is less etchant at the bottom of the gate line the top CD will decrease at a higher rate and thus, increased pressure will cause the SWA taper to increase.

Another mechanism to increase SWA taper is by increasing the bottom electrode power. By increasing the bottom electrode power the vertical etch rate of the top of the feature exposed to the plasma will increase since it is being bombarded with the ions.

Simultaneously more material sputters away from the space between the resist lines and redeposits on the sidewalls of resist. This leads to tapering of resist sidewalls.

For the polysilicon etch step, HBr/O_2 chemistry is typically used [112]. During this step, a bromine rich oxide-like passivation film forms on the sidewall and on the bottom of the space between the gate lines. Ions bombard the bottom of the feature, removing the passivation film from the bottom, while leaving the passivation layer on the sidewalls intact causing anisotropic etching. The concentration of O_2 used in this step affects both the SWA and CD. During the polysilicon etch step, atomic O concentration increases with O_2 flow and passivates the sidewalls and the thickness of passivation film increases with concentration. When the O_2 flow rate during polysilicon etch step is increased, the SWA bias will increase, where SWA bias is an increase in taper from litho to etch. This can be attributed to the re-deposition of the passivation film that was sputtered away from the bottom of the space, as is shown in Figure 6.8. Additionally, a chamber pressure increase during the polysilicon etch will cause the passivation layer to be sputtered away from the bottom of the feature and redeposited on the sidewall due to increasing O flux. This will cause an increase in SWA bias.

In summary, the litho CD and SWA affect the resist etch step and the resist etch features and polysilicon chemistry affect the etched CD and SWA of the polysilicon line. As shown these process interactions are complex and it is necessary to have a nonlinear model to control the final etched CD and SWA.

6.4.2 Mathematical Modeling with Design of Experiments

Experimental design requires a way to express the response of the system under study. Conceptually a generic representation of a model of that system is needed, independent of whether analytic, numeric, or experimental execution of the model

produces the response [113]. Based on physical analysis and engineering experience, design of experiments (DOE) can be built for statistical models which can connect MVs with each CV. Statistical models are generally used as they have the advantage of rapid development and use even if there is limited physical understanding of the system under study [114]. A general approach for DOE is described in Atkinson and Donev [115] while Hood and Welch [114] show DOE applications for semiconductor manufacturing. If the number of experiments increases, a more accurate model can be obtained, but at the expense of additional materials and time. Therefore, cost and availability will limit the wafers for DOE to the minimum. In order to reduce them as much as possible while minimizing inaccuracy, a well designed DOE is important. The most critical factor for such a DOE is the format of the predicted model (linear or non-linear). Generally in semiconductor manufacturing, a linear model is accepted. However, etch processes are not well described using a linear model because of the complex plasma behavior. Hence the relationship between the output characteristics of the model and the parameters must be explored [116]. Thus, nonlinear model formats with higher order and interaction terms are proposed before setting up the DOE. Literature review [113], [114], [115], [116], [117] and in depth discussions with process engineers would be a good way to predict a model format. This determines the combination of MVs in DOE runs. By using MVs with a range, the number of DOE wafers, and predicted model format, a DOE table for each wafer is created. Statistical software, such as JMP can be used to make this table. Care should also be taken to ensure that chamber conditioning is complete before running the DOE. The process modeling assumes the chamber state is stable between wafers and lots. After running experiments according to the DOE table, nonlinear models can be obtained by using a least squares technique via statistical software. The form of the base models can vary (linear, quadratic, with and without cross

terms, cubic, etc.). To reduce model complexity, terms with small coefficients can be deleted.

6.5 MULTIPLE INPUT MULTIPLE OUTPUT OPTIMIZATION AND CONTROL

Multiple inputs and multiple outputs (MIMO) control using dose and focus is becoming more common in lithography to control multiple variables. However, there are not many MIMO applications in etch processes, even though nonlinearity of etch processes can considerably affect those variables. For any etch process, there is more than one CV so that wafer uniformity and line density sensitivity can be maintained. Controlling multiple CVs such as etch profile, linewidth, and uniformity requires adjusting more than one MV for both center and edge profile control. In other words to control multiple CVs simultaneously, the system requires an MIMO approach that recognizes the interactions among CVs and MVs. The etch recipe will usually contain multiple steps, and each step can have different MVs. Obtaining a nonlinear model can be achieved by using data from a DOE and fundamental etch chamber behavioral knowledge. Theoretically, perfect control is achievable if there are the same numbers of control knobs as outputs. However, some of these output values might be constrained by manufacturing limits leading to less perfect control. Therefore, dynamic optimization algorithms should be added for MIMO control to support high volume manufacturing.

6.5.1 MMO Control Overview

Multivariable controllers are used when there are MIMOs in the control system. In some cases, the MIMO system can be decomposed into a series of SISO control loops (a multi-loop control system). Tuning of multi-loop control systems usually must be

done by trial and error [118] due to input-output interactions in the process. The principal difficulty with decomposing a multivariable system into a series of SISO control loops is that, while the dominant interactions are utilized, all other interactions are ignored. If there are interactions that are disrupting an otherwise efficient control system, then the effect of such interactions should be included in the controller design, especially for processes with many potential inputs and outputs, such as in plasma etching.

The preferred control strategy is a square system with each CV controlled by one MV, hence MIMO control is decomposed into several separate SISO control loops. However, plasma etching equipment inputs and outputs usually lead to a non-square system with unequal numbers of MVs and CVs. If a system has more CVs than MVs, it may not be possible to have perfect control of all CVs using the given MVs. Thus, some CVs may need to be eliminated to optimize performance. When you have more MVs than CVs, not all of the MVs need to be used for control and some MVs can be selected to maximize equipment performance.

6.5.2 Relative Gain Array Analysis

One way to evaluate whether single loop control is feasible is by using the relative gain array (RGA) proposed by Bristol [119]. Here one seeks the best SISO connections in a MIMO system. Rather than being based on system dynamics, it relies on a measure of the steady-state interactions between a given input/output pairing. By using the most sensitive SISO connections, control magnitudes can be minimized. The relative gain array can be obtained analytically, computationally, or experimentally. Consider a process with n controlled variables and n manipulated variables. The relative gain λ_{ij}

between a controlled variable, y_i, and a manipulated variable, u_j, is defined to be the dimensionless ratio of two steady-state gains:

$$\lambda_{ij} = \frac{\left[\frac{\partial y_i}{\partial u_j}\right]_{u_{k,k\neq j}}}{\left[\frac{\partial y_i}{\partial u_j}\right]_{y_{k,k\neq i}}} = \frac{gain(open-loop)}{gain(closed-loop)} \tag{6.1}$$

for $i = 1.2, \ldots, n$ and $j = 1, 2, \ldots, n$.

In Eq. (1) the symbol, $(\partial y_i / \partial u_j)_u$, denotes a partial derivative that is evaluated with all of the manipulated variables except u_j held constant. Thus, this term is the open-loop gain between y_i and u_j. Similarly, $\left(\partial y_i / \partial u_j\right)_y$ can be interpreted as a closed-loop gain that indicates the effect of u_j and y_i when all of the other feedback control loops are closed.

We define the relative gain array, denoted by Λ, as follows:

$$\Lambda = G \otimes (G^{-1})^T \tag{6.2}$$

where G is gain matrix and $\otimes$ denotes element-by-element multiplication.

The RGA has two important properties:

(a) It is normalized since the sum of the elements in each row or column is one.

(b) The relative gains are dimensionless, and thus not affected by choice of units or scaling of variables.

Based on experience with the technique, the system has minimal interactions for $0.5 < \lambda < 2.0$. Large or negative values of λ are problematic for multi-loop or multivariable control. Similar conclusions can be stated for more inputs/outputs.

Singular value analysis (SVA) is based on the steady-state multivariable gain matrix and can be used to solve the following important control problems [118]:

(a) selection of controlled and manipulated variables

(b) evaluation of the robustness of a proposed control strategy

(c) determination of the best multi-loop control configuration.

An important property of G is its singular values, which are the positive square roots of the eigenvalues of $G^T G$. The condition number, CN, of G is defined as the ratio of the largest and smallest nonzero singular values.

If G is singular, then it is ill-conditioned and by convention $CN \to \infty$. The concept of a condition number can also be extended to non-square matrices. The condition number provides useful information on the sensitivity of the matrix properties to variations in the elements of the matrices, which is related to control system robustness. Processes with poorly conditioned G matrices tend to require large changes in the manipulated variables in order to influence the controlled variables.

The preferred control strategy is a square system with each CV controlled by one MV, hence MIMO control is decomposed into several separate SISO control loops. However, plasma etching equipment inputs and outputs usually lead to a non-square system with unequal numbers of MVs and CVs. If a system has more CVs than MVs, it may not be possible to have perfect control of all CVs using the given MVs. Thus some CVs may need to be eliminated to optimize performance. When you have more MVs than CVs, not all MVs need to be used to control the CVs, and some MVs can be selected to maximize equipment performance. Relative gain array (RGA), non-square relative gain array (NRGA), and singular value analysis (SVA) can be applied together for selecting the best combinations of CVs and MVs for control. Chang and Yu [120] extended the RGA method of Bristol to non-square systems, which require the sum of elements in each row to be equal to unity (when there are more MVs than CVs), but the sum of column elements is not equal to one.

6.5.3 Recipe Optimization

While process insight can be obtained using interaction analysis, the previous methods use a linear gain matrix and the higher order and cross terms are ignored. If the etch process is highly nonlinear, however, linearization is most effective near the linearization center points. Another limitation of interaction analysis is that it does not consider constraints on the MVs. These constraints are based on a number of criteria, including machine and safety limitations, undesirable effects on uncontrolled outputs, and settings beyond which the unit process cost is known to increase [121]. Previous works have been published to overcome these limitations. Lachman-Shalem *et al.* applied CD control to lithography using a combination of genetic programming (GP) and nonlinear model predictive control (NMPC) [122]. The GP algorithm identifies the inputs that have the most impact on CD and generates the best multivariable empirical model for developing the NMPC controller. Chemali *et al*. investigated a MIMO R2R controller with linear models and kalman filtering scheme in order to control CD and side wall angle (SWA) in photolithography process by using dose and focus as control inputs [123]. Nonlinear optimization can be used to treat nonlinear relationships and constraints on the MVs and CVs in order to maximize performance of a multi-step etch process by changing the recipe after each run. The collection of MVs ("control knobs") is called a recipe [9]. A quadratic objective function utilizes weighting factors to prioritize each CV term in the objective function. In recipe optimization for run-to-run control, all of the MVs are allowed to vary for optimization, optimal points are found and verified, and controller is initialized with obtained optimal recipe [113]. Recipe optimization can be combined with run-to-run control to provide closed-loop control that maximizes a specific performance objective. Filters based on EWMA (exponentially weighted moving average) are used to update offset terms in each nonlinear equation of

CVs. Thus the controller is subsequently able to adjust the recipe to respond to equipment shifts and drifts [113].

6.5.4 Simulation Example

A commercial polysilicon gate etch process is used to demonstrate the effect of multivariable recipe optimization. The CVs are etch bias (EB), side wall angle bias (SWAB), and difference between center and edge CD (CDΔ). Incoming disturbance variables (DVs), namely the incoming CD and SWA, also affect CVs, so this interaction needs to be included in the model. After DOE and modeling steps are performed, all MVs are analyzed by interaction analysis. Table 6.1 shows all the variables used in the simulation. Since the actual MVs selected are proprietary they are not reported in this paper.

CVs		DVs		MVs
EB	y(1)	Pre_CD	x(1)	u(1)
SWAB	y(2)	Pre_SWA	x(2)	u(2)
CDΔ	y(3)	Pre_CDΔ	x(3)	u(3)

Table 6.1 Process variables for plasma etching

The three nonlinear models with higher order and interaction terms relating MVs and CVs are:

$$\begin{aligned} y(1) &= f_1(u(1), u(2), u(3), x(1), x(2), x(3)) + \text{offset}_1 \\ y(2) &= f_2(u(1), u(2), u(3), x(1), x(2), x(3)) + \text{offset}_2 \\ y(3) &= f_3(u(1), u(2), u(3), x(1), x(2), x(3)) + \text{offset}_3 \end{aligned} \tag{6.3}$$

To construct a quadratic objective function, target deviation CVs need to be defined:

$$\begin{aligned} t(1) &= x(1) - \text{target Post\_CD} \\ t(2) &= x(2) - \text{target Post\_SWA} \\ t(3) &= x(3) - \text{target Post\_CD}\Delta \end{aligned} \tag{6.4}$$

Using the models and target terms, a quadratic objective function suitable for nonlinear programming can be defined as:

$$\min_{u(1),u(2),u(3)} \{(\frac{t(1)-y(1)}{t(1)})^2 w_1 + (\frac{t(2)-y(2)}{t(2)})^2 w_2 + (\frac{t(3)-y(3)}{t(3)})^2 w_3\} \tag{6.5}$$

where the w_j are weighting factors. $u(1)$, $u(2)$ and $u(3)$ have upper and lower limits and can be included as inequality constraints.

$$\begin{aligned} a \le u(1) \le b \\ c \le u(2) \le d \\ e \le u(3) \le f \end{aligned} \tag{6.6}$$

An industrial reference data set with 450 wafers using W2W center CD control was used for simulation and typical disturbance patterns were identified. The simulation is designed to execute the same sequence as the gate etch process used in production. An IM unit measures the CD and SWA of multiple sites on each incoming wafer from litho. This information is then used by the optimizer unit to calculate the MVs. It determines the etch recipe by minimizing Eq. (6.9) with Eq. (6.10) using nonlinear programming. In particular, the MATLAB optimization tool box is used for this simulation. The details of nonlinear programming are beyond the scope of this paper. Next, the etching process is run with the MVs from the optimizer in the process unit. After the end of the etch process, the output CVs are measured using the IM unit. Errors (actual outputs minus model outputs) of each CV are calculated in an observer unit

and these values are used to update the MIMO model CV offsets using an exponentially weighted moving average (EWMA) filter [10]. Note that CV errors due to etch process disturbances [10] are assumed to be uncorrelated. These offset values are delivered to the optimizer unit to be used for compensating the disturbance for next run. This offset is used until a new update is available. This procedure continues until the final wafer is processed.

Figures 6.9 and 6.10 show simulation results for two CVs as a function of wafer number; the solid line shows the target for each CV.

Table 6.2 shows that the control performance measured by the mean square error (MSE) of center CD, SWA, and CDΔ is better than the current reference data which represents center CD control only. Using MIMO optimization allows much better control of SWA and CDΔ, compared to the "No Control" case of Figures 6.9 and 6.10, shown by the red x's for each wafer. It's important to note that the center CD control performance was not sacrificed, but improved in addition to adding two new control variables. Center CD of the reference data set and MIMO is not shown because they very near equal and already tightly control.

Data	MSE		
	CD control	MIMO	improvement
CD	0.43	0.33	23.8%
SWA	0.50	0.09	82.6%
CDΔ	2.07	0.70	66.3%

Table 6.2 Control performance for CD control vs. MIMO optimization

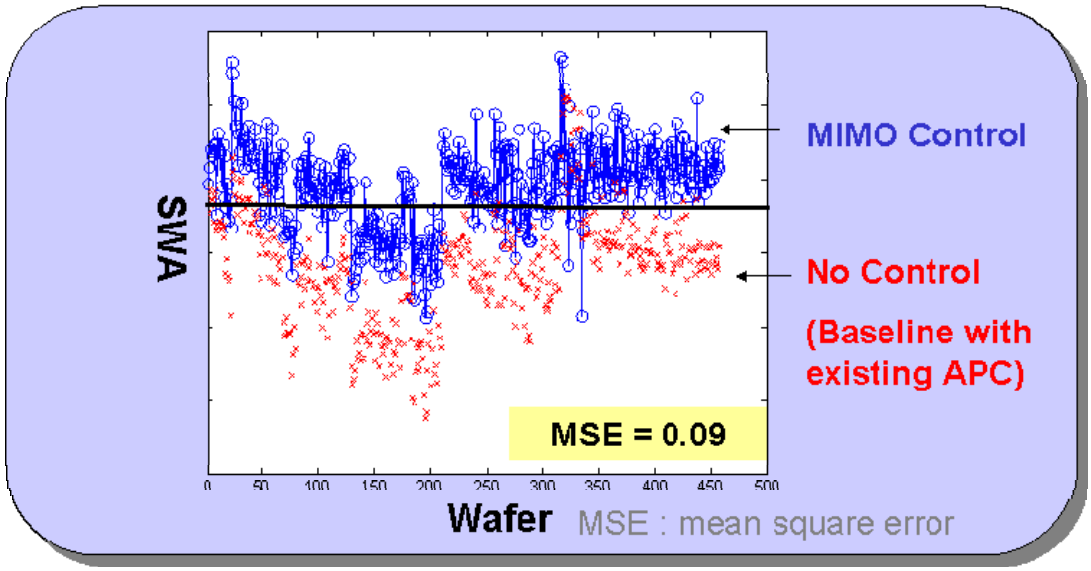


Figure 6.9 Simulation results: SWA

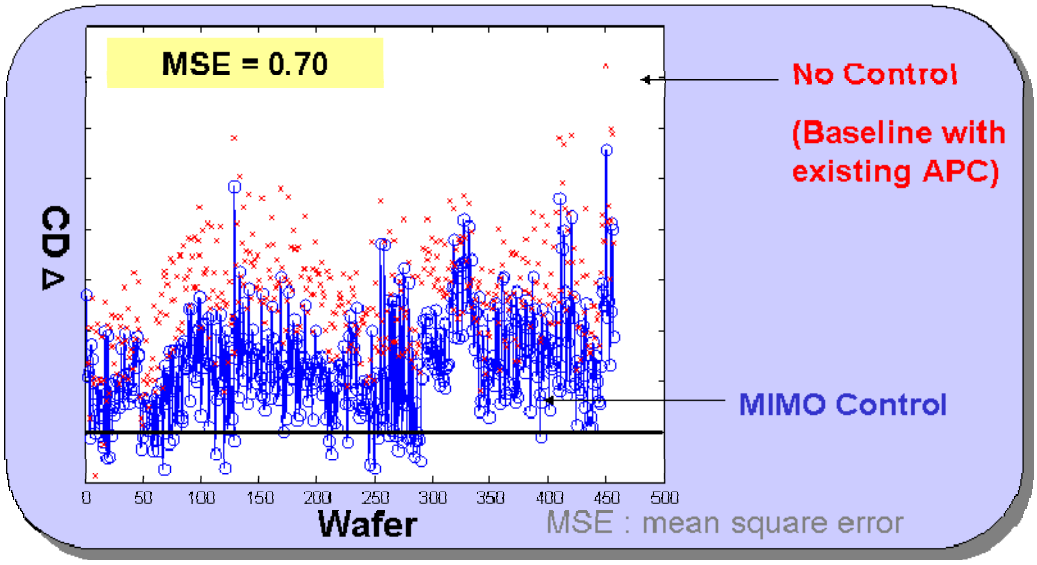


Figure 6.10 Simulation results: Post CDΔ

6.6 SUMMARY AND CONCLUSION

In this paper the importance of scatterometry-based integrated metrology was discussed as a data source for the litho cluster to adjust the across wafer CD uniformity and potentially control resist SWA by adjusting the photo process. As DOF windows shrink at future technology nodes, further investigation into the sources and contribution levels of litho SWA variation will be needed. Just as post-etch CD uniformity drives resist CD process control, similarly post-etch SWA may one day require more active resist SWA control in order to meet gate control requirements. Integrated metrology and scatterometry will no doubt be a key enabler of any additional control system needed for production.

Analysis of the etch process showed several interactions of etch process parameters and the incoming wafer state. These interactions are the basis of the etch process window and allow for active control of not only the post-etch CD, but the SWA and with-in wafer CD uniformity. These parameters are impacted by the incoming profile from the lithography step.

The complex interaction of litho CD and SWA with etch CD bias, SWA bias, and across wafer CD bias is impossible to decouple and understand without rigorous modeling and analysis. This requires a fundamental understanding of litho-etch process interactions. Additionally, understanding the interactions within each process step is critical since this is the level where litho and etch processes are controlled. By understanding these interactions a robust model can be built with simple extensions of current methods presently used in process development and process control. The multivariate methods that are used do not have to be proprietary to achieve world class production results.

In the MIMO simulations shown in this paper we have been able to show significant improvements in CD, SWA and within wafer CD uniformity. It is important to note that this potential improvement in process performance is possible with only a change to control methodology. Where the etch system takes advantage of the additional data provided by scatterometry and this data is used to control the knobs that were found during carefully executed DOEs. The fundamental interactions that enable this control exist within the multi-step plasma process.

As geometries shrink and new materials are used, multivariate control in etch is required to achieve better manufacturing performance. Better manufacturing performance can be accomplished by using new control methods, tighter integration of methods (control, measurement, and process), and understanding process control tuning; in some cases, new hardware (control knobs, MVs) is also needed to achieve the uniformity requirements.

CHAPTER 7

Conclusions and Future Work

7.1 CONCLUSIONS

The semiconductor industry has adopted the use of advanced process control (APC) to reach desired process goal of improved device yield. APC uses information about the materials to be processed, metrology data, and the desired output results to choose which model and control plan to employ. Metrology applications have become key enablers for many processing steps in semiconductor manufacturing. The economic advantage of effective metrology applications increases with the difficulty of the manufacturing process. The goal of this research intends to understand the relationship between metrology and APC in semiconductor manufacturing and develop an enhanced sampling strategy in order to maximize the value of metrology and control for critical wafer features.

Dynamic sampling is a method of changing the sampling frequency based on prior observations. The sampling rate is adjusted as the process progresses and increases when the process has moved away from target. Dynamic sampling is expected to be applicable to any process online since it is flexible and adjustable to the process dynamics. The performance of dynamic sampling depends on the characteristics of the process. When the process has little variation so every run is within the control limits, dynamic sampling has little effect but does not degrade the control performance. Dynamic sampling can be an optimal solution for the data set with step disturbances especially when there are data with large step disturbances.

In semiconductor manufacturing, high mix manufacturing environments have become more common because of the high capital costs associated with the tools and the limited capacity of the facility. The development of an optimal measurement sampling strategy under high mix manufacturing is required to optimize the sampling plan in order to quickly identify the sources of prediction errors and decrease the metrology cost and cycle time. Processing context describes the combination of equipment, product, and other factors that identify the processing environment that have measurable effects on the run. The online dynamic sampling algorithm is extended for high-mix manufacturing environments with non-threaded state estimations using Kalman filtering and qualification runs. Dynamic sampling has a synergistic effect with qualification runs for improving control performance.

Semiconductor manufacturing is characterized by a dynamic environment where the equipment variables change during the process step or between process steps. There are two major disturbances in semiconductor manufacturing – gradual drift and abrupt shift (step disturbance). Two distinct but isolated problems, drifts and step disturbances, are considered for robust control at the same time by using modified RDC with Bayesian detection. The modified robust drift cancellation (RDC) method takes actions to compensate for states having process drifts or normal noises, while the Bayesian EWMA (B-EWMA) method is preferred to control step disturbances. When an obvious drift is apparent in the process, the modified RDC with aggressive tuning has better performance; when step disturbances occur in the process, Bayesian detection can contribute to increased *Cpk* values.

With wafer-to-wafer (W2W) control the system can manage more process variation, which is compensated for by making a wafer-to-wafer recipe adjustment. Using integrated scatterometry with W2W feedback (FB) control can correct for step

disturbances and for variation of an etch chamber over time within a lot. It helps to refine and tune the control and measurement models for compensating for process model error quickly. Thus, W2W FB with integrated scatterometry shows better control performance as well as chamber matching than standalone lot-to-lot (L2L) FB by simulation using high volume manufacturing data set.

The etch system takes advantage of the additional data provided by scatterometry and this data is used to analyze the etch process. Analysis of the etch process shows several interactions of etch process parameters and the incoming wafer state. These interactions are the basis of the etch process window and allow for active control of not only the post-etch critical dimension (CD), but the side wall angle (SWA) and within-wafer CD uniformity. The complex interaction of CD, SWA, and CD uniformity is impossible to decouple and understand without rigorous modeling and analysis. This requires a fundamental understanding of litho-etch process interactions and the interactions within each process step. By understanding these interactions a robust model can be built with multivariate methods. As geometries shrink and new materials are used, multivariate control in etch is required to achieve better manufacturing performance.

7.2 RECOMMENDATIONS FOR FUTURE WORK

This work has provided some insight to understand the relationship between metrology and advanced process control in semiconductor manufacturing. As the semiconductor industry continues to grow and the integrated circuit technology advances, appropriate metrology applications and advanced process control algorithm will become more significant. It is increasingly important that information from metrology is used

properly with advanced process control. There are a number of topics that can be researched to extend this work in the future.

The process profit function was developed for economic optimal sampling. Since it is based on a unit process such as photolithography process, etch process, and thin film deposition process, it is hard to obtain accurate economic values for calculating the process profit function. It requires product cost, metrology cost, and yield loss. However, to estimate the economic values of those items in unit processes rather than whole manufacturing process is almost impossible. Thus, modifications of process profit function by using a different variable such as cycle time rather than cost would be good way to use it for real manufacturing.

Online dynamic sampling algorithm shows better control performance than uniform sampling in most cases. However, the results depend on the selection rules and tuning such as moving window size and confidence level which determines sampling frequency. Therefore an open question is how to update those parameters online. Historical data analysis is useful to obtain those but sometimes it is not realistic since the manufacturer might not want to keep all historical data in their database and there is not enough time to analyze it before resetting the run-to-run controller. Thus excluding the usage of an historical data set, alternative methods using online statistical analysis could help with development of more robust online dynamic sampling algorithm.

The applications of this online dynamic sampling algorithm together with qualification runs for high mix manufacturing were not tested with industrial data sets in this work. Thus, even though it shows better results, it does not guarantee successful applications to real manufacturing. Therefore testing this method with an appropriate industrial data set is a recommendation for future work.

When an obvious drift is shown in data set, the modified RDC with aggressive tuning has a better performance as expected. However, if there is no severe drift, aggressive tuning might overcompensate for process error, thus it shows worse performance than regular tuning. To detect drift using statistics like step disturbance detection with Bayesian approach will be very useful for drift compensation with an adjustment of tuning. Updating tuning parameters online according to process results is also a good option for better performance. Those two topics are recommended for future work.

Multiple inputs and multiple outputs (MIMO) control of a commercial polysilicon gate etch process is considered to control CD, SWA, and CD uniformity for better post etch profile and uniformity. However, more controlled variables (CVs) such as SWA uniformity and the thickness of etched layer could be added with more suitable manipulated variables (MVs). After additional design of experiment (DOE) and multivariate modeling, MIMO control and recipe optimization could be extended to have bigger system.

Recipe optimization with multivariable run-to-run control is used standard nonlinear programming using a quadratic objective function in order to find optimum points. Using less than five variables, the result and computation time to get a solution are almost the same whatever optimization methods are used. However, if complicated systems with many variables are considered, the optimization method might be an issue in terms of control performance and computing time. Therefore more studies about optimization with complex system are good topics for future work.

Bibliography

[1] T. F. Edgar *et al*., "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, Vol. 36, pp. 1567-1603, 2000.

[2] M. Quirk and J. Serda, *Semiconductor manufacturing technology*, Upper Saddle River, N.J.: Prentice Hall, 2001.

[3] G. S. May and C. J. Spanos, *Fundamentals of semiconductor manufacturing and process control*, John Wiley & Sons, Inc., 2006.

[4] H. H. Lee, *Fundamentals of microelectronics processing*, New York: McGraw-Hill, 1990.

[5] J. Moyne *et al.*, *Run-to-Run Control in Semiconductor Manufacturing*, CRC Press LLC, 2001.

[6] S. Joe Qin *et al*., "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *Journal of Process Control*, Vol. 16, No. 3, pp 179-191, 2006.

[7] S.W. Butler and J.A. Stefani, "Supervisory run-to-run control of a polysilicon gate etch using in situ ellipsometry," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 7, pp. 193–201, 1994.

[8] E. Sachs *et al*., "Run by run process control: Combining SPC and feedback control," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 8, pp. 26–43, 1995.

[9] S. W. Butler, "Process control in semiconductor manufacturing," *Journal of Vacuum Science Technology B*, Vol. 13 No. 4, pp. 1917-1923, 1995.

[10] H. Lee and T. F. Edgar, "Wafer-to-wafer feedback control for high volume polysilicon gate etch process using integrated scatterometry," *Proceedings of AEC/APC Symposium XIX*, Indian Wells, CA, September, 2007.

[11] K. Stoddard *et al.*, "Application of feed-forward and adaptive feedback control to semiconductor device manufacturing," *Proceedings of American Control Conference*, Baltimore, MD, June 1994.

[12] S. Leang et al., "A control system for photolithographic sequences," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, No. 2, pp. 191-207, 1996.

[13] G. Behm, *et al.*, "A Successful Integration of IBM's 300 mm Factory-level APC System with Tool-level Embedded Integrated Metrology and APC System," *Proceedings of 5th European AEC/APC Symposium*, Dresden, Germany, April, 2004.

[14] M. Sendelbach, *et al.*, "Integrated scatterometry in high volume manufacturing for polysilicon gate etch control," *Metrology, Inspection, and Process Control for Microlithography XX, Chas N. Archie, Editor, Proceedings of SPIE*, Vol. 6152, pp. 61520F, 2006.

[15] K. Bandy, *et al.*, "Polysilicon Gate Etch Process Control in High Volume Manufacturing Using Scatterometry-Based Integrated Metrology and Wafer-to-Wafer Process Control," *Proceedings of AEC/APC Symposium XVIII*, Westminster, CO, Sepember, 2006.

[16] S. W. Butler et al., "An intelligent model based control system employing in situ ellipsometry," *Journal of Vacuum Science and Technology*, A, Vol. 12, No. 4, pp. 1984-1991, 1994.

[17] E. D. Castillo and A. M. Hurwitz, "Run-to-run process control: Literature review and extensions," *Journal of Quality Technology*, Vol. 29, No. 2, pp. 184-196, 1997.

[18] B. Bunday et al., "Realizing value-added metrology," *Metrology, Inspection, and Process Control for Microlithography XVIII*, Chas N. Archie, Editor, *Proceedings of SPIE*, vol. 6518, pp. 65181K-1-65181K-21, 2007.

[19] M. Sendelbach and C. Archie, "Scatterometry measurement precision and accuracy below 70 nm," *Metrology, Inspection, and Process Control for Microlithography XVII, Daniel J. Herr, Editor, Proceedings of SPIE*, Vol. 5038, pp. 224-238, 2003.

[20] http:// www.itrs.net / Links / 2007Winter / 2007_Winter_Presentations / 09_Litho_2007_JP.pdf (2007 Lithography ITRS Roadmap Update)

[21] R. K. Nurani *et al.*, "In-Line Defect Sampling Methodology in Yield Management: An Integrated Framework," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, No. 4, pp. 506-517, 1996.

[22] R. Williams *et al.*, "Challenging the paradigm of monitor reduction to achieve lower product cost," *Proceedings of The 10th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Boston, MA, September, 1999.

[23] R. Williams *et al.*, "Optimized sample planning for wafer defect inspection," *Proceedings of IEEE International Symposium on Semiconductor Manufacturing Conference*, Santa Clara, CA, October, 1999.

[24] R. Babikian and C. Engelhard, "Statistical methods for measurement reduction in semiconductormanufacturing," *Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Boston, MA, 1998.

[25] C. Cheng and J. Moyne, "Intelligent Metrology and Control Strategies Applied to Lithography Processes," *Proceedings of AEC/APC Symposium XIV*, Snowbird, UT, September, 2002.

[26] T. Sonderman and M. Purdy, "Automated fab-wide control as a mechanism for maximizing IC manufacturing efficiency," *Proceedings of AEC/APC Symposium XVII*, Indian Wells, CA, September, 2005.

[27] A. Holfeld *et al*., "A fab-wide APC-sampling application," *Proceedings of 7th European AEC/APC Symposium*, Aix-en-Provence, France, March, 2006.

[28] R. C. Elliott *et al*., "Sampling plan optimization for detection of lithography and etch CD process excursions," *Metrology, Inspection, and Process Control for Microlithography XIV*, Neal T. Sullivan, Editor, *Proceedings of SPIE*, vol. 3998, pp. 527-536, 2000.

[29] R. K. Nurani *et al*., "Role of in-line sampling methodology in yield managemenet," *Proceedings of IEEE/UCS/SEMI International Symposium on Semiconductor Manufacturing*, Austin, TX, 1995.

[30] R. C. Elliott *et al*., "Critical dimension sample planning for sub-0.25 micron processes," *Proceedings of The 10th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, Boston, MA, September, 1999.

[31] C. Mouli and M. J. Scott, "Adaptive Metrology Sampling Techniques Enabling Higher Precision in Variability Detection and Control," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 12-17, 2007.

[32] A. Bousetta and A. J. Cross, "Adaptive Sampling Methodology for In-Line Defect Inspection," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 1-7, 2005.

[33] N. Patel, "Auto-Tuning and Dynamic Sampling under EWMA Control," *Proceedings of AEC/APC Symposium XIX*, Indian Wells, CA, September, 2007.

[34] J. Mao *et al.*, "Adaptive metrology sampling technique and its application to defect metrology," *Proceedings of AEC/APC Symposium XIX*, Indian Wells, CA, September, 2007.

[35] S. L. Riley, "Optical inspection of wafer using large-area defect detection and sampling," *Proceedings of IEEE International Workshop on Defect and Fault Tolerance in VLSI Systems*, Dallas, TX, 1992.

[36] D. B. Sullivan *et al.*, "Overlay metrology sampling capability analysis and implementation in manufacturing," *Proceedings of IEEE Conference and Workshop on Advanced Semiconductor Manufacturing*, pp. 208-212, 2004

[37] D. L. Dance and D. W. Jimenez, "Applications of cost-of-ownership," *Semiconductor International Magazine*, pp. 6-7, 1994

[38] V. A. Ames *et al*., *Semiconductor Manufacturing Productivity Overall Equipment Effectiveness (OEE) Guidebook Revision 1.0*, SEMATECH, 1995.

[39] SEMI E35-0701, *Cost of ownership for semiconductor manufacturing equipment metrics*, 2001.

[40] SEMI E79-0200, *Standard for definition and measurement of equipment productivity*, 2000.

[41] D. L. Dance *et al*., "Modeling the cost of ownership of assembly and inspection," *IEEE Transactions on Components, Packaging, and Manufacturing Technology – Part C*, Vol. 19, No. 1, pp. 57-60, 1996.

[42] D. L. Dance, "Modeling test cost of ownership," *Proceedings of 3rd International Conference on the Economics of Design, Test, and Manufacturing*, Austin, TX, 1994.

[43] D. L. Dance and P. Bryson, "Cost of ownership for inspection equipment," *Integrated Circuit Metrology, Inspection, and Process Control VII, Michael T. Postek, Editor, Proceedings of SPIE*, Vol. 1926, pp. 249-256, 1993.

[44] J. Stuber, "Device dependant run-to-tun control of transistor critical dimension by manipulating photolithography exposure settings," *Proceedings of AEC/APC Symposium XII*, Vail, CO, September, 2000.

[45] A. J. Pasadyn, *et al,* "Scheduling semiconductor manufacturing processes to enhance system identification," *Journal of Process Control*, to be published.

[46] J. Wang and P. He, "A Bayesian approach for disturbance detection and classification and its application to state estimation in run-to-run Control," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 20, No. 2, pp. 126-136, 2007.

[47] J. Wang and C. Bode, "Bayesian statistics and its application in semiconductor manufacturing," *Proceedings of AEC/APC Symposium XV*, Colorado Springs, CO, September, 2003.

[48] D. A. Hsu, "A Bayesian robust detection of shift in the risk structure of stock market returns," *Journal of the American Statistical Association*, Vol. 77, No. 377, pp. 29-39, 1982.

[49] E. A. Feinberg and A. N. Shiryaev, "Quickest detection of drift change for Brownian motion in generalized Bayesian and minimax settings," *Statistics & Decisions*, Vol. 24, No. 4, pp. 445–470, 2007.

[50] N. J. Gordon *et al*., "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEEE Radar and Signal Processing*, Vol. 140, No. 2, pp. 107-113, 1993.

[51] W. M. Bolstad, *Introduction to Bayesian Statistics*, Wiley-Interscience, 2004.

[52] S. Ross, *A first course in probability*, Upper Saddle River, N.J.: Prentice Hall, 2005

[53] G. Casella and R. L. Berger, *Statistical inference*, Brooks/Cole, 2001.

[54] J. Stefani and C. Gross, "Run by run process control of bilayer liners for copper interconnects," *Proceedings of AEC/APC Symposium XIV*, Snowbird, UT, September, 2002.

[55] J. Stefani and C. Murali, "Run by run APC deployment in 65nm thin film module," *Proceedings of AEC/APC Symposium XVIII*, Westminster, CO, Sepember, 2006.

[56] J. Stuber *et al*., "Gate CD control for the 65nm technology node," *Proceedings of AEC/APC Symposium XVIII*, Westminster, CO, Sepember, 2006.

[57] A. J. Pasadyn and T. F. Edgar, "Observability and State Estimation for Multiple Product Control in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 18, No. 4, pp. 592-604, 2005.

[58] A. J. Pasadyn, *et al.*, "Adaptive control of multiple product processes," *Process Control and Diagnostics, M. L. Miller and K. A. Ashtiani, Editors, Proceedings of SPIE*, Vol. 4182, pp. 22–30, 2000.

[59] A. J. Pasadyn *et al*., "Online parameter estimation for simultaneous control of multiple products and processes," *Proceedings of AEC/APC Symposium XII*, Vail, CO, September, 2000.

[60] A. J. Toprac *et al.*, "Solving the high-mix control problem," *Proceedings of AEC/APC Symposium XVI*, Westminster, CO, Sepember, 2004.

[61] A. J. Toprac and Y. Wang, "Advanced method for run-to-run control of photolithography overlay registration in high-mix semiconductor production,"

Data Analysis and Modeling for Process Control II, I. Emami, C. P. Ausschnitt, and K. W. Tobin, Editors, Proceedings of SPIE, Vol. 5755, pp. 1–7, 2005.

[62] D. King *et al*., "Feedback model evaluation of high-mix product manufacturing," *Metrology, Inspection, and Process Control for Microlithography XX, Chas N. Archie, Editor, Proceedings of SPIE*, Vol. 6152, pp. 615242, 2006.

[63] C. A. Bode *et al*., "Run-to-run control and state estimation in high-mix semiconductor manufacturing," *Annual Reviews in Control*, Vol. 31, No. 2, pp. 241-253, 2007.

[64] O. A. Vanli *et al*., "Model costext selection for run-to-run control," *Proceedings of AEC/APC Symposium XIX*, Indian Wells, CA, September, 2007.

[65] X. Wan *et al*., "Overlay advanced Process control for foundry application," *Metrology, Inspection, and Process Control for Microlithography XVIII, R. M. Silver, Editor, Proceedings of SPIE*, Vol. 5375, pp. 735-743, 2004.

[66] S. K. Firth, *et al*., "Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor process," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 19, No. 3, pp. 298-315, 2006.

[67] J. Wang, *et al*., "On state estimation if high-mix semiconductor manufacturing using a singular Gauss Markov model," *IEEE Transactions on Semiconductor Manufacturing*, to be published.

[68] A. V. Prabhu, T. F. Edgar, "A new state estimation method for high-mix semiconductor manufacturing process," *Journal of Process Control*, to be published.

[69] V. M Martinez *et al*., "Adaptive estimation and control of overlay tool bias," *Proceedings of AEC/APC Symposium XIV*, Snowbird, UT, September, 2002.

[70] D. Crow *et al*., "Application of feedforward reticle – offset for overlay APC in high part count fab," *Metrology, Inspection, and Process Control for Microlithography XVI, D. J. Herr, Editor, Proceedings of SPIE*, Vol. 4689, pp. 1151-1161, 2002.

[71] W. F. Ramirez, *Process Control and Identification*. Academic Press, 1994.

[72] S. A. Middlebrooks, "Use of observability matrix to design and optimize metrology measurement patterns," *Proceedings of AEC/APC Symposium XVI*, Westminster, CO, Sepember, 2004.

[73] F. L. Lewis, *Optimal Estimation with an Introduction to Stochastic Control Theory*, New York: John Wiley and Sons, 1986.

[74] C. K. Hanish, "Run-to-run state estimation in system with unobservable states," *Proceedings of AEC/APC Symposium XVII*, Indian Wells, CA, September, 2005.

[75] S. A. Middlebrooks and J. B. Rawlings, "Model predictive control of SiGe thin film chemical-vapor deposition," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 20, No. 2, pp. 114-125, 2007.

[76] S. A. Middlebrooks, "Optimal model-predictive control of overlay lithography implemented in an ASIC fab," *Advanced Process Control and Automation, M. Hankinson, C. P. Ausschnitt, Editor, Proceedings of SPIE*, Vol. 5044, pp. 12-23, 2003.

[77] G. Box and A Luceno, *Statistical control by monitoring and feedback adjustment*, John Wiley & Sons, Inc., 1997.

[78] H. B. Nembhard and M. S. Kao, "Adpative forecast-based monitoring for dynamic systems," *Technometrics*, Vol. 45, No. 3, pp. 208-219, 2003.

[79] G. Box and T. Kramer, "Statistical process monitoring and feedback adjustment – a discussion," *Technometrics*, Vol. 34, No. 3, pp. 251-267, 1992.

[80] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, England: J. Wiley and Sons, LTD, 2000

[81] A. Hu, *An Optimal Bayesian process control for flexible manufacturing processes*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[82] G. Box, and G. Tiao, "A further look at robustness via Bayes's theorem," *Biometrika*, Vol. 49, pp. 419-433, 1962

[83] V. Venkatasubramanian *et al*., "Review of process fault diagnosis—part I: quantitative model-based methods," *Computers Chemical Engineering*, Vol. 27, No. 3, pp. 293–311, 2003.

[84] V. Smidl, and A. Quinn, *The variational Bayes method in signal processing*, Berlin, Germany: Springer, 2006.

[85] J. Musacchio, *Run-to-run control in semiconductor manufacturing*, M.S.Thesis, University of California, Berkeley, 1997

[86] T. E. Tuncer, "Causal and Stable FIR-IIR Wiener Filters," *Proceedings of IEEE Workshop on Statistical Signal Process,* St. Louis, MO, September, 2003.

[87] Y. Yamamoto *et al.*, "Optimizing FIR approximation for discrete-time IIR filters," *IEEE Signal Processing Letter*, Vol. 10, No. 9, pp. 273-276, 2003.

[88] J. Scott Goldstein *et al.*, "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Transaction on Information Theory*, Vol. 44, No. 7, pp. 2943-2959, 1998.

[89] A. Chen and R.S. Guo. "Age-based double EWMA controller and its application to CMP processes," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 14, No.1, 2001.

[90] J. Musacchio et al., "On the utility of run to run control in semiconductor manufacturing," *Proceedings of IEEE International Symposium on Semiconductor Manufacturing Conference*, San Francisco, CA, October, 1997.

[91] F. Callier, and C. Desoer, *Linear System Theory*, New York: Springer Verlag, 1991.

[92] S. Boyd *et al*., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia: Society for Industrial and Applied Mathematics, 1994.

[93] A. Ingolfsson and E. Sachs, "Stability and sensitivity of an EWMA controller," *J. Quality Technol.*, Vol. 25, No. 4, pp. 271–287, 1993.

[94] P. Rohr and S. Hodges, "Implementation of scatterometry as a Production Line Monitor," KLA-Tencor Yield Management Seminar, July, 2004.

[95] J. Allgair, *et al*., "Spectroscopic CD offers higher precision metrology for sub-0.18 μm linewidth control," *KLA-Tencor, Yield Management Solutions*, pp. 8-13, 2002.

[96] K. Bandy, *et al*., "Polysilicon gate etch process control using integrated scatterometry and multiple-input wafer-to-wafer process control," *Proceedings of AEC/APC Symposium XIX*, Palm Springs, CA, September, 2007

[97] B. A. Rashap, *et. al*., "Control of semiconductor manufacturing equipment: real-time feedback control of a reactive ion etcher," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 8, No. 3, pp. 286-297, 1995.

[98] W. Kang and J. Mao, "Robust control of lithographic process in semiconductor manufacturing," *Data Analysis and Modeling for Process Control II, Proceedings of SPIE*, Vol. 5755, pp. 29-36, 2005.

[99] K. Lensing and B. Stirton, "Integrated metrology and wafer-level control," *Semiconductor International*, Vol. 29, No. 6, pp. 44-54, 2006.

[100] D. A. Barad, *et al.*, "Strong tracking filter for control of run-to-run semiconductor fabrication," *Proceedings of AEC/APC Symposium XVIII*, Denver, Co, October, 2006.

[101] D. H. Zhou, *et al.*, "A Suboptimal multiple fading factor extended Kalman filter," *Chinese Journal of Automation*, Vol. 4, pp. 145-152, 1991.

[102] R. P. Good and S. J. Qin, "On the stability of MIMO EWMA run-to-run controllers with metrology delay," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 19, No. 1, pp. 78-86, 2006.

[103] M. Sendelbach *et al.*, "Comparison of scatterometry, atomic force microscope, dual beam system, and XSEM to measure etched via depths," *Metrology, Inspection, and Process Control for Microlithography XIX*, *Richard M. Silver, Editor*, *Proceedings of SPIE*, Vol. 5752, pp. 272-287, 2005.

[104] M. Sendelbach *et. al.*, "Feedforward of mask open measurements on an integrated scatterometer to improve gate linewidth control," *Metrology, Inspection, and Process Control for Microlithography XVIII*, *Richard M. Silver, Editor, Proceedings of SPIE*, Vol. 5375, pp. 686-702, 2004.

[105] H. Lee and M. Funk, "Robust W2W control using integrated scatterometry in high volume manufacturing polysilicon etch," *Proceedings of AEC/APC Asia Symposium*, Kyushu, Japan, November, 2007

[106] A. Hazelton, *et al.*, "Immersion lithography in mass production: latest results for Nikon immersion exposure tools," *4th International Immersion Symposium Program*, Keystone, CO, October, 2007.

[107] C. P. Ausschnitt *et al*., "CD control of low k-factor step and scan lithography," *Optical Microlithography XIV*, *Christopher J. Progler, Editor*, *Proceedings of SPIE*, Vol. 4346, pp. 293-302, 2001.

[108] Levin *et al.*, "Integrated metrology: solutions for APC," *Proceedings of AEC/APC Asia Symposium*, Kyushu, Japan, November, 2007.

[109] K. Lensing *et al*., "Lithography process control using scatterometry metrology and semi-physical modeling," *Metrology, Inspection, and Process Control for Microlithography XVIII*, *Chas N. Archie, Editor, Proceedings of SPIE*, Vol. 6518, pp. 651804-1-651804-12, 2007.

[110] M. A. Lieberman and A. J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, John Wiley & Sons, Inc., New York, 1994.

[111] S. Rauf, "Model for photoresist trim etch in inductively coupled CF4/O2 plasma," *J. Vac. Sci. Technol. B*, Vol. 22, pp. 202-211, 2004.

[112] M. Armacost *et al.*, "*Plasma-etching processes for ULSI semiconductor circuits," IBM Journal of Research and Development, Plasma processing*, Vol. 43, No. 1, pp. 39-72, 1999.

[113] D. S. Boning and P. K. Mozumder, "DOE/Opt: A system for design of experiments, response surface modeling, and optimization using process and device simulation," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 7, No. 2, pp. 233-244, 1994.

[114] R. Guo and E. Sachs, "Modeling, optimization and control of spatial uniformity in manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 6, No. 1, pp. 41-57, 1993.

[115] A. C. Atkinson and A. N. Donev, *Optimal Experimental Designs*, Oxford University, New York, 1992.

[116] S. J. Hood and P. D. Welch, "Experimental design issues in simulation with examples from semiconductor manufacturing," *Proceedings of the 24th conference on Winter simulation,* pp. 255-263, 1992.

[117] G. S. May *et al.*, "Statistical experimental design in plasma etch modeling," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 8, No. 3, pp. 286-297, 1995.

[118] D.E. Seborg *et al.*, *Process dynamics and control*, 2nd ed., Wiley, New York, 2004.

[119] E.H. Bristol, "On a new measure of interactions for multivariable process control", *IEEE Transactions Automatic Control*, Vol. AC-11, pp. 133-135, 1966.

[120] J. Chang and C. C. Yu, "The relative gain for non-square multivariable systems", *Chemical Engineering Sci*ence, Vol. 45, pp. 1309-1323, 1990.

[121] A. E. Gower-Hall *et al.*, "Model-based uniformity control for epitaxial silicon deposition," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 15, No. 3, pp. 295-309, 2002.

[122] S. Lachman-Shalem *et al.*, "Nonlinear modeling and multivariable control of photolithography," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 5, pp. 310-322, 2002.

[123] C. El Chemali *et al.*, "Run-to-run critical dimension and side wall angle lithography control using the PROLITH simulator," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 17, pp. 388-401, 2004.

Vita

Hyung Joo Lee was born in Seoul, South Korea on April 14th, 1979, the son of Eun Kyu Lee and Jung Hee Lee. After graduating from Karak High School in 1998, he entered Seoul National University in Seoul, South Korea. For 26 months from January, 2000 he served as ammunition clerk near general outpost (GOP) for protecting his country against North Korea. In February of 2004 he received a Bachelor of Science in Chemical Engineering from Seoul National University. He then entered the graduate school of The University of Texas at Austin. Hyung Joo has taken an internship at Tokyo Electron America at Austin for 18 months while continuing to pursue his Ph.D. from The University of Texas. He will be employed as a process technology development engineer at Intel Corporation in Hillsboro, Oregon from October, 2008.

Permanent address: 3-105 Sindonga APT, Haengdang-2-dong, Seongdong-gu,
Seoul, South Korea, 133-766

This dissertation was typed by the author.