Founded 1905

CONTROL OF SEMICONDUCTOR MANUFACTURING: CMP & THICKNESS VARIATIONS

BY

LI DA (M.ENG.)
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2003

Acknowledgments

I would like to express my deepest gratitude to my supervisor, assistant professor Arthur Tay for his guidance through my M.E. study. Without his gracious encouragement and generous guidance, I would not be able to finish my work. His unwavering confidence and patience have aided me tremendously. His wealth of knowledge and accurate foresight have greatly impressed and benefited me. I am indebted to him for his care and advice not only in my academic research but also in my daily life. I would like to extend special thanks to Dr. Abdullah Al Mamun. His comments, advice, and inspiration played an important role in this piece of work.

Special gratitude goes to my friends and colleagues. I would like to express my thanks to my fellow partner, Mr. Ganesh, and my friends, Mr Lu Xiang, Ms Fu Jun, Ms Zhao Shao, Mr Yang Yongsheng, and many others in the advanced control technology lab. I enjoyed very much the time spent with them. I also appreciate the national university of Singapore for the research facilities and scholarship.

Finally, this thesis would not have been possible without the support from my family. The encouragement from my parents has been invaluable. My mother, Heying Zhou, is the one who deserves my deepest appreciation. I would like to dedicate this thesis to them and hope that they would enjoy it.

Li, Da

July, 2003

Contents

List of Figures

List of Tables

Summary

The most important variable in the semiconductor manufacturing process is the linewidth or critical dimension (CD), which is the single variable with the most direct impact on the device speed and performance. In this thesis, we discuss one of the key parameters that has an impact on the final CD : the depth-of-focus (DOF). Parameters and processes that affect the DOF will be discussed. One of the processes is chemical mechanical polishing (CMP) process, which has become an indispensable semiconductor processing module used in fabrication facilities worldwide to achieve the global planarization. This capability is absolutely required to increase the number of wiring levels in the integrated circuits without the limitation of DOF issue. To reduce the within-wafer-nonuniformity (WIWNU) in the CMP process, a combination of statistics process control (SPC) and advanced process control (APC), namely run-to-run control (R2R), is investigated. The lack of in-situ measurements of the products qualities of interest, in this case, the surface thickness uniformity, makes run-to-run control the only viable scheme. Run-to-run control is further necessitated by the non-stationery nature of most semiconductor processes. The literature contains many variations of R2R control schemes to control the CMP process such as Exponential Weighted Moving Average (EWMA), Predictor Corrector Controller (PCC), Optimizing Adaptive Quality Controller (OAQC), Model Predictive Controller (MPC). In this thesis, we analyze the performance of these R2R control schemes and propose a self-tuning predictor-corrector controller (SPCC). This allows for automatically adjusting the forecasting

parameters in the face of changing process noise and disturbances. Simulation results depicts an order of magnitude improvement in terms of removal rate and WIWNU when compared to conventional R2R controllers.

A related problem that affects the DOF is wafer warpage. We propose in this thesis an approach to predict the wafer warpage by monitoring the bake plate temperature during the baking of the wafer in the microlithography sequence.

Chapter 1

Introduction

1.1 Motivation

Advances in modelling and control are required to meet future technical challenges in microelectronics manufacturing. The implementation of closed-loop control on key unit operations has been limited due to a dearth of suitable in-situ measurements, variations in process equipments and wafer properties. An advanced control framework for integrating factory control and equipment scheduling, supervisory control, feedback control, statistical process control and fault detection/diagosis in microelectronics is urgently needed to meet the development of the whole chip-based industry.

In this thesis, we mainly discuss the baking of wafer in the microlithography process and chemical mechanical polishing (CMP) process affect the depth of focus (DOF) . DOF is defined as the range of focus which keeps the resist profile of a given feature within all specifications (linwidth, sidewall angle, and resist loss) while maintaining at least the specified exposure latitude. Variation in DOF affects the critical dimension uniformity. To improve the DOF, both the thickness variation issue in the baking process and within-wafer-non-uniformity (WIWNU) in the CMP process are discussed into details. To solve the thickness variation, in another

term, wafer warpage, an automatic fault detection methodology is proposed and a physical model of the baking system is presented. Airgap estimation will be done through experiments in the future. As for the WIWNU, a combination of statistics process control (SPC) and advanced process control (APC), namely run-to-run control, is investigated and we contribute our self-tuning PCC Controller with a magnitude of improvement in terms of WIWNU. With the aid of the automatic fault detection in thickness variation and improved WIWNU in CMP process, the DOF requirement is met to the next-generation device manufacturing.

1.1.1 Overview of semiconductor manufacturing

The semiconductor manufacturing industry is arguably the fastest evolving major industry in the world. Semiconductors, sometimes referred to as computer chips or integrated circuits (ICs), contain numerous electrical pathways which connect thousands or even millions of transistors and other electronic components. These transistors store information on the semiconductors, either by holding an electrical charge or by holding little or no charge. An integrated circuit consists of several layers of carefully patterned thin films, each layer is chemically altered to achieve the desired electrical characteristics. Although the design of integrated circuits is normally done by electrical engineers, these device are manufactured through a series of physical and/or chemical batch unit operations similar to the way that specially chemical are made; From 30 to 300 process steps are typically required to construct a set of circuits on s single crystalline substrate called a wafer. The wafers are 100-300mm in diameter, 400-700 μm thick are served as the substrate upon which microelectronic circuits (device) are built. Circuits are constructed by depositing thin films (0.01-10μm) of material of carefully controlled composition in specific patterns and then etching these films to exacting geometries (0.35-10μm).

Success in the industry requires constant attention to the state of art in process

Table 1.1. Silicon integrated circuit technology roadmap

Year	2003	2004	2007	2010	2013	2016
DRAM 1/2 PITCH (nm)	100	90	65	45	32	22
MPU / ASIC 1/2 PITCH (nm)	100	90	65	45	32	22
Maximum reference clock speed (MHz)	2500	2500	2500	2500	5000	5000
Wafer size (mm)	300	300	300	300	450	450
Interconnect levels	8	9	10	10	11	11

tools, process chemistries and physics, and techniques for processing and process improvement. In the US, a technology "roadmap" for design features of integrated circuits has been promulgated by the international technology roadmap of semiconductor (ITRS) (2003); Table 1 gives the ITRS roadmap for the design parameters over a 13-year period. As the wafer diameter increases from 12 to 18 in (300-450mm) and the DRAM 1/2 PITCH shrinks from 90-22 nm, more chips will be placed on each wafer. The manufacturing costs can ne lowered by 25-40%, which will require fewer factories to meet chip demand. The new "fabs" that will be constructed in the 21st century will incorporate increased robotic handling, utilize a high level of in-situ diagnostics instead of post-process testing, and employ realtime process control to achieve much higher levels of accuracy and reduced variation in key quality variables.

The two major fronts along which product advancements are made in this industry are minimum feature size and wafer critical dimension. At the time of this writing, the "state-of-the-art" minimum feature size was in the 130 to 100 nm range, while processing on 300mm wafers was becoming more prevalent. As feature size shrink are wafer size increase, the industry must innovate to maintain acceptable product yield, throughput, and overall equipment effectiveness (OEE). Some manufacturing capability attributes, such as non-product wafer (NPW) usage and wafer scrap, must actually be improved in the transition to larger wafer sizes because of the increased value of 300mm wafers (raw and processed). Faults introduced in any stage of manufacturing will often only show up in final elec-

tronic testing, and the consequent device loss may be (cost-wise) quite devastating especially with the large-diameter wafers.

Although a number of solutions,including improved equipment design and process innovation, will continue to aid in making these transitions cost effective, it has become clear that they are no longer sufficient. Specifically, it has become generally accepted that process and wafer quality sensing and subsequent process tuning will be required to complement these equipment and process improvements. The microelectronics industry has adopted a broad definition of advanced process control (APC) which is considered to include four components, namely fault detection, fault classification, fault prognosis, process control. In its most basic mode, APC monitors the process and determines the necessary manipulated variable action. The controller also monitors the adjustments to ensure they satisfy operating constraints and generates the necessary alarms. Recently, the definition of APC has been expanded to include not just a single machine, but to encompass the entire fab. The ultimate motivation for APC in microelectronics manufacturing is improved device yield. A typical semiconductor manufacturing process can have several hundred unit operations, any of which could be a yield limiter if a given unit operation is out of control. It is difficult to evaluate the potential yield for a given a lot before the wafers reach the end of production line. Therefore, it is essential that each one of steps in the manufacturing process be operated as closely as possible to the specification for the operation. In the APC process, fault detection and diagnosis can potentially be automated, reducing the time spent in these operations. The main form of process tuning that is being implemented as a standard process and equipment control solution is run-to-run (R2R) control, which is now a proven and available technology, and has become a critical component of the success of existing and next generation fabrication facilities.

1.1.2 Process monitoring, fault detection and diagnosis

According to SEMATECH, overall equipments effectiveness (OEE) numbers typically show process tools actively producing product wafer only 30 % of time. Running test wafers and both scheduled and unscheduled downtime represent almost another 30%. Fully optimized OEE can theoretically double the wafers output in a fab. Equipment monitoring and fault detection is at a early state of implementation within the semiconductor industry. Through the equipment analysis, a better understanding is obtained on the affects of aging, cleaning and maintenance cycles on process tool performance. Emerging equipment monitoring and fault detection technologies can provide fabs with all these capabilities by tracking key signals critical to the process step.

Current fault detection is often off-line. The same data can be used to compare tool-to-tool, wafer-to-wafer, or lot-to-lot variations. It can also be used to evaluate the effects of maintenance or clean cycles on yield, develop programs of conditions based vs time-based maintenance, and analyze "what-if" scenarios to improve process performance via equipment remodeling. Benefits from implementing these strategies in production including improving the predictability and quality of process results, reducing scrap and decreasing the number of test wafers. The capability of handling these functions from anywhere in the fab will be necessary as companies head toward high speed network data access.

Fault detection identifies the problematic conditions or faults by examining the fab data. Once a faulty behavior is identified, the diagnosis process attempts to identify its possible cause. This process is usually done by a team of engineers. Anyway, fault detection and diagnosis can potentially be automated, at least partially, reducing the time spent in these operations. Fault detection and diagnosis represent a very high level of data processing. Typical fault detection and diagnosis used first principle model or empirical models, sophisticated statistical analysis and

symbolic processing. Techniques like neural network, principle component analysis and expert systems are beginning to be used to perform automated fault detection and diagnosis.

1.1.3 CMP and thickness variation in microlithography

Lithography is the key technology in semiconductor manufacturing, because it is used repeatedly in a process sequence that depends on the device design. It determines the device dimensions, which affect not only the device's quality but also its product amount and manufacturing cost. It is a kind of art made by impressing, in turn, several flat embossed slabs, each covered with greasy ink of a particular color, onto a piece of paper. The various colors or levels must be accurately aligned with respect to one another within some registration tolerance. Several methods can be used to make ultra large scale integration (ULSI) circuit patterns on wafers such as optical lithography, electron lithography, x-ray lithography and ion-beam lithography. The most common process is to make the the master photomask using an electron beam exposure system and replicating its image by optical printers. The exposing radiation is transmitted through the "clear" part of a mask. The opaque part of the circuit pattern blocks some of the radiation. The resist, which is sensitive to the radiation and has resistance to the etching, is coated on the wafer surface. The mask is aligned within the required tolerance on the wafer; then radiation is applied through the mask, the resist image is developed, and the layer underneath the resist is etched.

In general, dynamic random access memories (DRAMs) have been used as the indicator of progress in ULSI technology. The most advanced DRAM currently in mass production is the 1G bit type with $0.15\mu m$ geometry. If we consider the resolution, optical lithography is considered almost impossible to use for devices that has less than $0.2\mu m$ geometry because of its resolution limit. We have only

two choices: electron beam direct writing or an x-ray technology. However optical lithography still has a margin over next-generation device manufacturing because some commercially available resists can resolve down to $0.2\mu m$, optical lithography still has a margin over next-generation manufacturing. Therefore design of focus (DOF), which is defined as the total amount od defocus allowed without violating a given linewidth tolerance, has become very tight because the wavelength has been reduced and the numerical aperture (NA) has been increased. There are two process namely baking process and CMP process who can affect the DOF. Once we can solve the thickness variation issue present in the baking process and reduce the within-wafer-non-uniformity (WIWNU) of the wafer, we can increase our DOF to meet the manufactuability requirements.

In this thesis, we first discuss a new way, an online fault detection in the process of baking of semiconductor wafer, whose aim is to help us to detect the thickness variation of the wafer. This process is included in photolithography step in semiconductor manufacturing process. A general requirement of these systems is a capability to reject the load disturbance induced by a placement of a cold substrate on the bake plate. When a wafer or reticle is placed on the bake plate, the temperature of the bake plate drops and then is gradually rejected by the heater controller. The temperature disturbance is gainfully used to estimate the airgap between the wafer and the heater surface. Warpage of the wafer can affect device performance, reliability and line-width control during various microlithographic patterning steps. There are a few factors accounting for the wafer warpage or bowing during its processing, one of the reasons is exactly due to stress by CMP processing on the wafer surface causes wafers to bend. Current methods for measuring wafer warpages include capacitive measurement probes, shadow moire techniques, pneumatic-electro-mechanical systems. However, these are off-line methods. Hence they increase process steps, the relevant equipment and operation cost and prolong the production cycle. In chapter 2, we have presented a physical model of the bak-

ing system and an automatic airgap estimation will be done through experiments in the future. In real life application, we can directly get how many wafers with the warpage are out of our specification, therefore they need not further processed or inspected. It is cost-effective and labour-saving. Comparatively, it is easily implemented on line and does not increase system complexity and equipment cost.

Although chemical-mechanical planarization (CMP) has been used for years to produce smooth damage-free silicon wafer surfaces, it has only recently become an essential step in the device fabrication sequence. CMP is being used to provide unprecedented planarity of inter-layer dielectric silicon dioxide and in lithography-limited sub-micron trench isolation (Warnock, 1991). CMP improved on the alternate planarization techniques in many ways. The basic process is to deposit the silicon oxide thicker than the final thickness you want and polish the material back until the step heights are removed. This gives you a good flat surface for the next level. In addition, the process can be repeated for every level of wiring that is added. CMP is the only technique that performs global planarization of the wafer. This is absolutely required to increase the number of wiring levels in the integrated circuits. Prior to CMP, DOF issues due to global planarization problems limited the total number of IC wiring levels to 3 - 4. With CMP, current state of the art IC production is able to achieve 7 - 8 wiring levels.

The control of CMP is chronically poor,arising from poor understanding of the process, degradation (wear out) of polishing pads, inconsistency of the slurry, variation in pad physical properties, and the lack of in-situ sensors (Boning *et al.*, 1996). The nature of the polishing environment means that it is not possible to obtain real-time measurements of the surface planarity although indirect end-point detection methods do exists (Bibby and Holland, 1998), however they are still not reliable enough to be used in actual manufacturing environment. Ex-situ measurement of surface thickness and uniformity are thus required to characterize the process. Due to the lack of in-situ measurements of surface thickness and the non-

stationery nature of the CMP process , Run-to-run controllers have been widely used. We compare the traditional run-to-run control algorithms like EWMA, PCC, OAQC and herein contribute our self-tuning PCC Controller with a magnitude of improvement in terms of WIWNU. With the aid of the automatic fault detection in thickness variation and improved WIWNU in CMP process, the DOF requirement is met to the next-generation device manufacturing.

1.2 Contribution

In this thesis, we first discuss about the automated fault detection in microlithography process and then investigated the advanced process control, namely run-to-run control in CMP process. The aim of this thesis is to employing advanced techniques to improve DOF in microlithography process to meet the future device manufacturing requirements.

In the early part of this thesis, we put forward one method by combining the dynamics of bake plate and wafer, we thus present a physical model of the baking system and airgap estimation will be done through experiments in the future.

As for the process control in CMP process, we focus on the APC control methodology. We compared the traditional EWMA controller, PCC Controller and OAQC controller and then proposed a self-tuning PCC controller. The performance of the EWMA controller and PCC controller depends heavily on the proper selection of the weighting factors or forecasting parameters. That makes it inconvenient to be used in a manufacturing environment where the nature of the disturbance is usually unknown. The issue of the difficulty in choosing EWMA weights for a MIMO process is raised by Smith and Boning (Smith and Boning, 1996). The effects of drift, noise, target change, and model error on the optimal EWMA weight are investigated. The authors distill these various process disturbance and model error down to a single disturbance state of noise and drift.

This state can be mapped to optimal EWMA weights using an artificial neural network (ANN) . During controller operation, the disturbance state is estimated and passed to the neural network that determines the optimal EWMA weights to be used by the observer. Some probabilistic tuning methods have also been proposed for the EWMA controller but they are yet to be developed for the PCC controller (Hamby *et al.*, 1998). For PCC controller, though an extra degree of freedom is obtained using two EWMA filters, tuning of the second EWMA filter is not as intuitive as it is for a single EWMA controller. Therefore, the ability to dynamically update the two weighting factors is necessary to achieve the best performances of the controller. If the variability of the adjustments can be neglected, the following approach for tuning PCC controllers was suggested by Del Castillo (Castillo, 2002). There is a trade-off between the magnitude of the transient effect and the long-run(asymptotic) variance when choosing the weights. In the long run, the PCC eliminates the offset and the process will be on the target on average. Therefore, the author proposes an optimization based approach for choosing the appropriate weights for PCC controller. The difficulty and uncertainty of tuning the forecasting parameters in the PCC controller is the motivation of this thesis. In this thesis, a methodology for self tuning the two forecasting parameters by using variable step size least mean square estimation in PCC controller is developed and discussed in full detail. Simulation results depicts an order of magnitude improvement in terms of the removal rate and non-uniformity when compared to conventional R2R controllers.

1.3 Organization

This thesis is organized as follows. Chapter 2 first proposes a prediction of wafer warpage in photolithography semiconductor process, we present a physical model of the baking system as the first step. Chapter3 further presents a brief view of

chemical mechanical polishing process. The blanket wafer performance metrics is also introduced in this chapter and the first investigation of the modelling of CMP process is discussed as well. Chapter4 further investigates the process modelling for CMP process by using design of experiments and response surface model and then discussed model-based run-to-run control algorithm, including EWMA, PCC, OAQC and MPC, A performance benchmark is then used to compare the control effect among these model based controllers. In chapter5, we propose a self-tuning PCC controllers to automatically select the optimal weighting factors in PCC controller, and via the simulation of the benchmark problem, we prove that our self-tuning PCC controllers can achieve an order of magnitude improvement in terms of the removal rate and non-uniformity when compared to conventional R2R controllers. Conclusions are drawn in Chapter 6.

Chapter 2

Modeling of wafer warpages during baking in microlithography

2.1 Introduction

Photolithography may be considered as the most critical step in the semiconductor manufacturing process. It is estimated that lithography accounts for nearly one third of the total wafer fabrication cost (Quirk and Serda, 2001). As shown in Figure 2.1 the micro-lithography sequence includes numerous baking steps such as the soft bake, post-exposure bake, and post-develop bake (hard bake). In some cases, additional bake steps are employed. Thermal processing of semiconductor wafers is commonly performed by placement of the substrate on a heated plate for a given period of time. The heated plate (or chuck) is held at a constant temperature by a feedback controller that adjusts the (resistive or radiant, in the case of susceptors) heater power in response to a temperature sensor embedded in the plate near the surface. The wafer is placed on proximity pins. Processes that utilize this thermal approach span a large temperature range and include photoresist processing, chemical vapour deposition (CVD) and rapid thermal annealing (Campbell, 1996), (Schaper *et al.*, 1994).

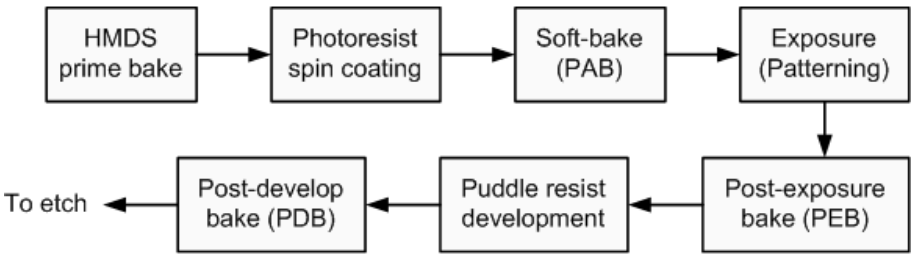

Figure 2.1. The photo-resist processing and exposure steps are used in the lithography sequence

A general requirement for these systems is an ability to reject the load disturbance induced by placement of a cold substrate on the bakeplate (Ho *et al.*, 2000). When a wafer or reticle is placed on the bakeplate, the temperature of the bake plate drops and then is gradually rejected by the heater controller. In this chapter, we show how the temperature disturbance can be gainfully used to estimate the airgap between the wafer and the heater surface. A warped wafer and a flat wafer would give different airgap. Wafer warpages is a common feature in semiconductor manufacturing. For example, silicon nitride deposition can induce strong tensile stress on the silicon wafer, this stress also contributes to the wafer warpage which has an effect on the breakdown voltage of DRAM device, warpage of about 80 μm is reported in Song et al. (I.S. *et al.*, 1999). In silicon thinning and stress relief, wafer warpages up to 260 μm is observed (Hendrix *et al.*, 2000). Gettering and dislocation density in silicon wafers also affect wafer warpages, Kishino et al. (Kishino *et al.*, 1993) reported warpage range from $40 - 120$ μm for a change of 1000 cm^{-3} in dislocation density. Fukui and Kurita (Fukui *et al.*, 1997) reports InP wafer warpage of 60 μm induced during its processing. The processing of semiconductor wafers in the manufacturing of integrated circuits requires that the wafer to be extremely flat (Sheats and Smith, 1998), (Thompson *et al.*, 1994). This is especially critical during the lithography process (Exposure step) where knowledge of wafer flatness is extremely important. Warpage is also a critical issue in thin wafer processing in the smart card industry (Hendrix *et al.*, 2000).

Table 2.1. Product critical level post-exposure bake requirements

Year of First Product Shipment	1999	2001	2004	2007	2010	2013
Technology Node (nm)	180	130	90	65	45	33
Post-Exposure Bake (PEB) sensitivity (nm/^{0}C)	5	4	2	2	1	1

The information of wafer warpage is critical for precise temperature control, equipment design, process optimization and routine monitoring. Temperature uniformity control is an important issue in photoresist processing with stringent specification as shown in Table 2.1. Of these, the most important or temperature sensitive step is the post-exposure bake step (PEB). Temperature metrology during PEB is vital for reducing critical dimension (CD) variability and effective profile control in deep ultra-violet (DUV) lithography. The precise temperature control can promote chemical modifications of the exposed portions of the photoresists (Sheats and Smith, 1998). Excessive temperature variations will affect the kinetics of the acid catalytical reaction in the resist. For such chemically-amplified photoresists, the temperature of the substrate during this thermal step has to be controlled to a high degree of precision for CD control. For commercially available DUV resist systems, a representative post exposure bake latitude for CD variation is about 5nm /°C. Requirements call for temperature to be controlled within 0.1°C at temperatures between 70°C and 150°C for 1 or 2 minutes to reduce CD inconsistencies.

Sturtevant et al. (Sturtevant *et al.*, 1993) reported a 9% variation in CD per 1°C variation in temperature for a DUV photoresist. APEX-E resist has been shown to display a sensitivity close to 12 nm/°C, and UVIIHS 4 to 10 nm/°C. A number of recent investigation also shows the importance of proper bake plate operation on CD control (Crisalle *et al.*, 1998) (Mohondro and Gaboury, 1993). According to the International Technology Roadmap for Semiconductors (2000), the PEB

resist sensitivity to temperature will be more stringent for each new lithography generation as depicted in Table 2.1. By the year 2010, the PEB resist sensitivity is expected to be only 1 nm/°C; making temperature control even more critical. To meet future temperature requirements for advanced lithography processes, it is important to reduce temperature variation of the baking process.

There are several origins of wafer warpage which can be induced during its processing (Fukui *et al.*, 1997): 1. stress by mechanical processing (e.g. CMP or backgrinding) on the wafer surfaces causes wafers to bend, sometimes it will cause wafer to be fractured; 2. stress by heating processing (e.g. rapid thermal processing) also causes wafers to bend, the mismatch in the coefficient of thermal expansion among different layers in silicon substrates induces some distortion at the wafer level; 3. stress are also induced during slicing and lapping of wafers. Current methods for measuring wafer warpages include capacitive measurement probes, shadow Moire techniques (Wei *et al.*, 1998), pneumatic-electro-mechanical systems. An innovative alternative for full-field, whole-wafer measurement is developed using a laser light source and the modified shadow moire technique (Wei *et al.*, 1998). The shadow moire method does not require wafers to be contacted or rotated, thus reducing the vibration and enhancing the fidelity of measurement. In addition, the whole wafer surface can be obtained through fringe patterns which can then be analyzed by computers to automate the measuring process.

However, these are off-line methods. Hence they increase process steps, the relevant equipment and operation cost and prolong the production cycle. In this paper, we put forward one method by combining the dynamics of bakeplate and wafer to modelling heat dynamics based on temperature measurement of bakeplate and wafer. We will conduct the parameter estimation and validate our simulation through the experiments in the future.

This chapter is organized as follows. In Section 2.2, we present a physical model of the baking system. Conclusions are presented to propose the future work and

assess the future application of advanced systems techniques to the lithography process in Section 2.4.

2.2 Modelling of baking process in microlithography

Conventional thermal systems utilize separate bake plates and chill plates to accomplish the baking steps. These units are comprised of large thermal mass systems that are held constant at the set-point temperature. The substrate is placed on the bake or chill plate. The substrate typically rests about 5 *mils* (thousandths of an inch) from the surface of the plate on small pins, as opposed to direct contact, to prevent contamination. The plates are typically single or dual zones systems. To further understand the limitations of these conventional bake systems, a mathematical model is developed for the bake operation.

2.2.1 Thermal modeling

We assume that the substrate used for baking is a silicon wafer. Spatial distributions of temperature and other quantities in a silicon wafer is most naturally expressed in a cylindrical coordinate system. The assumed system consists of three main parts, the bakeplate, airgap and wafer. The bakeplate is also assumed to be cylindrical in shape with the same diameter as the wafer. The system is discretized spatially into N radial elements as shown in Figure 2.2.

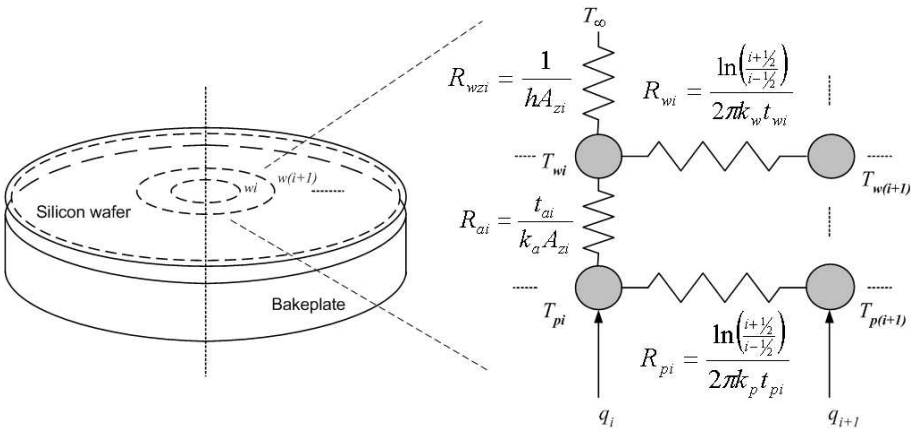

Figure 2.2. Thermal model of wafer-bakeplate system

The energy balance equations for the system are as follows:

$$
\begin{aligned}
C_{p1}\dot{\theta}_{p1} &= -\frac{\theta_{p1}-\theta_{p2}}{R_{p1}} - \frac{\theta_{p1}-\theta_{w1}}{R_{a1}} + q_1, \\
C_{pi}\dot{\theta}_{pi} &= \frac{\theta_{p(i-1)}-\theta_{pi}}{R_{p(i-1)}} - \frac{\theta_{pi}-\theta_{p(i+1)}}{R_{pi}} - \frac{\theta_{pi}-\theta_{wi}}{R_{ai}} + q_i, \qquad 2 \le i \le N-1, \\
C_{pN}\dot{\theta}_{pN} &= \frac{\theta_{p(N-1)}-\theta_{pN}}{R_{p(N-1)}} - \frac{\theta_{pN}}{R_{pN}} - \frac{\theta_{pN}-\theta_{wN}}{R_{aN}} + q_N, \\
C_{w1}\dot{\theta}_{w1} &= \frac{\theta_{p1}-\theta_{w1}}{R_{a1}} - \frac{\theta_{w1}-\theta_{w2}}{R_{w1}} - \frac{\theta_{w1}}{R_{wz1}}, \\
C_{wi}\dot{\theta}_{wi} &= \frac{\theta_{w(i-1)}-\theta_{wi}}{R_{w(i-1)}} + \frac{\theta_{pi}-\theta_{wi}}{R_{ai}} - \frac{\theta_{wi}-\theta_{w(i+1)}}{R_{wi}} - \frac{\theta_{wi}}{R_{wzi}} \qquad 2 \le i \le N-1, \\
C_{wN}\dot{\theta}_{wN} &= \frac{\theta_{w(N-1)}-\theta_{wN}}{R_{w(N-1)}} + \frac{\theta_{pN}-\theta_{wN}}{R_{aN}} - \frac{\theta_{wN}}{R_{wN}} - \frac{\theta_{wN}}{R_{wzN}}.
\end{aligned}
$$

where

$\theta_{pi} = T_{pi} - T_\infty$	:	$i\ th$ plate element temperature
$\theta_{wi} = T_{wi} - T_\infty$	:	$i\ th$ wafer element temperature
C_{pi}	:	thermal capacitance of $i\ th$ plate element
C_{wi}	:	thermal capacitance of $i\ th$ wafer element
R_{pi}	:	thermal conduction resistance between the $i\ th$ and $i+1\ th$ plate element
R_{wi}	:	thermal conduction resistance between the $i\ th$ and $i+1\ th$ wafer element
R_{wzi}	:	thermal convection loss of the ith wafer element
R_{ai}	:	thermal conduction resistance between the $i\ th$ plate and $i\ th$ wafer element
q_i	:	heat flux into the $i\ th$ plate element

The various thermal resistances and capacitances are given by

$$
\begin{aligned}
R_{pi} &= \frac{\ln\left(\frac{i+1/2}{i-1/2}\right)}{2\pi k_p t_p} \quad (K/W) && 1 \le i \le N-1 \\
R_{pN} &= \frac{1}{h(\pi D t_p)} \quad (K/W) \\
R_{wi} &= \frac{\ln\left(\frac{i+1/2}{i-1/2}\right)}{2\pi k_w t_w} \quad (K/W) && 1 \le i \le N-1 \\
R_{wN} &= \frac{1}{h(\pi D t_w)} \quad (K/W) \\
R_{wzi} &= \frac{1}{h A_{zi}} \quad (K/W) \\
R_{ai} &= \frac{t_a}{k_a A_{zi}} \quad (K/W) \\
C_{pi} &= \rho_p c_p \left(t_p A_{zi}\right) \quad (J/K) && 1 \le i \le N \\
C_{wi} &= \rho_w c_w \left(t_w A_{zi}\right) \quad (J/K) && 1 \le i \le N \\
A_{zi} &= \pi \Delta r^2 \left[i^2 - (i-1)^2\right] \quad (m^2) && 1 \le i \le N
\end{aligned}
$$

where A_{zi} is the cross-sectional area of element i normal to the axial heat flow. t_p, t_w and t_a are the bakeplate thickness, wafer thickness and airgap between the wafer and the bakeplate. ρ_p and ρ_w are the density of the bakeplate and wafer respectively. c_p and c_w are the specific heat capacity of the bakeplate and wafer respectively. The width of each element is given by $\Delta r = D/2/N$. Some numerical

values are helpful to analyze the dynamics. We consider the commercial aluminum hot plate for photoresist processing shown in Figure 2.3.

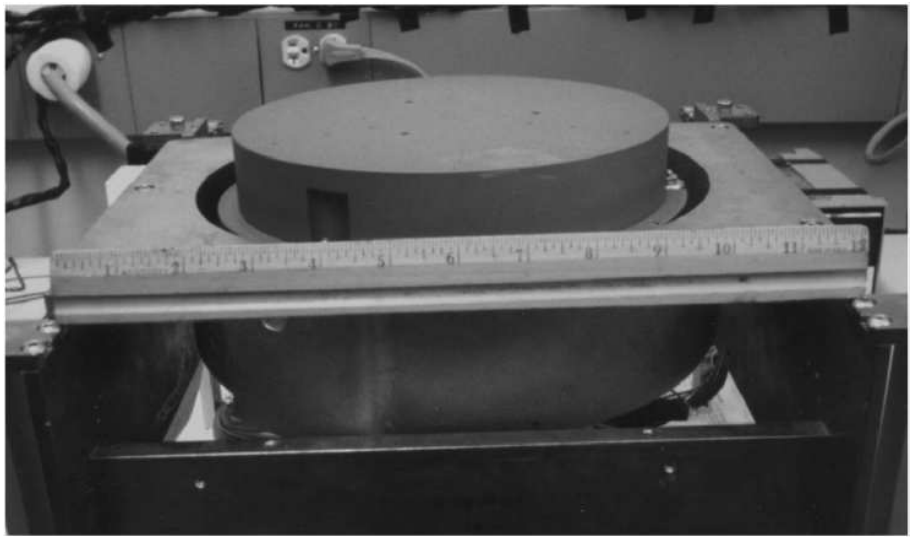

Figure 2.3. Commercial bake-plate.

The system dimensions used in the modeling are as follows. The diameter of the wafer/airgap/heater-plate, $D = 0.3\ m$ (300 mm wafer); the wafer thickness, $t_w = 8.6 \times 10^{-4}\ m$; and the bakeplate thickness, $t_p = 1.78 \times 10^{-2}\ m$. Most thermophysical properties are temperature dependent. However, for the temperature range of interest from 15^0C to 150^0C, it is reasonable to assume that these thermophysical properties remained fairly constant. Average values are used. The thermal properties of pure aluminum, silicon and air are obtained from (Ozisik, 1985) and (Raznjevic, 1976).

2.2.2 Simulation

The performance of conventional bake systems can be analyzed by simulating the above energy balance equations. Figures 2.4 and 2.5 shows the wafer and bakeplate temperature profile when a flat 300 mm wafer is dropped onto a fixed, uniform temperature bakeplate 100^0C above ambient temperature. As expected, the temperature at the edge is lower than the center. The maximum drop in bakeplate

temperature is also fairly uniform except at the edge which has a larger drop as shown in Figure 2.5 (b).

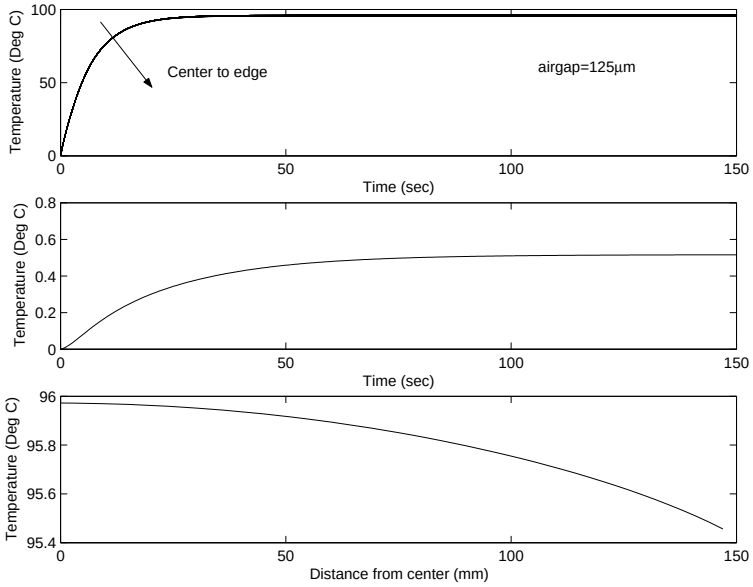

Figure 2.4. Flat wafer dropped on a bakeplate: (a) Wafer temperature, (b) Peak-to-peak wafer temperature nonuniformity, (c) Steady-state wafer temperature.

Figures 2.6 and 2.7 shows the wafer and bakeplate temperature profile when a warped 300 mm wafer is dropped onto a fixed, uniform temperature bakeplate 100^0C above ambient temperature. The wafer is warped in a bow-shaped (airgap between the wafer and bakeplate varies from $5-7$ $mils$ from center to edge). Notice that the wafer temperature nonuniformity is now more severe compared to that of a flat wafer. Such temperature nonuniformity is undesirable for temperature sensitive photoresist processing, we do need a nonuniform bakeplate temperature to give a uniform wafer temperature. Of interest here is that there is significant difference in the maximum drop in bakeplate temperature across the bakeplate as shown in Figure 2.7 (b). By monitoring the bakeplate temperature profile, we are able to estimate the wafer warpage. This can be achieved easily in practice by embedding temperature sensors in the bakeplate.

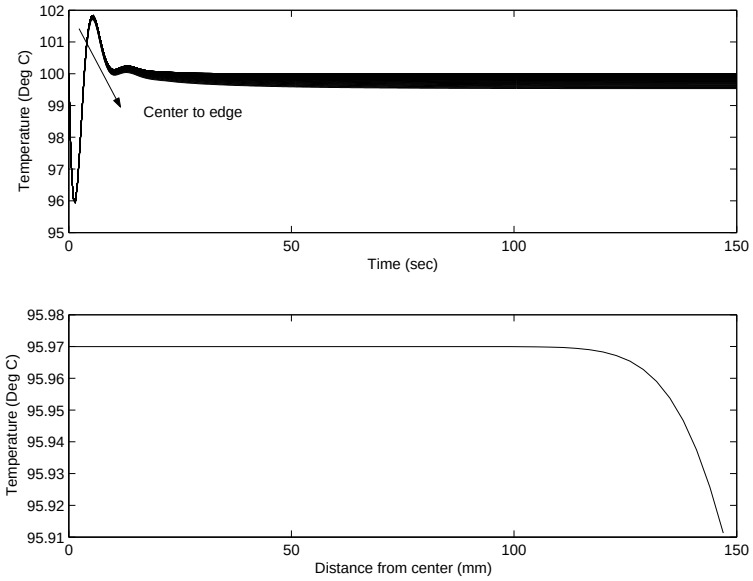

Figure 2.5. Flat wafer dropped on a bakeplate: (a) Bakeplate temperature profile, (b) Maximum temperature drop across bakeplate.

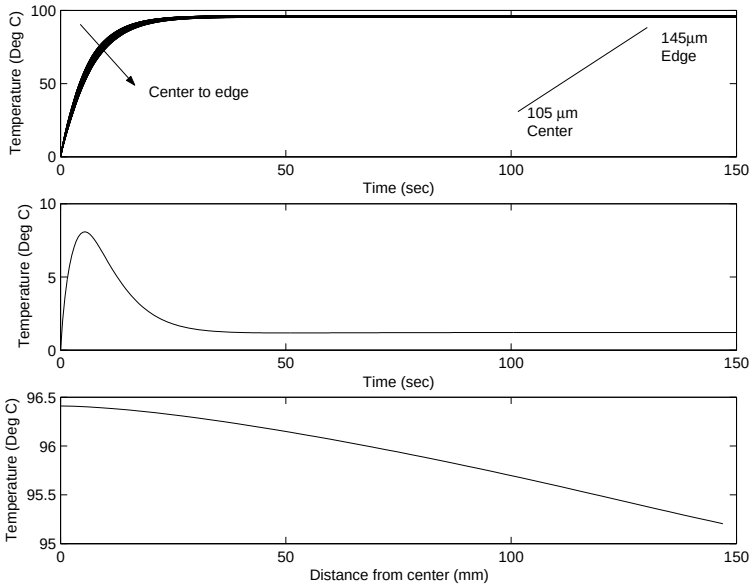

Figure 2.6. Warped wafer(center to edge airgap: $105\mu m \rightarrow 145\mu m$) dropped on a bakeplate: (a) Wafer temperature, (b) Peak-to-Peak wafer temperature nonuniformity, (c) Steady-state wafer temperature.

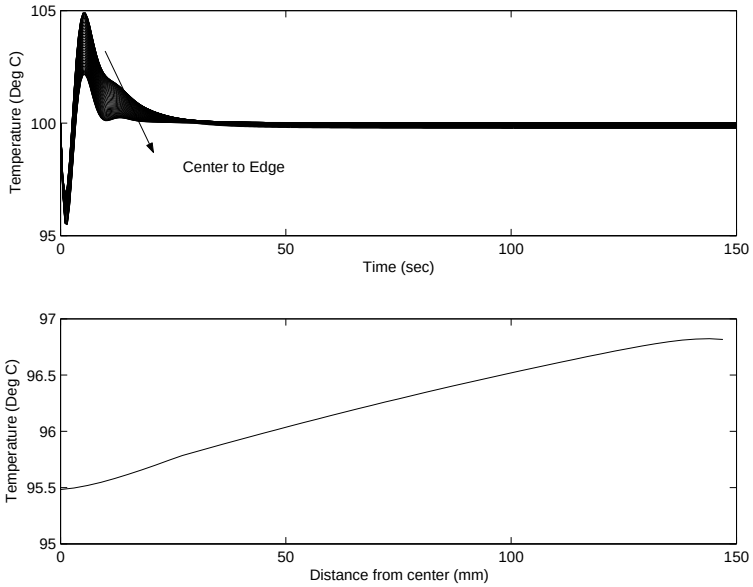

Figure 2.7. Warped wafer (center to edge airgap: $105\mu m \rightarrow 145\mu m$) dropped on a bakeplate: (a) Bakeplate temperature profile, (b) Maximum temperature drop across bakeplate.

2.3 Predicting wafer warpage

In this section, we show how the temperature disturbance can be used to estimated the airgap between the wafer and the heater surface. As we have seen in Figure 2.5, the plate temperature drops to a minimum before the PID controller rejects the disturbance and returns to the steady state. If we investigate the temperature drop at the center of hotplate, and regard it as the average of the temperature drop across the whole hotplate, we can actually predict the wafer warpage by inspecting the maximum temperature drop.

From the model we have set up, we change the number of the airgap parameters in the modeling and then get the different temperature profiles. We then plot the maximum temperature drop versus different airgaps in Figure 2.8,

From Figure 2.8, we can see that with the rise of airgap, the maximum temperature drop of the hotplate is decreased because the heat convection and conduction between the hotplate and wafer are decreased due to the larger airgap. Actually

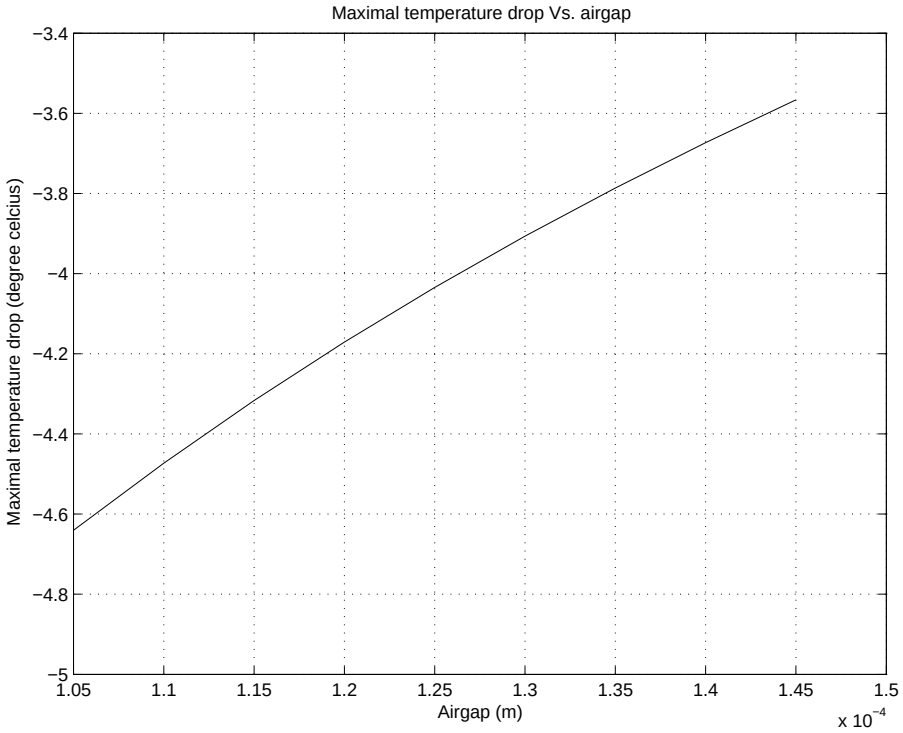

Figure 2.8. Maximum temperature drop of hotplate versus airgap

we can use this property to predict the wafer warpage just by simply checking out the maximum temperature drop of the hotplate in the baking wafer process. Thus the fault detection is easily implemented on-line, automated and therefore cost-effective and labour saving. For example, in Figure 2.9, a schematic of the system under consideration is shown. The system consists of 3 basic sections: the heater, the airgap and the silicon wafer. The silicon wafer is placed on the pin of the hotplate instead of being attached to the hotplate directly. Assume that we have measured that the height of the pin is 3 μm, and then we put a cold wafer to the hotplate, the temperature of the bake plate drops and then is gradually rejected by the heater controller. Actually we can inspect the maximum temperature drop and then put it into Figure 2.8 to check the corresponding airgap, if this airgap is equivalent to the height of the pin, in other words, $airgap = 3\mu m$, then we can judge that this wafer is unwarped. Instead, if the airgap we have inspected

from corresponding maximum temperature drop is less than the height of the pin, for example, only $1\mu m$, then we can say that the wafer is warped in the baking process, as shown in Figure 2.10.

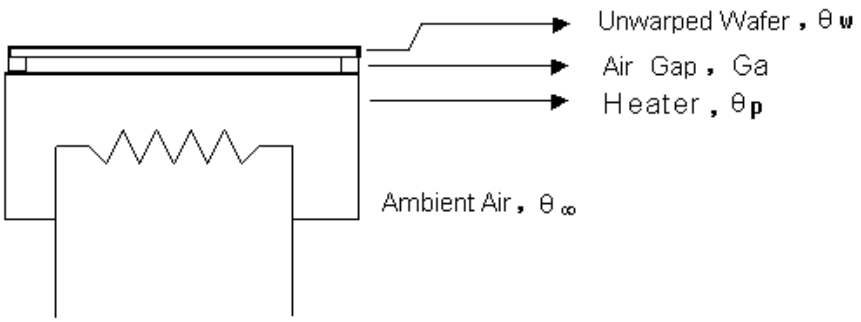

Figure 2.9. Unwarped wafer profile

Airgap estimation based only on the temperature measurement from bake plate and wafer is easily implemented and we only need the maximum wafer temperature drop for the first time and it will reduce contamination induced by putting temperature sensors on wafer in the subsequent processes. Therefore, it is cost effective for semiconductor manufacturer to get the wafer warpage arising from heating process. It will be greatly helpful to improve the CD control due to the temperature variation throughout heating process. The basic underlying principle is that we combine the maximum bake plate temperature drop and airgap together in our modeling part and build up the corresponding relation between them. In our prediction, what we have to emphasize is that the airgap we obtained is the average warpage on wafer surface (only one point in the center). Warpage profile on whole surface can be mapped out if we have multiple sensors embedded in the bakeplate.

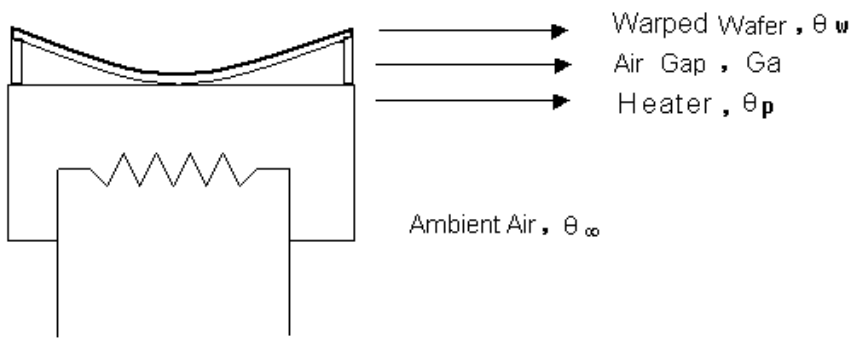

Figure 2.10. Warped wafer profile (deflexed)

2.4 Conclusion

The lithography manufacturing process will continue to be a critical area in semiconductor manufacturing that limits the performance of microelectronics. Enabling advancements by computational, control and signal processing methods are effective in reducing the enormous costs and complexities associated with the lithography sequence. In this thesis, we have presented a physical model of the baking system and airgap estimation will be done through experiments in the future.

Chapter 3

The chemical and mechanical polishing process

Although chemical-mechanical polishing (CMP) has been used for years to produce smooth damage-free silicon wafer surface, it has only recently become an essential step in the device fabrication sequence. CMP has been used in the global planarization of oxide and tungsten process and now is being used to provide unprecedented planarity of inter-layer dielectric silicon dioxide, copper and in lithography limited sub-micron trench isolation (Warnock, 1991). It is projected that the observed effectiveness of the CMP process will lead to the widespread use of this process at various stages of integrated circuit (IC) fabrication, for a variety of high performance and application specific ICs, and for a variety of materials.

3.1 Introduction

The CMP process involves a silicon wafer, attached to a carrier by vacuum, being pressed face down into a polishing pad. The polishing environment is flooded with a colloidal slurry which physically enhances abrasion and helps prevent redeposition of the oxide or metals. The polish table is rotated while the wafers also

rotate about their axis and orbit about the polish table ((Boning *et al.*, 1996)). Figure 3.1 shows a schematic of a simple CMP machine.

Figure 3.1. Baseline CMP experiment

Abrasive particles in the slurry cause mechanical damage on the sample surface, loosening the material for enhanced chemical attack or fracturing off the pieces of surface into a slurry where they dissolve or are swept away. The process is tailored to provide enhanced material remove rate from high points on surfaces (compared to low areas), thus affecting the planarization (Steigerwald *et al.*, 1997). Note chemistry alone will not achieve planarization because most chemical actions are isotropic. Mechanical grinding alone, theoretically, may achieve the desired planarization but is not desirable because of extensive associated damage of the material surfaces. Here it is pointed out that there are three main players in this process (Steigerwald *et al.*, 1997):

- The surface to be polished;
- The pad-the key media enabling the transfer of mechanical forces to the surface being polished; and
- The slurry- that provides both chemical and mechanical effects.

Most of the input and output variables to CMP could be categorized in one of the above groups. Temperature, pressure, relative velocity of the surface being polished with respect to pad (which is usually rotating), and pre- and post-CMP cleaning that may affect the final acceptance criteria for the polished surface are other parameters that play important roles.

What is so unique about CMP? CMP achieves planarization of the nonplanarized surfaces. Nonplanarized surface topography is a result of the fabrication process that ends up with a deposition of the film on a previously patterned surface, with a pattern generated by an etching. The generation of surface topography by several deposition, pattern etch and planarization processes have been examined by Pai et al. Only CMP is universally applicable to cause global planarization. There are several advantages of CMP as follows (Steigerwald *et al.*, 1997):

- Achieves global planarization.
- Universal or materials insensitive-all types of surfaces can be planarized.
- Useful even for multi-material surfaces.
- Reduces severe topography allowing for fabrication with tighter design rules and additional interconnection levels.
- Provides an alternate means of patterning metal eliminating the need of reactive ion etching or plasma etching for difficult-to-etch metals and alloys.
- Leads to improved metal step coverage.
- Helps in increasing reliability, speed and yield of sub-0.5 μm devices/circuits.
- Expected to be a low cost process.
- Does not use hazardous gases in dry etching process.

The most important advantage is that CMP process achieves global planarization which is essential in building multilevel interconnections. However, there are also several cost advantages to using CMP. The increase in processing complexity required by many planarization schemes increases both cost and defect densities. Alternatively, CMP planarization involves only one step and often reduces or eliminates some defects. Nonplanarity defects such as metal stringers, which form when the thick metal film at the edge of a step is not completely etched, and poor step coverage are eliminated by global planarization. Because CMP levels the wafer surface, film particles from previous processing can be readily removed. Indeed, companies often find a reduction in defect densities upon implementing CMP processes, in spite of post-CMP cleaning treatments being a relatively undeveloped area. Reduced defect densities translate to increased die yields and decreased die cost. However, there are also some disadvantages of CMP, which arise from the fact that CMP is a new process which remains unoptimized. As a result of the process immaturity, process windows are narrow, requiring an increased level of wafer metrology to obtain the desired results. Anyway, the promise of global planarity leading to improved performance is likely to make the required investment extremely cost effective.

3.2 The CMP variables and manipulations

The CMP process is quite complicated, involves large number of variables. This section lists and discusses these variables in two categories: output variables and input variables. It must be noted that many input variables will primarily affect either the chemical or mechanical component. However, there is a danger in assigning a variable as being either strictly a chemical or mechanical variable. The chemical and mechanical components are inseparable, so that variables cannot be listed as affecting only the mechanical component or only the chemical component.

For example, velocity and pressure can be thought of as primarily mechanical variables. However, changing the velocity and/or pressure will affect slurry transport across the wafer and also the thickness of the fluid layer between the pad and wafer. Slurry transport and fluid layer thickness will then affect the diffusion of chemical reactants and products to and from the wafer surface, which in turn affects reaction rates. Pressure may affect the abrasive size and shape, pad performance, the film stack and then effect old preexisting wafer curvature. This section examines these variables and how the variable is expected to influence the CMP process and the final result.

3.2.1 Output variables in CMP

Polish Rate: Units of (nm/min) or (μm/min). Polish rate is the film thickness removed divided by the polish time. Higher polish rates lead to shorter process times and are thus desirable. However, if the polish rate is too high, the process is difficult to control. Note that the polish rate can be significantly higher for wafers with topography than for un-patterned wafers. This is because the contact area with pad is smaller for wafers with topography.

Planarization Rate: Planarization rate is the time it takes to reduce the topography of the wafer surface to the desired level. In the CMP of oxide and other ILDs, because the end goal is surface planarization not simply material removal, the planarization rate is as important a metric as polish rate.

Surface Quality: Surface quality is an indication of the expected yield and reliability of the interconnections. A rough ILD film is more susceptible to low breakdown strength and high leakage. A rough metal film is more susceptible to corrosion and electromigation. Roughness is minimized by properly balancing the chemical and mechanical components of the CMP process. High , particle densities lower die yields. Particle densities may be reduced by using an effective post-polish

clean sequence and by choice of slurry constituents. A high degree of corrosion resistance of metal films is required to ensure reliability. High corrosion resistance is ensured by forming a passivating film on the metal during or immediately after the CMP step.

3.2.2 Input variables in CMP

Slurry chemicals: A large variety of materials (metals, alloys, insulators, semiconduct -ors, etc.) are being polished. Each has a different chemistry as far as chemical interactions with the slurry is concerned. Slurry chemicals affect primarily the chemical component, e.g., etch rate. However, chemical reactions modify the mechanical properties of the film, pad, and abrasive surfaces, which in turn affect the mechanical component.

Slurry Abrasive: The slurry abrasive provides the mechanical action of CMP. Size and concentration slurry have a different effect on mechanical abrasion. However, the abrasive can also have a chemical effect as in the case of glass polishing with ceria abrasive where the ceria forms a chemical bond with the glass surface or in the case of alumina, which seems to create surface defects on SiO_2 films polished in pH, in the range of 5 to 8.

Slurry flow rate: Units of (liters/min) or (ml/min). The rate at which slurry is delivered to the center of the pad. Slurry flow rate affects how quickly new chemicals and abrasives are delivered to the pad and reaction by-products and used abrasive are removed from the pad. The slurry flow rate also affects how much slurry in on the pad and therefore will affect the lubrication properties of the system.

Temperature: Units of (^oC). Because CMP is in part a wear process, temperature increases are to be expected. Temperature can also be controlled to some extent by maintaining the temperature of the polish table with recirculating water

or by heating the slurry and measuring the temperature at the pad. The primary effect of temperature is on reaction rates. However, dramatic changes in the temperature of the surface will affect the mechanical properties of the film.

Pressure: Units of (kPa) or (psi). Pressure is the load applied to the wafer divided by the wafer area. Note that if the surface is rough or has topography, the contact area is less than the geometric area, and hence the pressure is increased until such time as the surface is made smooth. Mechanical abrasion rate is proportional to pressure. Pressure also affects planarization.

Pad Velocity: Units of rotations per minute (rpm) or (cm/sec). If the wafer is rotated off the axis of the pad , which is common, the pad velocity is also the average relative velocity of the pad with respect to the wafer. Mechanical abrasion rate is also proportional to velocity. Velocity also affects slurry transport across the wafer and the transport of the reactants and products of chemical reactions to and from the wafer surface.

Wafer Velocity: Units of rotations per minute(rpm) or (cm/sec). The velocity of the wafer affects the average velocity across the wafer. If the pad and wafer rotational velocities match, the average velocity is the same at every point on the wafer.

Pattern Geometries: Feature size and pattern density affect localized pressure distribution and therefore affect the removal rate at the feature scale. Small features polish quicker than large features, and small pattern densities polish faster than large pattern densities. Feature size and pattern density thus affect planarization rates in ILD polishing and metal dishing and ILD erosion in metal polishing.

Polish Pad: The polish pad affects virtually all of the above listed output variables and interacts with most of the input variables.

Pad Conditioning: Pad conditioning techniques improve and stabilize performance.

Wafer Curvature: Wafer curvature affects the distribution the applied load

across the wafer. If a wafer is bowed up in the center, for example, more of the applied load is distributed to the center of the wafer, and therefore the pressure is greater in the center. This also contributes to feature size dependence that is variable across the wafer.

Wafer Size: The diameter of the wafer being polished will play a very significant role not only in determining the force, relatively velocity on different areas of the wafer, but also the feed rate of the slurry and integrity of the abrasive under the wafer. A CMP process for large size wafers will thus face the significant problem of the uniform supply of slurry under the wafer.

3.3 Blanket wafer performance metrics

The performance of the CMP process is gauged by several different metrics. In particular, the removal rate (RR) of material on blanket sheet film wafers is often used to judge how quickly a process will remove step heights on patterned wafers. Processes with higher removal rates are generally considered better. The RR is determined by measurements of the oxide film thickness before and after polishing at each several sites on the wafer.

The "removal rate" metric most often used is the average of the amount removed at each site, divided by the fixed polish time. Differences between polish rates at the center and the edge of the wafer may arise due to wafer asymmetry, non-constant relative pad velocity from the edge to the center, non-uniform slurry and by-product transport under the wafer, wafer bowing due to pressure or tool design, or machine drift with tool or pad age of any of these parameters. As a result, the uniformity of the polishing process across the surface of the wafer is also of a concern. In order for all devices on the wafer to be polished to the same amount, the Within-wafer non-uniformity (WIWNU) of a polished unpatterned blanket wafer is desired to be small (typically 5% or less). The calculation of the WIWNU metric

varies in the industry (Smith and Boning, 1999). One common calculation used is the standard deviation of the amount removed (AR) over the sites on the wafer, divided by the average AR over the several sites, times 100. Other approaches include the standard deviation of the removal rate or post-polish thickness profiles. These two blanket wafer metric are generally used to develop CMP processes, as well as to monitor the CMP process on a lot to lot basis. In addition, particle and scratching tests are performed on unpatterned wafers. Particles and wafer scratching caused by CMP can create severe failures in manufactured circuits (Kim *et al.*, 1999), and thus must be carefully monitored.

3.4 An introduction to CMP process problems

The key knowledge of the chemical, structural, and mechanical properties of the surface to be polished establishes the polishing parameter space including the chemistry and mechanical force. CMP of a single material is thus easier compared to that of a surface consisting of different materials spaced at different surface coverage. A complex set of phenomena occurs that control this feature size dependence, the most important of which is related to the elastic behavior of the pad. Ideally one would expect he pad to be rigid and chemically inert so that it can carry abrasives and chemicals all across the surface being polished. For real situations pads are not rigid, leading to several issues: changes occurring in pad properties as polishing continues, cyclic changes, solvent/chemical effects on rigidity, and erosion. Similarly pads are not chemically and physically inert materials, thus leading to the following changes in surface and possibly bulk chemistry of the pad ingredients (changes affected by mechanical forces and changes that affect mechanical properties) , surface bonding between abrasive and pad, electrochemical effects, and the necessity to recondition or regenerate the pad to cause reproducible polishing behavior. Thus there is a need for understanding these changes in pads as a function

of the use and during actual use.

The slurry is the third important key player among the three above. Slurries provide both the chemical action through the solution chemistry and the mechanical action through the abrasives. High polishing rates, planarity, selectivity, uniformity, post-CMP ease of cleaning including environmental health and safety issues, shelf-life,and dispersion ability are the factors considered to optimize the slurry performance. Finally the last important step of the complete CMP-process sequence is the cleaning. Removal of the slurry from the surface without leaving any macroscopic, microscopic, or electrically active defects is very important in making the process useful.

As for the recipe parameters, typically, there are three principal parameters in a CMP recipe including the down force, the platen rotation speed, and the carrier rotation speed. Another variable in CMP recipe is the back pressure. Usually, if the non-uniformity problem is identified to be a center-slow-edge-fast process, back pressure can be used to push the back of a wafer and accelerate the center polish rate. Thus the uniformity can be improved accordingly.

A pad conditioner or pad dresser, is used to condition the polish pad to retrieve polish rate. If this is not done, the surface of a pad can become glazed and the pad austerity lost. A pad conditioner consists of diamond grit or similar silicon carbide materials. These, extremely hard materials can scrape off the topmost layer of a pad during conditioning; if properly deployed, A pad conditioner can help flatten a polish pad and improve polish uniformity. Failure will lead to the surface being roughened and the non-uniformity worsened. No matter how good the consumable, recipes and equipment are in CMP, there is an intrinsic non-uniformity problem, in which a wafer is always polished more at its edge and less at its center. This is due to the fact that the relative velocity between a rotating wafer and a rotating platen is larger for positions at the edges than those at the center. Hence a polish profile can be generated on the polish pad. The areas contacting the wafer center

are less polished and the areas contacting the wafer edge are more polished. Once this kind of polish profile is formed on a polish pad, the normal polish uniformity is lost. In other words, the intrinsic non-uniformity problem can trigger an extrinsic problem on the polish pad. The only way to counter this is using edge and less at its center. This is due to the fact that the relative velocity between a rotating wafer and a rotating platen is larger for positions at the edges than those at the center. Hence a polish profile can be generated on the polish pad. The areas contacting the wafer center are less polished and the areas contacting the wafer edge are more polished. Once this kind of polish profile is formed on a polish pad, the normal polish uniformity is lost. In other words, the intrinsic non-uniformity problem can trigger an extrinsic problem on the polish pad. The only way to counter this is to use a pad conditioner. A pad conditioner can intentionally condition more at some places and less at others. Another concern with the use of pad conditioner is the down force during conditioning. This down force must be as low as possible, as long as the polish rate remains stable. If the down force is set too high, the resultant high wear rate shortens the pad life. Once the grooves on the pad are worn out, the pad can no longer deliver slurry. The pad conditioner can intentionally condition more at some places and less at others. Another concern with the use of pad conditioner is the down force during conditioning. This down force must be as low as possible, as long as the polish rate remains stable. If the down force is set too high, the resultant high wear rate shortens the pad life.

In the end,the primary purpose of using CMP in back-end interconnect processes is to planarize the surface. A question arises, however, as to how much sacrificial thickness maybe required for polishing away to planarize the surface. Intuitively, the more the thickness polished, the better would be the planarity achieved; however, at the same time, the across-wafer final non-uniformity becomes worse.

3.5 CMP modeling

Modelling blanket wafer polishing is the first step in understanding the polish characteristics and in exploring the appropriate control algorithm. Polish process optimization and control depend on accurate delineation of the roles of macroscopic process parameters, such as down force and relative speed, on polish rate and uniformity. Such analysis is highly simplified for blanket wafer polishing. Successful CMP process modelling also entails a good understanding of the polish mechanism, which is easier to infer from blanket wafer polishing.

The earliest glass polishing model, which can be applied to oxide dielectric polishing, was proposed by Preston (Preston, 1927), using the first principle modelling. According to the model, the polish rate at any position on the wafer is given by equation 3.1:

$$\frac{\triangle H}{\triangle t} = -K_p(\frac{L}{A})\frac{\triangle S}{\triangle t} \tag{3.1}$$

where K_p is the Preston's coefficient, L is the applied load, A is the contact area, and $\triangle S$ is the relative distance travelled between pad and wafer position of interest. The model assumes mechanical abrasion and the chemical effects are lumped into the coefficient. It is usually written as in equation 3.2:

$$\frac{dH}{dt} = -K_p PV \tag{3.2}$$

where P and V are local pressure and relative velocity, respectively. Cook(Cook, 1990) has attempted an extension to Preston's model by incorporating a mechanism in the model. He assumes the slurry particles are responsible for polishing, then models their penetration into the surface as an Hertzian penetration problem. The polish rate is then given by equation 3.3:

$$\frac{\triangle H}{\triangle t} = -\frac{1}{2E}PV \tag{3.3}$$

which is identical to Equation 3.1with Preston's coefficient replaced by $\frac{1}{2E}$ where E is the Young's modulus of the surface being polished.Based on mechanistic models of particle-based wear, other non linear dependencies on pressure and velocity have also been proposed (Tung, 1997).

An controller uses a process model to approximate what change in process setting is necessary to counteract an observed drift in the process output and/or changes in the incoming wafer characteristics. A model used for control does not need to be perfect not as detailed as a simulation model. Thus, mostly black box models for the process are developed and used for control. To obtain a black box model, design of experiments (DOE) is done to identify the important process variables that affect the process output and give the polynomial model. A linear regression fit is then obtained as given in equation 3.4

$$Y_i = AX_i + C_{i-1} + \varepsilon_i + \delta_i \tag{3.4}$$

where Y_i is the output at batch i, A is the process gain. The process noise, ε_i, is assumed to be normally distributed white noise. The parameters δ_i and C_{i-1} in the above equation represent the drift and the estimate of the intercept, respectively. Most of Run to Run controllers (Boning *et al.*, 1996) (Campbell, 1999) (Bulter and Stefani, 1994) (Castillo and Hurwitz, 1997) employ this kind of process model in their control algorithms.

3.6 Conclusion

In this chapter, we introduced a critically important semiconductor process step both in the front-end process and back-end semiconductor process, Chemical Mechanical Polishing Process. We further set this process as the interest of our run-to-run controller and therefore discussed about its output variables and in-

put variables. In addition, Wafer performance metrics are also investigated with a introduction to CMP process problems. The chapter ends with a brief introduction to the first principle modelling of CMP process. It is just the inherent character of CMP process, its non-stationery process disturbance and Ex situ measurement of output due to the nature of polish environment, leading to the popular investigation of run-to-run controller in the use of controlling CMP process, which will be discussed in the next chapter.

Chapter 4

Run to run control in CMP process

4.1 Introduction

The newer approach to solving process problems in the semiconductor manufacturing industry is through a combination of SPC and automatic process control (APC) known as run-to-run control (R2R) (Sachs *et al.*, 1995) (Castillo and Hurwitz, 1997). Run-to-run controllers generally are model-based controllers coupled with an observer of some type. The final element of the run-to-run controller is the control law which specifies how the recipe for the process should be updated (Castillo and Hurwitz, 1997). The task of R2R controller is thus to update the recipe of the manipulated variables for the next wafer by compensating for process changes without increasing the variability of the product. The literature contains both single and multivariate R2R controllers applied to CMP process, such as the MIT Gradual Model Exponentially Weighted moving average (EWMA) controller (Boning *et al.*, 1996), Predictor Corrector Controller (PCC) (Bulter and Stefani, 1994), Optimizing Adaptive Quality Controller (OAQC) (Castillo and Yeh, 198) and Model Predictive R2R controller(MPR2RC) (Campbell, 1999). A

comparison is made between the various R2R controllers proposed in the literature in this section. Before the comparison of Run-to-Run control algorithms, Design of Experiments (DOE) and Response surface modelling (RSM) are also introduced to provide the process model for controller use. Often a model used for control does not need to be perfect. If a controller is not successful in cancelling all of the output drift after one run, feedback will tell the controller to change the input again so that in the second run, the output will be moved even closer to target. This iterative behavior is the fundamental reason why feedback control is robust to a modest amount of error in the control model, and it is also the reason why feedback control can counteract unmodeled disturbance.

4.2 CMP process modeling

4.2.1 Design of experiments

Experimental design methods have found broad application in many disciplines. In fact, We may view experimentation as part of the scientific process and as one of the ways we learn about how systems or process work. Generally, we learn through a series of activities in which we make conjectures about a process, perform experiments to generate data from the process, and then use the information from the experiment to establish new conjectures, which lead to new experiments, and so on. Experiment design is a critically important tool in the engineering world for improving the performance of a manufacturing process. It also has extensive application in the development of new processes and play a major role in engineering design activities, where new products are developed and existing ones improved. Some application of experimental design in engineering design include

- Evaluation and comparison of basic design configurations.
- Evaluation of material alternatives.

- Selection of design parameters so that the product will work well under a wide variety of field conditions, that is, so that the product is robust.
- Determination of key product design parameters that impact product performance.

The aim of DOE in this thesis is that we want to find the most influential input variables among the input variables to CMP process and further set up a linear regression model as our process model.

In general, experiments are used to study the performance of processes and systems. The process can be represented by the model shown in figure 4.1. We can

Figure 4.1. General model of a process or system

usually visualize the process as a combination of machines, methods, people,and other resources that transform some input (often a material) into an output that has one or more observable responses (Montgomery, 1996). Some of the process variables x_1, x_2, ..., x_p are controllable, whereas other variables z_1, z_2, ..., z_q are uncontrollable. The objectives of the experiment may include the followings:

- Determining which variable are most influential on the response y.

- Determining where to set the influential x's so that y is always near the desired nominal value.
- Determining where to set the influential x's so that variability in y is small.
- Determining where to set the influential x's so that the effects of the uncontrollable variables z_1, z_2, ..., z_q are minimized.

The general approach to planning and conducting the experiments is called the strategy of experimentation. There are several strategies that an experimenter could use. When there are several factors in the experiments, the correct approach to dealing with it is to conduct a factorial experiment. This is an experiment strategy in which factors are varied together, instead of one at a time. Generally, if there are K factors, each at two levels, the factorial design would require 2^k runs. In our case, since CMP is a MIMO process, all four parameters-Relative velocity, back-pressure, profile, down-force could be investigated in a 2^4 full factorial design.

In a 2^k factorial design, it is easy to express the results of the experiment in terms of a regression model in the following forms 4.1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{4.1}$$

where y is the response variable, the x's are a set of regressor or predictor variables, the β's are the regression coefficients and ε is an error term, assumed to be normally distributed between 0 and σ^2. In general, the regression coefficients in these models are estimated using the method of least squares; that is , the β's are chosen so as to minimize the sum of the squares of the errors (the ε's) .

However, as the number of factors in a 2^k factorial design increases, the number of runs required for a complete replicate of the design rapidly outgrows the resource of most experiments. If the experiments can reasonably assume that certain high-order interactions are negligible, then information on the main effects are

Table 4.1. 2^{4-1} design matrix for WCMP process-DOE

TTspeed	TRspeed	Downforce	BSP	RR	Nu%
50	50	311	138	2389	4.32
50	50	414	276	2707	3.64
50	100	311	276	2247	3.35
50	100	414	138	2988	4.2
90	50	311	276	3420	9.9
90	50	414	138	4385	11.1
90	100	311	138	3607	6.55
90	100	414	276	4672	15.7

low-order interactions may be obtained by running only a fraction of the complete factorial experiment. In our experiments, we have four factors, each at two levels, are on interest. But the experiments cannot afford to run all $2^4 = 16$ treatment combinations and we also really assume the second-order interaction are negligible, then it suggests a one-half fraction of a 2^3 design. A 2^3 factorial design in the 4 parameters-table speed, top-ring speed, down-force and back pressure was performed as described in Table 4.1

4.2.2 Response surface modelling

In the case of 2^k design, it is extremely easy to find the least square estimates of the β's. The least square estimate of any regression coefficient β is simply one-half of the corresponding factor effect estimate. If the output variable y is plotted with respect to the two process variables x_1 and x_2 in a 3D plane, it is called response surface plot, and the regression model used to generate the graph is called first-order response surface model.

The regression model built up in the DOE data in the last section is obtained as follows, where we use relative velocity to take place of table speed velocity and

top-ring velocity thus only 3 factors are in our interest.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 12.1628 & 7.4976 & -0.5851 \\ 0.0651 & 0.0255 & 0.0166 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 765.9439 \\ -3.9931 \end{bmatrix} \tag{4.2}$$

where x_1 is the relative speed, x_2 is the down-force, and x_3 is the back pressure.

Figure 4.2, 4.3 shows the response surface plot obtained by the DOE data listed in 4.2:

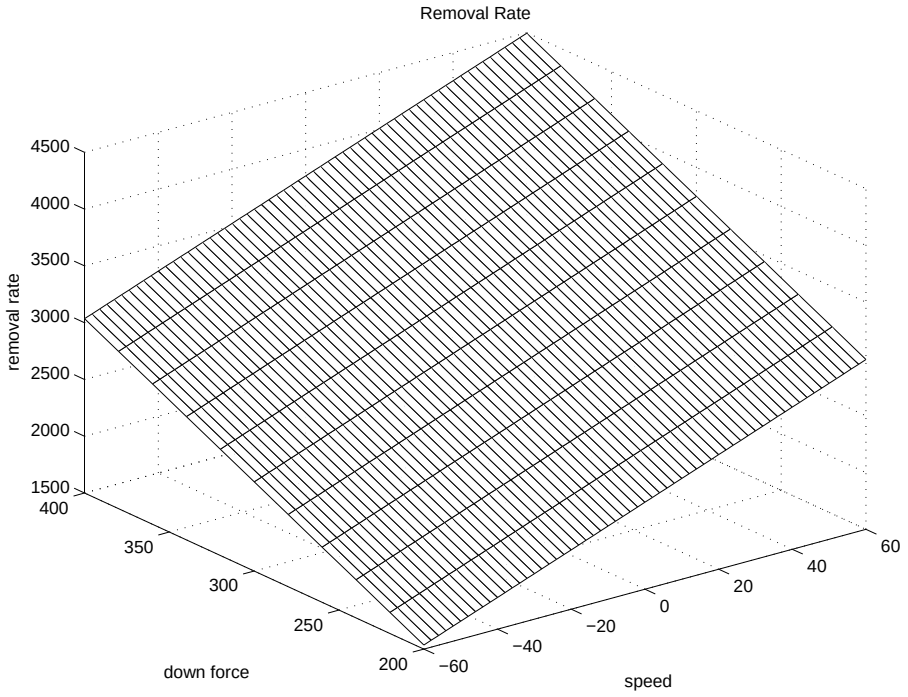

Figure 4.2. Response surface models for removal rate

Inspection of the response surface makes interpretation of the results of an experiment very simple. From the response surface plot (RSM) , the experimenter might select an optimum set of conditions for doing the process. It can also be used in process robustness studies and process improvement and optimization. Thus the objective of every designed experiment is a quantitative model of the process.

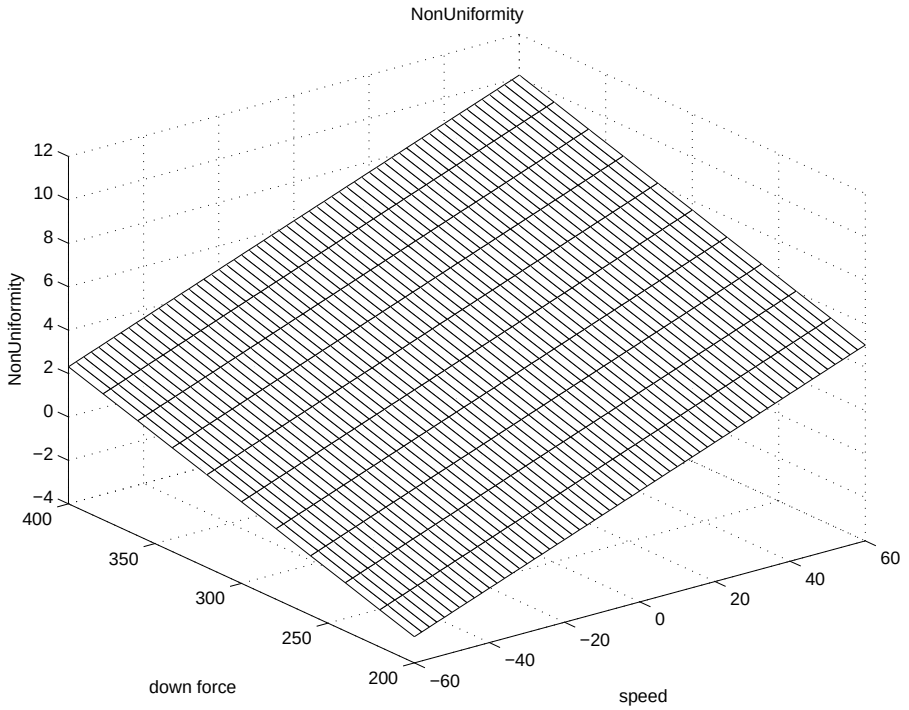

Figure 4.3. Response surface models for non-uniformity

4.3 Model-based run to run control algorithm

There are many different methods that a run-to-run controller can be formulated in order to perform the necessary control tasks.However, most all controllers will have a similar structure, regardless of the detail. The final element of the Run-to-Run controllers is the control law which specifies how the recipe for the process should be updated (Castillo and Hurwitz, 1997).

Model-based controllers coupled with a filter are often used as Run-to-Run controller. Four such controllers (EWMA, PCC, and OAQC) are briefly explained in this section. Performances of these schemes are compared using simulation results.

4.3.1 EWMA controller

EWMA controller is identical to the Internal Model Control (IMC) scheme. Block diagram in Figure 4.4 explains the operation of this controller. The error between the outputs of the process and its model is used as feedback to rectify the process so that the error is minimized.

Figure 4.4. Block diagram of EWMA controller

The process is represented by the following linear regression model,

$$Y_i = AX_i + C_{i-1} + \varepsilon_i + \delta_i \tag{4.3}$$

where Y_i is the output at batch i, A is the process gain. We assume that the process gain, A, is constant and can be obtained from off-line DOE data. The control input (recipe) , X_i, at the i^th batch is obtained from the past errors upto the $(i-1)^th$ batch. The deviation of the output from the desired value is assumed to be caused by process noise and equipment-related drift, such as wear and tear of the pad. The process noise, ε_i, is assumed to be normally distributed white noise. The parameters δ_i and C_{i-1} in the above equation represent the drift and the estimate of the intercept, respectively.

The EWMA controller only adapts the offset term C_i based on the exponential smoothing of the previous estimates of the offset. The intercept is updated recursively by a filter of the form:

$$C_i = \omega(Y_i - AX_i) + (1 - \omega)C_{i-1} \tag{4.4}$$

where ω is the exponential weighting factor or tuning parameter of the filter, which takes a value between 0 and 1 based on the desired properties. Small values of ω are appropriate for systems with small deterministic drifts and relatively larger process noise. On the other hand, highly correlated output errors are better compensated for using higher values of the weighting factors.

The control recipe for EWMA controller is a plant inverse of the form:

$$X_{i+1} = \frac{T - C_i}{A} \tag{4.5}$$

where T is the target of the control process. Extension to MIMO systems may require optimization to find the constrained least-squared solutions to the control law.

From an advance process control (APC) point of view, the most interesting theoretical result is the ability to interpret the EWMA as an integral controller with a measurement delay of one run. It has been show (Sachs *et al.*, 1995) that for a pure gain system,the resulting control law can be written as:

$$X_i = -\frac{\omega}{A}\sum_{i=1}^{i-1}(Y_i - T) + X_1 \tag{4.6}$$

Hence an EWMA controller can be regarded as an optimal PID controller for a second-order dynamic process under the ARIMA (1,1) disturbance (Box and Jenkins, 1994). Box and Jenkins also show that an EWMA-based controller is a minimum mean square error (MMSE) controller when the underlying process

disturbance follows the ARIMA (1,1) process.

4.3.2 Predictor corrector controller

Predictor Corrector Control (PCC), is an extension of the standard EWMA scheme (Bulter and Stefani, 1994). The controller, shown in Fig 4.5, includes a prediction filter in addition to the smoothing filter of EWMA. Unlike EWMA controller which assumes locally constant value for the intercept, this control law predicts the future changes in the value of the intercept.

Figure 4.5. Block diagram of PCC controller

The equations describing the two filters of PCC are given below:

$$\begin{aligned} C_i &= \omega_1(Y_i - AX_i) + (1-\omega_1)C_{i-1} \\ P_i &= \omega_2(Y_i - AX_i - C_{i-1}) + (1-\omega_2)P_{i-1} \end{aligned} \tag{4.7}$$

where ω_1, ω_2 are the weights for the first and second EWMA equations, respec-

tively, and P_i is used to compensate for the error incurred by C_i. In other words, P_i is the drift speed, which is used to compensate for the drifting process. $C_i + P_i$ is then used to estimate the offset at run i+1 and the process recipe at i+1 becomes

$$X_{i+1} = \frac{T - (C_i + P_i)}{A} \tag{4.8}$$

This modification of the EWMA controller makes it possible to compensate for the lag in target tracking when a process in undergoing a drift. With appropriate choice of ω_1, ω_2, PCC controller can remove the drift completely from the process (Bulter and Stefani, 1994). However, tuning the second filter is not as intuitive as a single EWMA filter.

Similarly, from the APC standpoint, the estimate for $C_i + P_i$ in the run i+1 can be rewritten as (Chen and Guo, 2001):

$$C_i + P_i = \omega_I \sum_{j=1}^{i} e_j + \omega_{II} \sum_{j=1}^{i} \sum_{k=1}^{j} e_k + C_0 + (\omega_1 i + 1) P_0 \tag{4.9}$$

where $e_k = Y_k - T$,$\omega_I = \omega_1 + \omega_2 - \omega_1\omega_2$,$\omega_{II} = \omega_1\omega_2$. This is equivalently an Integral-double-Integral(I-II)controller. In this I-II controller, the process recipe is proportional to the summation of the output errors and to the summation of summations of the output errors. This controller can be shown to be an MMSE controller for the processes subject to ARIMA (2,2) disturbance (Box and Jenkins, 1994).

4.3.3 OAQC controller

The Optimizing Adaptive Quality Controller (OAQC) is designed to seek and maintain optimum operating conditions for a MIMO nonlinear quadratic process (Castillo and Yeh, 198). It consists of two elements:

1) An online recursive least squares estimator for identification of controller

parameters.

2) A Run to Run controller for control regulation.

Figure 4.6 shows the control diagram of OAQC controller.

Figure 4.6. Block diagram of OAQC controller

In OAQC, second-order MIMO Hammerstein transfer function of the model for one step ahead forecast is of the following form

$$\hat{Y}_i = LY_{i-1} + Mi + NX \tag{4.10}$$

where N is process gain matrix, M is the constant drift speed matrix.

A two-step process is used to update the control recipe at the following run. First, the model parameters L, M, and N are updated using recursive least square estimation algorithm. Defining $\theta_i = [L\ N\ M]^T$ and $\phi_i = [Y_{i-1}\ X_i\ i]^T$ such that $\hat{Y}_{i+1} = \theta_i^T * \phi_i$. We can estimate the parameters for output at run j by using the formulae shown below (Chamness *et al.*, 2001).

$$
\begin{aligned}
K_i^{[j]} &= \frac{P_{i-1}^{[j]}\phi_i^{[j]}}{\lambda + \phi_i^{[j]T} P_{i-1}^{[j]}\phi_1^{[j]}} \\
e_i^{[j]} &= y_i^{[j]} - \phi_i^{[j]T}\hat{\theta}_{i-1}^{[j]} \\
\hat{\theta}_i^{\,[j]} &= \hat{\theta}_{i-1}^{[j]} + K_i^{[j]} e_i^{[j]} \\
P_i^{[j]} &= [I - K_i^{[j]}\phi_i^{[j]T}]\frac{P_{i-1}^{[j]}}{\lambda} + R_i^{[j]}
\end{aligned}
\tag{4.11}
$$

where $R_i^{[j]} = K_i^{[j]T} P_{i-1}^{[j]} \phi_i^{[j]I}/dim$, $dim = p + n + 1$, p is the number of outputs, n is the number of control recipes, and I is the identity matrix with dimension, dim.

The optimal controller that minimizes the objective function $\phi = \|T - \hat{y_i}\|_W^2 + \|X_i - X_{i-1}\|_\tau^2$ is given by the following 4.12:

$$
X_i = (N^T W N + \tau)^{-1}(N^T W(T - Mi - LY_{i-1}) + \tau X_{i-1}) \tag{4.12}
$$

where τ and W are weighting matrices in the objective function.

The OAQC controller provides the optimum estimation of the non-linear regression model such that the best control action is achieved. Once optimal operating conditions is reached, the OAQC maintains the process under control running at that conditions. Hence the OAQC acts both as an "optimizer" and as a "controller". Moreover, the OAQC accounts for process nonlinearity because of recursive parameter estimation, and the initialization of the control model need not be very accurate.

4.3.4 MPC controller

It has also been proposed (Mullins *et al.*, 1997) that run-to-run control be implemented using linear model predictive control (Muske and Rawlings, 1993). The LMPC run-to-run controller uses a traditional state-space model for the process.

However, instead of time as the independent variable of the state-space model, batch number is the independent variable. A series of batches is equivalent to the time samples of a continuous process. By modelling a series of batches,dynamic behavior can be incorporated into the model. Now disturbance are modelled as dynamic behavior of the system, instead of unknown disturbances. Like the PCC, this allows offset free target tracking if there is no model mismatch. There are several advantages of the MPC formulation that the authors cite as motivating factors of their proposal. These include direct extension to higher order dynamics, MIMO systems, system with time-delay, and systems with input and output constraints. Also, the tuning objectives of the MPC algorithm allows weighting of inputs, outputs, and input rates of change for optimal specification of closed loop performance. The authors illustrated the use of MPC in simulation of CMP process.

Although model predictive control has traditionally been applied for real-time control of continuous processes, it can be easily be applied to the run-to-run control problem with- out major modification (Campbell and Toprac, 2001). For a discrete parts manufacturing process, the in- dependent variable in the state-space model becomes run number, rather than time. This algorithm uses a kalman filter to determine the drift of the process and the process model is used to predict the future behavior of the system (Campbell and Toprac, 2001) (Campbell, 1999) .

$$\begin{aligned} y_i &= Du_i + d_i \\ d_{i+1} &= di + \alpha_i \end{aligned} \tag{4.13}$$

In standard state-space form, we have,

$$\begin{aligned} x_{i+1} &= Ax_i + w_i \\ y_i &= Cx_i + Du_i \end{aligned}$$

where,

$$A = \begin{bmatrix} I & I \\ 0 & I \end{bmatrix}, C = [I \ \ 0], x_i = \begin{bmatrix} d_i \\ \alpha_i \end{bmatrix}$$

and w_i is normally distributed noise in the process states. Kalman filter gain is then computed and states are estimated by the following equation 4.14

$$\begin{aligned} \hat{x}_{i+1} &= A\hat{x}_i + J(y_i - \hat{y}_i) \\ \hat{y}_i &= C\hat{x}_i + Du_i \end{aligned} \tag{4.14}$$

The unconstrained solution for model predictive control (MPC) algorithm is given by

$$u_i = D^T(DD^T)^{-1}(T - C\hat{x}_i - Du_{i-1}) + u_{i-1} \tag{4.15}$$

4.4 Performance analysis

Performances of the three control algorithms are compared through simulation using the benchmark process model (plant) described by (Ning *et al.*, 1996):

$$y[n] = C + f(u[n]) + \varepsilon_n + \delta_n \tag{4.16}$$

where δ_n is a linear drift with constant drift speed $\delta = [-17 \ \ 1.5]'$ and ε_n is a normally distributed white noise with mean zero and covariance

$$\Lambda = \begin{bmatrix} 665.64 & 0 \\ 0 & 5.29 \end{bmatrix} \tag{4.17}$$

$f(u[n])$ is a full second-order polynomial function of the inputs with the following form:

$$f(u[n]) = \sum_{i=0}^{3}\sum_{j=0}^{3}\beta(i,j)u(i)u(j) \tag{4.18}$$

where

$$\beta = \begin{bmatrix} 1386.5 & 381.02 & -112.19 & 3778.8 & -21.301 & 8.7159 & 24.953 \\ 1520.8 & 2365.6 & 2923.5 & 281.66 & -3.9419 & -1.0754 & 1.406 \\ 37.082 & -17.642 & -11.974 & -164.99 & 28.150 & 249.17 & 0.025067 \\ 0.33797 & -72.274 & -94.222 & -26.175 & -13.505 & 36.691 & 32.929 \end{bmatrix}$$

The simulation model for the EWMA and PCC controllers have the form given in Equation 4.3, and with the outputs: Removal Rate and Within Wafer Non-uniformity. We used

$$A = \begin{bmatrix} 5.018 & -0.665 & 16.34 & 0.845 \\ 13.67 & 19.95 & 27.52 & 5.25 \end{bmatrix} \quad C = \begin{bmatrix} -138.21 \\ -627.32 \end{bmatrix} \tag{4.19}$$

The target removal rate (T) and the target Non-uniformity are assumed to be 1800 and 300, respectively. Initial non-uniformity is set to 150. Simulation of the process is done using the three control schemes. The EWMA controller is simulated with $\omega = 0.6$ and the PCC with $\omega_1 = 0.6$ and $\omega_2 = 0.3$. Controlled response (Removal rate and Non-uniformity) is shown in figure 4.7 and figure 4.8.

The results obtained with the three control schemes are compared quantitatively by calculating the mean squared error (MSE) between the response and the target for each run.

It is evident from Fig 4.7 that all three algorithms provide good control of the simulated CMP process (with linear drift and normally distributed white noise). Process drift is well compensated for keeping the removal rate near the target value. The best result is obtained for both removal rate and non-uniformity using PCC.

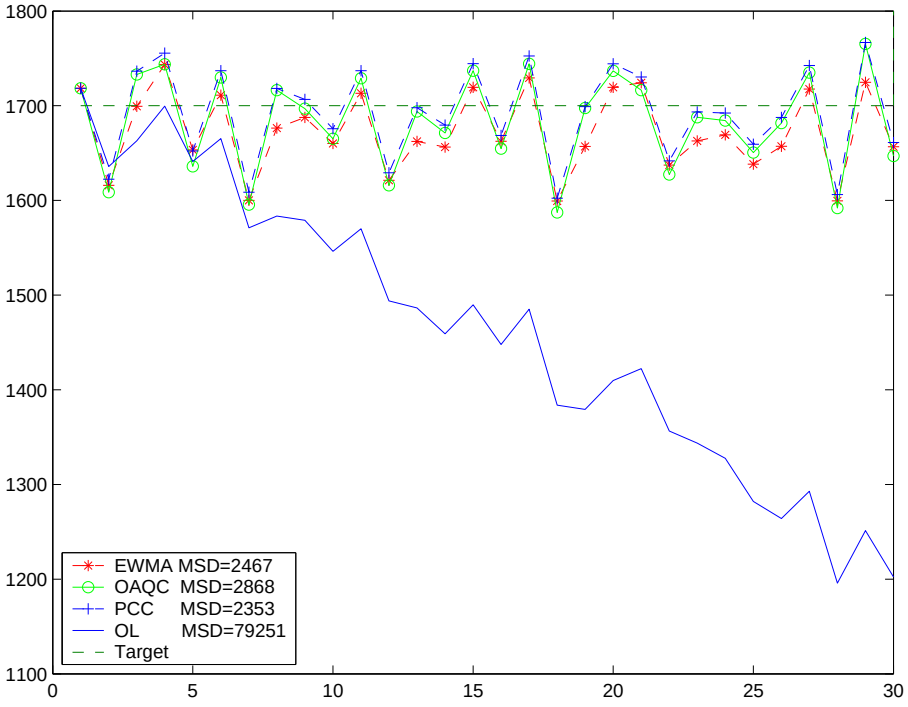

Figure 4.7. Removal rate comparison of three R2R control algorithms

Table 4.2. Comparison of results using PCC, EWMA and OAQC

MSE	Removal Rate	Non-Uniformity
PCC	2353	16
EWMA	2467	18
OAQC	2868	19

The mean square error for both removal rate and non-uniformity is summarized in the following table.

The weighting factors used in the PCC algorithm are fixed parameters. However, the drift can be removed completely from the process only with the appropriate choice of these factors. T. Smith used a self-tuning EWMA controller with the help of artifical neural network (ANN) function approximation which dynamically updates the controller parameters (Smith and Boning, 1996). Such strategy needs significantly large amount of training data to find out the functional mapping between the disturbance state of the process and the corresponding weighting factors

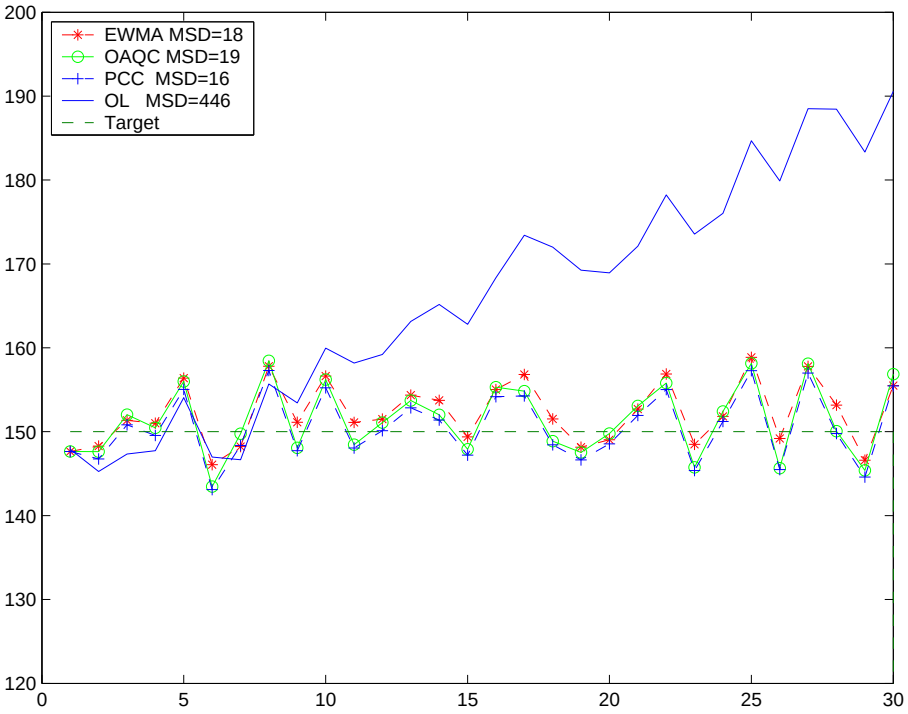

Figure 4.8. Non-uniformity comparison of three R2R control algorithms

of EWMA. Poor training of the ANN may cause fluctuations in the weighting factors, resulting in additive noise and poor tracking of drift.

4.5 Conclusions

In this chapter, process model is first obtained via DOE and RSM. Then different run-to-run controllers which have a similar structure coupled with a filter are compared and further simulated to evaluate their performance in CMP process control. Of these controllers, simple EWMA and PCC are the most used in semiconductor manufacturing while the optimum value of the weights of their algorithms is always difficult to select. A solution for automatically tuning the weights of the EWMA and PCC is developed in the next chapter.

Chapter 5

Self-tuning PCC controller

5.1 Introduction

Proper choice of controller parameters is critical to the performance of the system. Usually process engineers tune the controller according to the disturbance state (magnitude of drift and random noise), this should be a very tiring work.Additi - onally, in PCC controller, there exist two weighting factors. Two tuning parameters allow more flexibility in tuning than EWMA. However, tuning the second filter is not as intuitive as a single EWMA controller. Therefore the ability to dynamically update the double EWMA filter weighting factors is important and essential for maximum controller performance.

To properly select the weighting factors in PCC controller, first let us see how one might choose a value for ω in single EWMA filter situation. It is important to consider first how ω affects the control output. A high value of ω increases the impact of the current model error on the control action. Hence a high value of ω has a fast dynamic response. On the other hand, if there is noise in the process, a high value of ω would also cause control actions to increase the variance to the output. In other words, it increases the process noise. On the contrary, small value of ω smoothes the previous model errors and then has less sensitivity to the noise

with relatively slow dynamic response and less control action over true disturbance. Therefore it would seem that given a fixed amount of noise and expectations on the disturbances, one could choose an optimal value for this parameter. However, for any value of ω,results reveal that there is always a steady-state error (Boning *et al.*, 1995). This is caused by the fact that each time the EWMA updates the model to compensate for the amount the process has drifted, the process drifts again in the following run. Similar weighting consideration and simulation can be applied to controllers that eliminate this offset, which motivate the use of PCC control. However, there is still difference between the true offset and the predicted offset due to the change of the noise and drift. Here we propose a minimum variance controller to realize the adaptive optimization.

5.2 Adaptive filter theory

The performance of the EWMA controller and PCC controller depends heavily on the proper selection of the weighting factors or forecasting parameters. In this chapter, a methodology for self tuning the two forecasting parameters by using variable step size least mean square estimation in PCC controller is developed and discussed in full detail.

5.2.1 Introduction

Clearly, depending upon the time required to meet the final target of the adaption process, which we call convergence time, and the complexity/resources that are available to carry out the adaption, we can have a variety of adaption algorithms and filter structures. The term filter is commonly used to refer to any devices or system that takes a mixture of particles/elements from its input and process them according to some specific rules to generate a corresponding set of particles/elements as its output. In the context of signals and systems,particles/elements are the fre-

quency components of the underlying signals and, traditionally, filters are used to retain all the frequency components that belong to a particle band of frequencies, while rejecting the rest of them, as much as possible. Filters may be either linear or non-linear. We only consider linear filters in discrete time signals. Thus, all the signals will be represented by sequences, such as x(n). Figure 5.1 depicts a general schematic diagram of a filter emphasizing the purpose for which it is used in different problems addressed/discussed in this book. In particular, the filter is

Figure 5.1. Schematic diagram of a little emphasizing its role in reshaping the input signal to match the desired signal

used reshape certain input signals such a way that its output is a good estimate of the given desired signal. The process of selecting the filter parameters (coefficients) so as to achieve the best match between the desired signal and the filter output is often done by optimizing an appropriate defined performance function. The performance function can be defined in a statistical or deterministic framework. In the statistical approach, the most commonly used performance function is the mean-square value of the error signal. For stationery input and desired signals, minimizing this mean square error results is the well-known Wiener filter, which is said to be optimum in the mean square sense. In the deterministic approach, the usual choice of performance function is a weighted sum of the squared error signal. Minimizing this function results in a filter which is optimum for the given set of data.

As mentioned in the above, the filter required for estimating the given signal

can be designed using either the stochastic or deterministic formulations. In the deterministic formulation, the filter design requires the computation of certain average quantities using the given set of data that the filter should process. On the other hand, the design of Wiener filter requires a priori knowledge of the statistics of the underlying signals and a large number of realizations of the underlying signal sequence which is not practical in the reality. So we use adaptive filters to solve this problem. The most commonly used structure in the implementation of adaptive filters is the transversal structure, depicted in Figure 5.2. Here, the adaptive filter

Figure 5.2. Adaptive transversal filter

has a single input, x(n), and an output, y(n). The sequence d(n) is the desired signal. The output, y(n), is generated as linear combination of the delayed samples of the input sequence, x(n), according to the equation

$$y(n) = \sum_{i=0}^{N-1} w_i(n)x(n-i) \tag{5.1}$$

where the $w_i(n)$ are the filter tap weights (coefficients) and N is the filter design.

We refer to the input samples, x(n-i), for i=0, 1, ..., N-1, as the filter tap inputs. The tap weights, $w_i(n)s$, which may vary in time, are controlled by the adaption algorithm.

According to the Wiener filter theory, which comes from the stochastic framework, the optimum coefficients of a linear filter are obtained by minimizing of its mean-square error (MSE). As already noted, the minimization of MSE requires certain statistics obtained through ensemble averaging,which may not be possible in practical applications. To come up with simple recursive algorithms, very rough estimates of the required statistics are used. In fact, the celebrated least-mean-square (LMS) algorithm, which is the most basic and widely used algorithm in various adaptive filtering applications. It turns out that this very rough estimate of the MSE, when used with a small step-size parameter in searching for the optimum coefficients of the Wiener filter, leads to a very simple and yet reliable adaptive algorithm. The main disadvantage of the LMS algorithm is that its convergence behavior is highly dependent on the power spectral density of the filter input. When the filter input is white, i.e. its power spectrum is flat across the whole range of frequencies, the LMS algorithm converge fast. However, when certain frequency bands are not well excited, some slow modes of convergence appear, resulting in very slow convergence compared with the case of white input. Another problem is step-size parameter. A large step-size parameter may be required to minimize the transient time of the LMS algorithm. On the other hand, to achieve a small misadjustment a small step-size parameter has to be adopted. In next section, the variable step size LMS (VSLMS) algorithm which is introduced in this section is an effective solution to this problem (Boroujeny, 1998).

Table 5.1. Summary of the LMS algorithm

Input:	Tap-weight vector,w(n), Input vector,x(n), and desired output, d(n)
Output:	Filter output,y(n), Tap-weight vector update,w(n+1).
1. Filtering:	$y(n) = w^T(n)x(n)$
2. Error estimation:	$e(n) = d(n) - y(n)$
3. Tap-weight vector adaptation:	$w(n+1) = w(n) + 2\mu e(n)x(n)$

5.2.2 Variable step-size LMS algorithm

Before we introduced the Variable step size LMS Algorithm, let us review roughly what the LMS is. The LMS algorithm was first proposed by Widrow and Hoff in 1960 and is the most widely used adaptive filtering algorithm, which can be attributed to its simplicity and robustness to signal statistics. Table 5.1 summarizes the LMS algorithm. The major problem of the LMS recursion is its slow convergence when the underlying input process is highly colored.

We can also notice that the step-size parameter, μ, plays a significant role in controlling the performance of the LMS algorithm. There are conflicting requirements between the fast convergence speed and a small mis-adjustment, so a compromised solution has to be adopted.

The VSLMS algorithm works on the basis of a simple heuristic that comes from the mechanism of the LMS algorithm. Each tap of the adaptive filter is given a separate time-varying step-size parameter and the LMS recursion is written as in

equation 5.2

$$w_i(n+1) = w_i(n) + 2\mu_i(n)e(n)x(n-i), for 1mmi = 0, 1, ..., N-1, \qquad (5.2)$$

where $w_i(n)$ is the ith element of the tap weight vector $w(n)$ and $\mu_i(n)$ is its associated step-size parameter at iteration N. The adjustment of the step-size parameter $\mu_i(n)$ is done as follows. The corresponding stochastic gradient term $g_i(n) = e(n)x(n-i)$ is monitored over successive iterations of the algorithm and $\mu_i(n)$ is increased if the latter term consistently shows a positive or negative direction. As the weights converge to their optimum values, the change of the signs is detected and the step size parameters are gradually reduced to some minimum values. To ensure that the step-size parameters do not become too large or too small, upper and lower limits should be specified for each step-size parameter. Following the above argument, the VSLMS algorithm step-size parameters, the $\mu_i(n)s$, may be adjusted using the following recursion:

$$\mu_i(n) = \mu_i(n-1) + \rho sign[g_i(n)]sign[g_i(n-1)] \qquad (5.3)$$

where ρ is a small positive step-size parameter. This results in the following alternative step-size parameter update equation:

$$\mu_i(n) = \mu_i(n-1) + \rho g_i(n)g_i(n-1) \qquad (5.4)$$

The derivation to determine the range of the step-size parameters that ensure the stability of the VSLMS algorithm is very difficult, because of the time-variation of the step-size parameters. Here, we adopt a simple approach by assuming that the step-size parameters vary slowly so that for the stability analysis they may be assumes fixed and use the analogy between the resulting VSLMS algorithm equations and the conventional LMS algorithm to reach a result which has been

found to be reasonable. The set of update equations 5.2 may be written in the vector form as

$$\mathbf{w(n+1)} = \mathbf{w(n)} + \mathbf{2}\mu\mathbf{(n)e(n)x(n)} \tag{5.5}$$

where $\mu(n)$ is a diagonal matrix consisting of the step-size parameters $\mu_0(n)$, $\mu_1(n)$, ..., $\mu_{N-1}(n)$. Equation 5.5 may further be rearranged as

$$\mathbf{v(n+1)} = (\mathbf{I} - \mathbf{2}\mu\mathbf{(n)x(n)x^T})\mathbf{v(n)} + \mathbf{2}\mu\mathbf{(n)e_0(n)x(n)} \tag{5.6}$$

where $v(n) = w(n) - w_0$ is the weight-error vector. Now we may argue that to ensure the stability of the VSLMS algorithm, the scalar step-size parameter μ should be replaced by the diagonal matrix $\mu(n)$. This leads to the inequality

$$\mathbf{tr}[\mu\mathbf{(n)R}] < \frac{\mathbf{1}}{\mathbf{3}} \tag{5.7}$$

as a sufficient condition which assures the stability of the VSLMS algorithm. Although the inequality 5.7 may be used to impose some dynamic bounds on the step-size parameters $\mu_i(n)$ as the adaptation of the filter proceeds, this leads to a rather complicated process. Instead, in practice we usually prefer to limit all $\mu_i(n)s$ to the same maximum value, say μ_{max}. The minimum bound that may be imposed on the variable step-size parameter, the $\mu_i(n)s$, can be as low as zero. However, in actual practice a positive bound is usually used so that the adaptation process will be on all the time and possible variations in the adaptive filter optimum tap weights can always be tracked. Here, we use the notation μ_{min} to refer this lower bound. Table 5.2 summarizes an implementation of the VSLMS algorithm.

Table 5.2. Summary of an implementation of variable step-size LMS algorithm

Input:	Tap-weight vector, w(n), Input vector, x(n), Gradient terms $g_0(n-1), g_1(n-1), ..., g_{N-1}(n-1)$, Step-size parameters, $\mu_0(n-1), \mu_1(n-1), ..., \mu_{N-1}(n-1)$ and desired output, d(n)
. Output:	Filter output, y(n), Tap-weight vector update, w(n+1), Gradient terms $g_0(n), g_1(n), ..., g_{N-1}(n)$, and updated step-size parameters, $\mu_0(n), \mu_1(n), ..., \mu_{N-1}(n)$

1. Filtering:

$$y(n) = w^T(n)x(n)$$

2. Error estimation:

$$e(n) = d(n) - y(n)$$

3. Tap-weight vector adaptation:

For i=0,1,...,N-1

$g_i(n) = e(n)x(n-i)$

$\mu_i(n) = \mu_i(n-1) + \rho sign[g_i(n)]sign[g_i(n-1)]$

$if \mu_i(n) > \mu_{max}, \mu_i(n) = \mu_{max}$

$if \mu_i(n) < \mu_{min}, \mu_i(n) = \mu_{min}$

$w_i(n+1) = w_i(n) + 2\mu_i(n)g_i(n)$

end

5.3 Self-tuning PCC controller strategy

We use a recursive algorithm to optimize the weighting factors of a PCC controller. The objective is to minimize the mean square error between the measured offset and the estimated offset using the variable step size LMS algorithm. The algorithm hinges on simple representation of the optimal double EWMA filter weighting factors.

Let us consider the simple process model given in last chapter and the standard PCC controller.

$$\begin{aligned} C_i &= \omega_1(Y_i - AX_i) + (1-\omega_1)C_{i-1} \\ P_i &= \omega_2(Y_i - AX_i - C_{i-1}) + (1-\omega_2)P_{i-1} \end{aligned}$$

Then the control recipe is,

$$X_{i+1} = \frac{T - (C_i + P_i)}{A} \tag{5.8}$$

Choice of the two weighting factors is not as intuitive as single EWMA controller. We use an adaptive self-tuning algorithm to find the optimum values for the weighting factors. The block diagram in figure 5.3 describes the architecture of the proposed controller.

Figure 5.3. Block diagram of self-tuning PCC controller

We now derive the recursive algorithm. PCC controller equation can also be re-structured as

$$C_i = C_{i-1} + \omega_1(Y_i - T + P_{i-1}) \tag{5.9}$$

$$P_i = P_{i-1} + \omega_2(Y_i - T) \tag{5.10}$$

If C_m and P_m are the measured values for the intercept and drift speed, respectively, then

$$C_m = C_i + e_1 \tag{5.11}$$

$$P_m = P_i + e_2 \tag{5.12}$$

Equation 5.11 and 5.12 can be re-written as

$$C_m = C_{i-1} + \omega_1(Y_i - T + P_{i-1}) + e_1$$
$$C_m - C_{i-1} = \omega_1(Y_i - T + P_{i-1}) + e_1 \tag{5.13}$$

and

$$P_m - P_{i-1} = \omega_2(Y_i - T) + e_2 \tag{5.14}$$

Combining these two equations,we get

$$C_m - C_{i-1} + P_m - P_{i-1} = \omega_2(Y_i - T) + e_2 + \omega_1(Y_i - T + P_{i-1}) + e_1 \tag{5.15}$$

Writing these equations using vector notation,

$$d(i) = WX + e(i) \tag{5.16}$$

where $d(i) = C_m - C_{i-1} + P_m - P_{i-1}$,W=$[\omega_2\ \omega_1]$, $X = [Y_i - T\ Y_i - T + P_{i-1}]^T$. Then the variable step size algorithm (Boroujeny, 1998) is used to minimize the error $e(i) = d(i) - WX$ as follows:

$$g(i) = e(i)(Y_i - T) \tag{5.17}$$

$$\mu(i) = \mu(i-1) + \rho g(i)g(i-1) \tag{5.18}$$

$$\omega_1(i) = \omega_1(i)^{-1} + 2\mu(i)g(i)^{-1} \tag{5.19}$$

$$\omega_2(i) = \omega_2(i)^{-1} + 2\mu(i)g(i)^{-1} \tag{5.20}$$

subject to the constraint

$$if\ \omega_1(i) > \omega_{1max}, \omega_1(i) = \omega_{1max}$$
$$if\ \omega_1(i) < \omega_{1min}, \omega_1(i) = \omega_{1min} \tag{5.21}$$

$$if\ \omega_2(i) > \omega_{2max}, \omega_2(i) = \omega_{2max}$$
$$if\ \omega_2(i) < \omega_{2min}, \omega_2(i) = \omega_{2min} \tag{5.22}$$

The algorithm can be explained as follows. The gradient $g(i)$ in Equation 5.17 is monitored over successive iterations of the algorithm and $\mu(i)$ is increased if the latter term consistently shows a positive or negative direction. This happens when the adaptive filter has not yet converged. As the adaptive filter tap weights converge to some vicinity of their optimum values, the average of the stochastic gradient terms approaches zero and hence they change signs more frequently. This is detected by the algorithm and the corresponding step size parameters are gradually reduced to some minimum values. If the process and the environment change and the algorithm begins to hunt for a new optimum point, then the gradient term will indicate consistent (positive or negative) directions, resulting in an increase in the corresponding step size parameters. To ensure that the weights do not go out of bounds, they are restricted between a maximum and a minimum value. The algorithm searches for the optimum value within these bounds. The feedback factor μ, ρ should be selected carefully to ensure proper functioning of the algorithm.

5.4 Simulation results

Performance of the self tuning control design is evaluated using simulation with the benchmark process model given in (Chamness *et al.*, 2001), (Ning *et al.*, 1996), which has been already employed in chapter III. The EWMA and PCC controllers are used for comparison. All the three algorithms are simulated under the same conditions: i.e. the same noise, drifting, disturbance, model and model error. The output matrix consists of removal rate and non-uniformity. The process is buried in normally distributed white noise ε_k and linear drift δ_k. The control recipe includes four parameters (speed, down force, back pressure and pad profile). Mean Square Error (MSE) is calculated and taken as the main metric for evaluation.

5.4.1 Linear process model

The process model (Equation 5.23) is used here as both the real process model and the internal model of the controller.

$$Y_k = \theta u_k^T + C + \varepsilon_k + \delta_k \tag{5.23}$$

where δ_k is a linear drift with constant drift speed $\delta = [-17\ 1.5]'$ and ε_k is a normally distributed white noise with mean zero and covariance

$$\Lambda = \begin{bmatrix} 665.64 & 0 \\ 0 & 5.29 \end{bmatrix}$$

and θ, the system gain and C, the intercept are given by the following values

$$\theta = \begin{bmatrix} 50.18 & -6.65 & 163.4 & 8.45 \\ 13.67 & 19.95 & 27.52 & 5.25 \end{bmatrix}, \qquad C = \begin{bmatrix} -1382.60 \\ -627.32 \end{bmatrix}.$$

The simulation results for the self-tuning PCC (SPCC) as compared with

EWMA and PCC is shown in Figure 5.4.

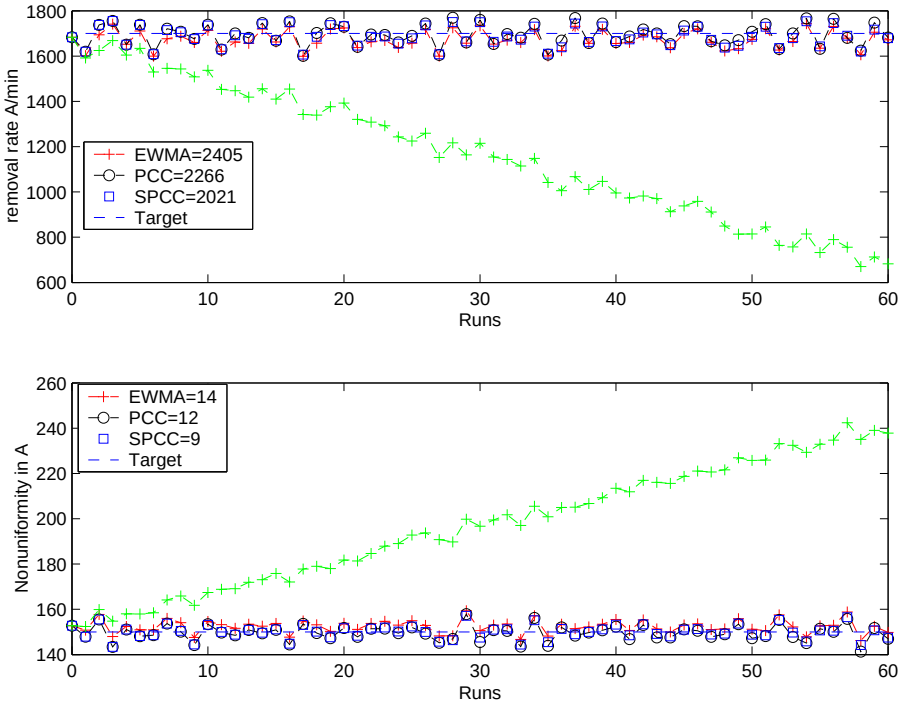

Figure 5.4. Comparison of MSE between the controllers for a linear perfect model under drift

From Table 5.3, it can be seen that the compensation effect of SPCC is much better than EWMA and PCC as measured by the MSE for the simulated number of runs. A 25% improvement in the within-wafer non-uniformity is achieved for the SPCC when compared to the PCC (i.e. MSE decreases from 12$\mathring{A}$ to 9$\mathring{A}$). The weights of ω_1 and ω_2 are 0.6 and 0.2 respectively.

But when there is some model error, which is common in real applications, the SPCC has faster convergence characteristic as shown in Figure 5.5, where the process model was taken as 80% of each parameters of the real process. SPCC is shown to track the target much faster than EWMA and PCC under this circumstance. Again, the MSE has decreased from 2266$\mathring{A}/min$ to 1967$\mathring{A}/min$ and 13$\mathring{A}$ to 11$\mathring{A}$ for the removal rate and within-wafer non-uniformity respectively between the PCC and SPCC controllers. The weights of ω_1 and ω_2 are 0.6 and 0.3 respectively.

MSE	Perfect model		Imperfect model		Under impulse disturbance		Nonlinear model	
	[1]RR	[2]NU	RR	NU	RR	NU	RR	NU
EWMA	2405	14	2357	18	13760	1564	2365	14
PCC	2266	12	2266	13	10946	1571	2197	12
SPCC	2021	9	1967	11	8760	1492	1985	9

[1]RR : Removal rate (Å/min)
[2]NU : Within-wafer Non-uniformity (Å)

Table 5.3. Comparison between EWMA, PCC and SPCC for CMP model under different conditions

Simulations are also performed for the case when there is a large impulse disturbance during the operation as shown in Figure 5.6.

This type of disturbance usually occurs when the pad is changed. The simulation result shows that PCC and SPCC is better than EWMA because EWMA is suitable only for process with slowly varying drifts. It is hard for it to compensate for large variance in several runs. The impulse disturbance in the simulation was experimented by changing the model parameters of the real process. The resulted shifting value equals to the target value plus 310 (Deng *et al.*, 1999).

In Figures 5.7 and 5.8, the variation of the weighting factors in the SPCC and SEWMA is shown. In order to compare the variation of weighting factors in SPCC and SEWMA controllers, we selected the same initial ω_1 value. Figure 5.7 shows that the two weighting factors in the SPCC controller decreased to a steady-state value. In the case of SEWMA, the weighting factor has increased, which would induce additional process noise.

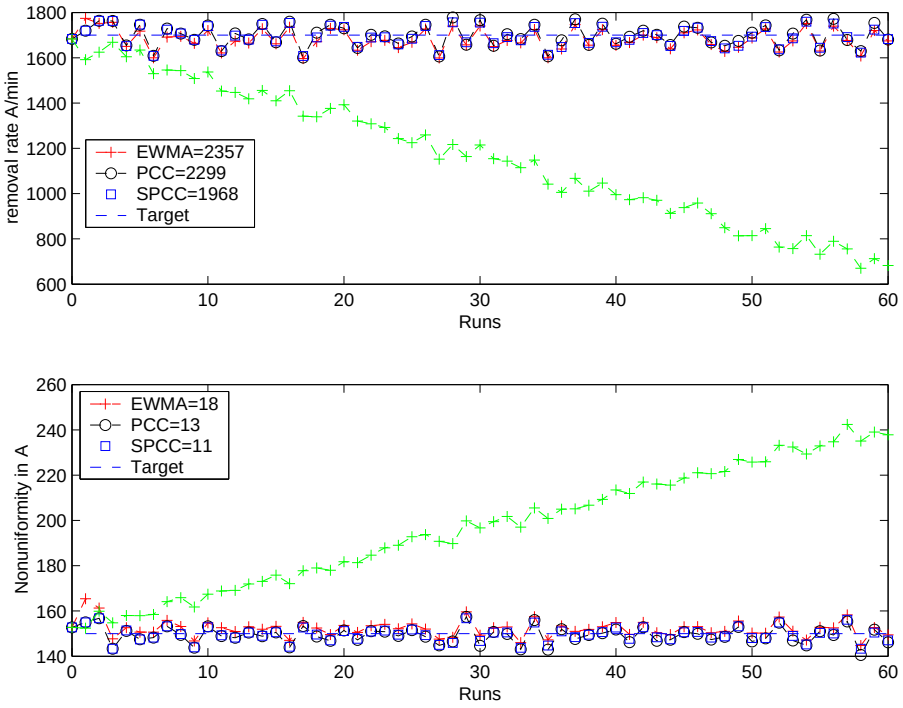

Figure 5.5. Comparison of MSE between the controllers for a imperfect model under drift

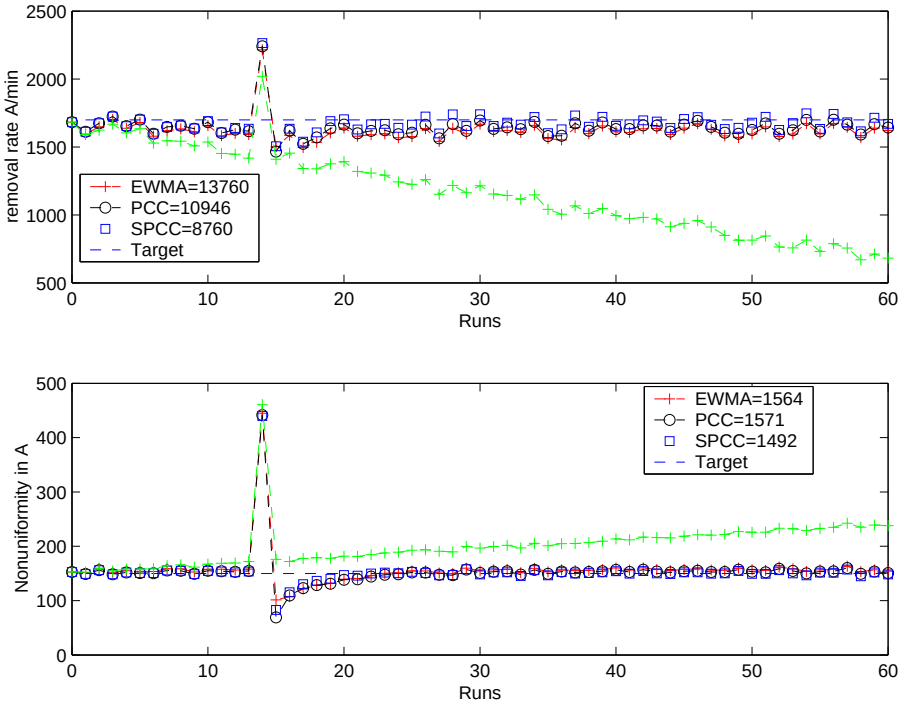

Figure 5.6. Comparison of MSE between the controllers for a impulse disturbance

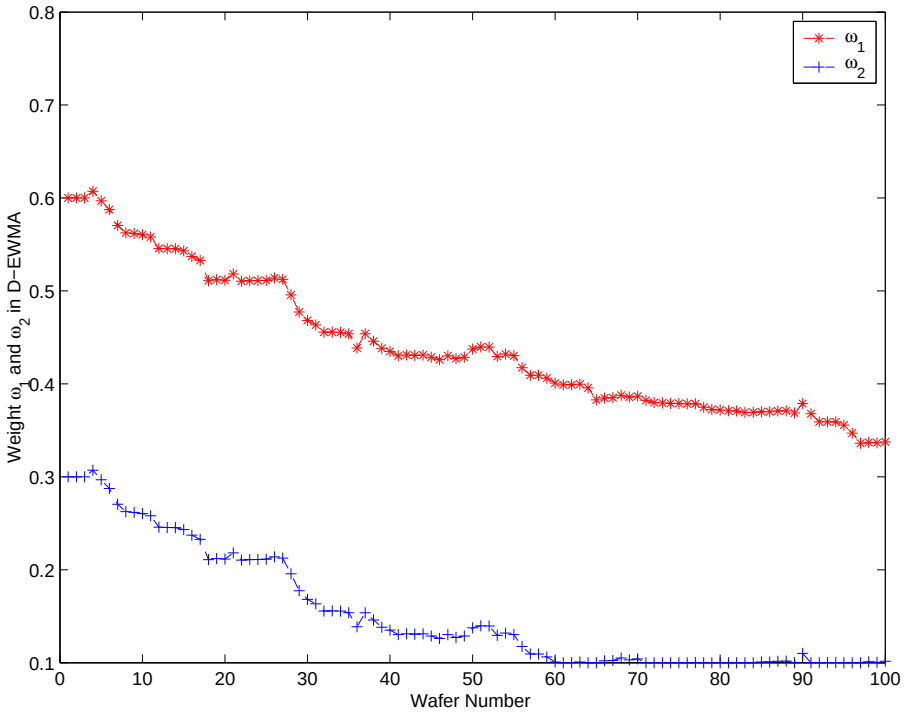

Figure 5.7. Weighting factors variation in SPCC controller

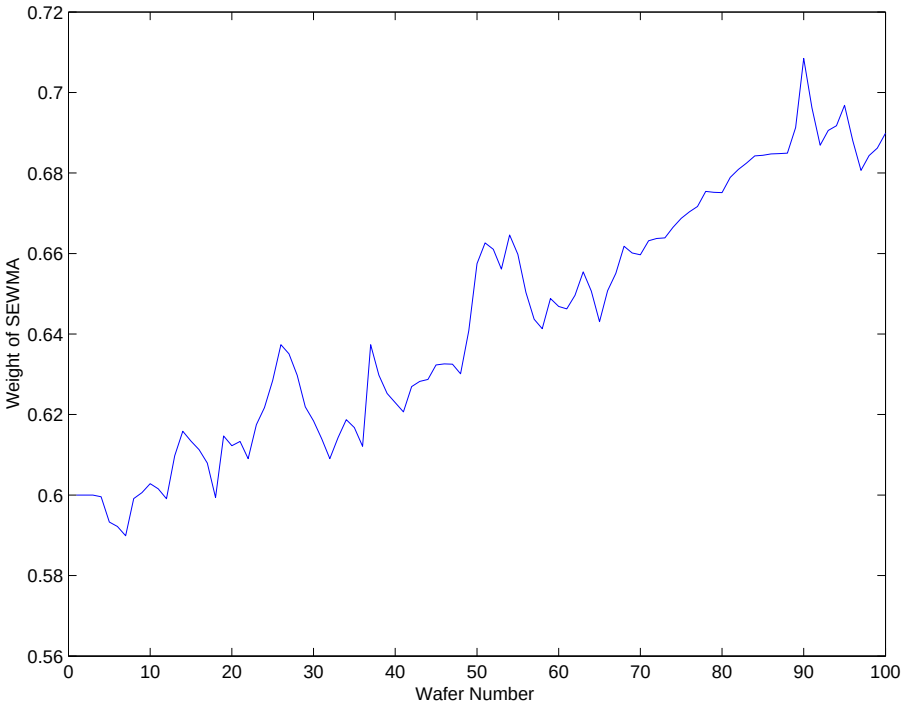

Figure 5.8. Weighting factors variation in self-tuning EWMA controller

5.4.2 Nonlinear process model

Simulation was done by assuming that the real process is non-linear and assuming a linear model for control purposes. The nonlinear model that is used is benchmark process model as given in (Chamness *et al.*, 2001), (Ning *et al.*, 1996). The real process model is given as

$$Y_k = C + f(X_k) + \varepsilon_k + \delta_k$$

where δ_k is a linear drift with constant drift speed $\delta = [-17\ 1.5]'$ and ε_k is a normally distributed white noise with mean zero and covariance

$$\Lambda = \begin{bmatrix} 665.64 & 0 \\ 0 & 5.29 \end{bmatrix}$$

and $f(u_k)$ is a full second-order polynomial function of the inputs with the following form:

$$f(u_k) = \sum_{i=0}^{3} \sum_{j=0}^{3} \beta_{i,j} u_i u_j$$

where

$$\beta = \begin{bmatrix} 1386.5 & 381.02 & -112.19 & 3778.8 & -21.301 & 8.7158 & 24.953 \\ 1520.8 & 2365.6 & 2923.5 & 281.66 & -3.9419 & -1.0754 & 1.406 \\ 37.082 & -17.642 & -11.974 & -164.99 & 28.150 & 249.17 & 0.025067 \\ 0.33797 & -72.274 & -94.222 & -26.175 & -13.505 & 36.691 & 32.929 \end{bmatrix}$$

The simulation results of the control effect of the algorithms are shown in Figure 5.9.

From Table 5.3, it can be inferred that SPCC provides a better performance even though the process is non-linear. We see an improvement in the MSE of 10%

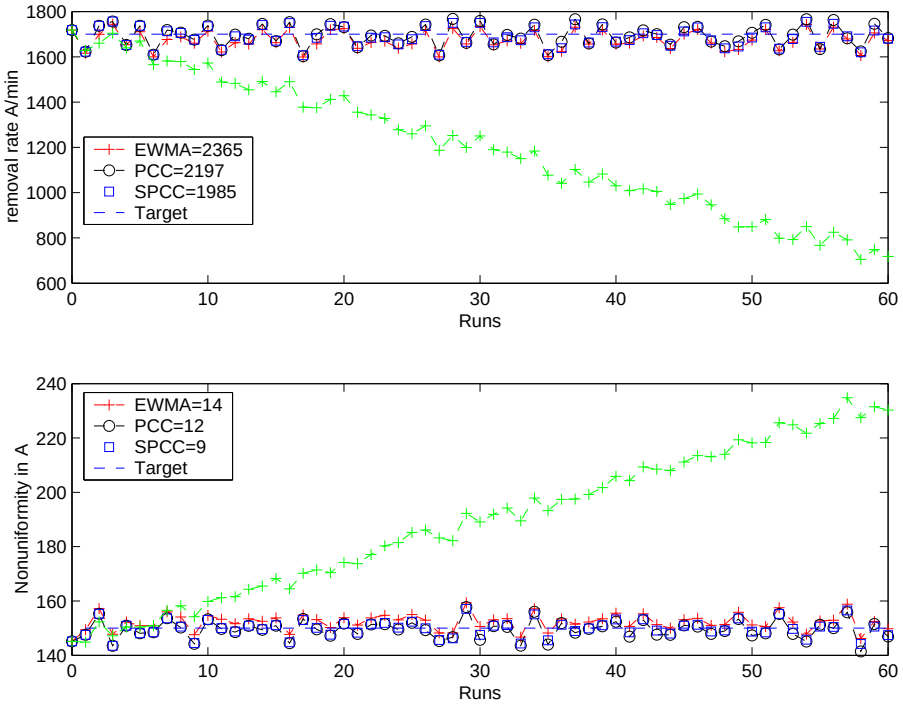

Figure 5.9. Comparison of MSE between the controllers for a non-linear model

in removal rate and 25% in within-wafer non-uniformity between the SPCC and PCC controllers.

5.5 Conclusions

A scheme for adaptively tuning the weights of the PCC controller is proposed in this chapter. This allows for automatically adjusting the forecasting parameters in the face of changing process noise and disturbances. Simulation results depicts an order of magnitude improvement in terms of the removal rate and non-uniformity when compared to conventional R2R controllers.

Chapter 6

Conclusions

6.1 Findings and conclusions

Lithography is the key technology in semiconductor manufacturing, because it is used repeatedly in a process sequence that depends on the device design. It determines the device critical dimensions, which affect not only the device's quality but also its product amount and manufacturing cost.To meet future technical challenges in microelectronics manufacturing especially in lithography process, it requires the Advanced Process Control (APC), namely a set of automated methodologies to achieve desired process goals on operating individual process steps. It is commonly considered to include 4 components named fault detection, fault classification, fault prognosis and process control. This paper reviews APC methodology in semiconductor processing, and covers the key unit operations of lithography and chemical-mechanical planarization. In this thesis, we mainly discuss about the wafer baking process in the lithography semiconductor manufacturing process and Chemical Mechanical Polishing (CMP) process which will affect the Depth of Focus (DOF). To improve the DOF, both the thickness variation issue in the baking process and Within-Wafer Non-Uniformity (WIWNU) in the CMP process are discussed into details. To solve the thickness variation, in another term, wafer

warpage, an automatic fault detection methodology is proposed and we have presented a physical model of the baking system and air-gap estimation will be done through experiments in the future. As for the Within-Wafer Non-Uniformity in the CMP process, a combination of Statistics Process Control (SPC) and APC, namely Run-to-Run Control, is investigated. A proposed auto-tuning Run to Run control strategy is further presented and discussed. CMP has been used in the global planarization of oxide and tungsten process and now is being used to provide unprecedented planarity of inter-layer dielectric silicon dioxide, copper and in lithography limited sub-micron trench isolation. It is projected that the observed effectiveness of the CMP process will lead to the widespread use of this process at various stages of integrated circuit (IC) fabrication, both in the front-end semiconductor manufacturing process and back-end semiconductor manufacturing process for a variety of high performance and application specific ICs, and for a variety of materials.

Due to the lack of in-situ measurements of surface thickness and the process environment buried in drift as well, run-to-run controller has proven higher potential to reduce the process variance in many discrete semiconductor manufacturing process. In this thesis, we only investigate the use of various run-to-run control scheme in CMP process, evaluating and analyzing their performance in a benchmark problem, chapter 3. All of these controller can largely reduce the within-wafer non-uniformity. Of these controllers, simple EWMA and PCC are the most used in semiconductor manufacturing while the optimum value of the weights of their algorithms is always difficult to select. A solution for automatically tuning the weights of the EWMA and PCC is developed in the chapter 4. We use a recursive algorithm to optimize the weighting factors of a PCC controller. The objective is to minimize the mean square error between the measured offset and the estimated offset using the variable step size LMS algorithm. The algorithm hinges on simple representation of the optimal double EWMA filter weighting factors. Simulation

results depict an order of magnitude improvement in terms of the removal rate and non-uniformity when compared to conventional R2R controllers. With the aid of the automatic fault detection in thickness variation and reduced WIWNU via run-to-run control in CMP process, the DOF requirement is met to the next-generation device manufacturing.

6.2 Suggestion for future work

6.2.1 Multi zone wafer warpage estimation

In this thesis, we have presented a physical model of the baking system, further the estimation of thickness variation will be done through experiments in the future. Therefore the wafer need not further processed or inspected and this technique will prove to be cost-effective and labour saving. It will be greatly helpful to improve the CD control due to the temperature variation throughout the process.

6.2.2 Integral control of different performance metrics in CMP process

Goals on various metrics such as removal rate, within-wafer non-uniformity, within-die non-uniformity and wafer-to-wafer non-uniformity must be addressed in the future for us to find an integral controller. This thesis has only taken the wafer to wafer and within wafer non-uniformity problem in CMP. As the critical dimension reduces much in the future, we have to pay more attention to the within-die non-uniformity. Variations in oxide thickness within a die can cause across die capacitance variations that can lead to timing problems in the device. An integrated controller for wafer to wafer, within wafer and within die variations is thus increasingly becoming important.

Author's Publications

List of publications

[1] Da Li, Varadarajan Ganesh Kumar, Arthur Tay, Abdullah Al Mamun, Weng Khuen Ho.Run-to-Run Process Control for Chemical and Mechanical Polishing in Semiconductor Manufacturing. *Proc. 17th IEEE International Symposium on Intelligent Control(ISIC2002), Oct. 2002,Vancouver, Canada*, pp740-745.

[2] Varadarajan Ganesh Kumar, Da Li, Arthur Tay, Abdullah Al Mamun, Weng Khuen Ho.Control of Chemical Mechanical Polishing in Microelectronic Manufacturing.*Proceedings of The 4th Asian Control Conference, Singapore, Sep. 2002*

List of submissions

[3] Arthur Tay, Abdullah Al Mamun, Varadarajan Ganesh Kumar,Da Li, Weng Khuen Ho. Control of Chemical and Mechanical Polishing in Microelectronic Manufacturing. *submitted to Asian Journal of Control*

Bibliography

Bibby, T. and K. Holland (1998). Endpoint detection for cmp. *Journal of Electronic Materials* **27**(10), 1073–1081.

Boning, D. S., A. Hurwitz, J. Moyne, W. Moyne, S. Shellman, T. Smith, J. Taylor and R. Telfyan (1996). Run by run control of chemical mechanical polishing. *IEEE Transactions on Components Packaging and Manufacturing Technology (C)* **19**(4), 307–314.

Boning, D. S., W. Moyne, T. Smith, J. Moyne and A. Hurwitz (1995). Practical issues in run by run process control. In: *IEEE/SEMI 1995 Advanced Semiconductor Manufacturing Conference and Workshop*. Cambridge,MA. pp. 201–208.

Boroujeny, B. F. (1998). *Adaptive filters: theory and applications*. Chichester. New York:Wiley.

Box, G. and M. Jenkins (1994). *Time Series Analysis-Forcasting and Control*. Prentice Hall. Englewood Cliffs, NJ.

Bulter, S. W. and J. Stefani (1994). Supervisory run-to-run control of polysilicon gate etch using in situ ellipsometry. *IEEE Transactions on Semiconductor Manufaturing* **7**(2), 193–201.

Campbell, S.A. (1996). *The Science and Engineering of Microelectrtonic Fabrication*. Oxford Univ. Press. London, U.K.

Campbell, W. J. (1999). Model predictive Run-to-Run Control of Chemical Mechanical Planarization. PhD thesis. University of Texas at Austin.

Campbell, W. J. and A. J. Toprac (2001). A survey of run to run control algorithms. In: *AEC/APC Symposium XIII.*

Castillo, E. D. (2002). *Statistical Process Adjustments for Quality Control*. Probability and Statistics. Wiley. NewYork.

Castillo, E. D. and A. Hurwitz (1997). Run-to-run process control: Literature review and extensions. *Journal of Quality Technology* **29**(2), 184–196.

Castillo, E. D. and J. Y. Yeh (198). An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes. *IEEE Transactions on Semiconductor Manufaturing* **11**(2), 285–295.

Chamness, K., G. Cherry, R. Good and J. Qin (2001). A comparison of r2r control algorithms for the cmp with measurements delays. In: *AEC/APC Symposium XIII.*. pp. 1–4.

Chen, A. and R. S. Guo (2001). Age-based double ewma controller and its application to cmp process. *IEEE Transactions on Semiconductor Manufaturing* **14**(1), 11–19.

Cook, L. M. (1990). Chemical processes in glass polishing. *J. Non-Crystalline Solids* **120**, 152–171.

Crisalle, O., C. Bickerstaff, D. Seborg and D. Mellichamp (1998). Imporvement in photolithography performance by controlled baking. *Proc. of SPIE* **921**, 317–325.

Deng, H., C. Zhang and J. S. Baras (1999). Run-to-run control methods on the dhobe algorithm. Technical Research Report ISR T.R.99-65. University of Maryland.

Fukui, T., H. Kurita and N. Makino (1997). Warpage of inp wafers. In: *International Conference on Indium Phosphide and Related Materials*. pp. 272–275.

Hamby, E. S., P. T. Kabamba and P. P. K. gonekar (1998). A probabilistic approach to rull-to-rull control. *IEEE Transactions on Semiconductor Manufaturing* **11**(4), 654–669.

Hendrix, M., S. Drews and T. Hurd (2000). Advantages of wet chemical spin-processing for wafer thinning and packaging applications. In: *26th IEEE/CPMT International Electronics Manufacturing Technology Symposium*. pp. 229–236.

Ho, W.K., A Tay and C. Schaper (2000). Optimal predictive contorl with constraints for processin gof semiconductor wafers on large thermal-mass heating plates. *IEEE Transactions on Semiconductor Manufaturing* **13**, 88–96.

I.S., S.H. Ko, K.I. Suh, J.H. Kim, Y.S. Kim, D.D. Lee, S.I. Kim and D.J. Ahn (1999). Warpage effect on breakdown voltage of dram device. In: *6th International Conference on VLSI and CAD*. pp. 437–440.

Kim, D., S. Kim, Y. Lee, S. Kim and K. Suh (1999). Study of micro-scratch on oxide film in vlsi circuit. In: *Proc. of 1999 VLSI-MIC*. pp. 283–287.

Kishino, S., H. Yoshida and H. Niu (1993). Optimizing gettering conditions for vlsi chips using simple yield model. *IEEE Transactions on Semiconductor Manufaturing* **6**(3), 251–257.

Mohondro, R. and R. Gaboury (1993). Charaterizing coat, bake, and develop process. *Solid State Technology* **7**, 87–90.

Montgomery, D. C. (1996). *Design and Analysis of Experiments-4th ed.*. John Wiley and Sons, Inc. New York:Wiley.

Mullins, J. A., W. J. Campbell and A. D. Stock (1997). An evaluation of model predictive control in run to run control processing in semiconductor manufacturing. In: *process,equipment,and materials control in integrated circuit manufacturing III. SPIE*. A.Ghanbari & A. Toprac. pp. 182–189.

Muske, K. R. and J. B. Rawlings (1993). Model predictive control with linear models. *AIChE Journal* **39**(2), 262–287.

Ning, Z., J. R. Moyne, T. Smith, D. Boning, E. D. Castillo, J. Y. Yeh and A. Hurwitz (1996). A comparative analysis of run-to-run control algorithms in the semiconductor manufactring industry. In: *IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop*. pp. 375–381.

Ozisik, M. N. (1985). *Heat Transfer - A Basic Approach*. McGraw-Hill International.

Preston, F. W. (1927). The theory and design of plate glass polishing machines. *J. Soc. Glass Technol.* **11**, 214–256.

Quirk, M. and J. Serda (2001). *Semiconductor Manufacturing Technology*. Prentice-Hall. Englewood Cliffs, NJ.

Raznjevic, K. (1976). *Handbook of Thermodynamic Tables and Charts*. Hemisphere Publishing Corporation.

Sachs, E., A. Hu and A. Ingolfsson (1995). Run by run process control: Combining spc and feedback control. *IEEE Transactions on Semiconductor Manufaturing* **8**(1), 26–43.

Schaper, C., M. Moslehi, K. Saraswat and T. Kailath (1994). Modelling, identification and control of rapid thermal processing systems. *J.Electrochemical Soc.* **141**(11), 3200–3209.

Sheats, J.R. and B.W.E. Smith (1998). Microlithography science and technology. Technical research report. Marcel Dekker Inc.

Smith, T. and D. Boning (1996). A self-tuning ewma controller utilizing artificial neural network function approximation techniques. *International Electronics Manufacturing Symposium* pp. 355–364.

Smith, T. and D. Boning (1999). A study of within-wafer non-uniformity metrics. In: *4th Intl. Workshop on Statistical Metrology*. Kyoto, Japan.

Steigerwald, J. M., S. P. Murarka and R. J. Gutmann (1997). *Chemical and Mechanical Planarization of Microelectronic Materials*. A wiley-interscience publication. John Wiley and Sons, Inc. NewYork.

Sturtevant, J., S. Holms, T. Vankessel, P. Hobbs, J. Shaw and R. Jackson (1993). Post-exposure bake as a process-control parameter for chmically-amplified photoresists. *SPIE Integrated Circuit Metrology, Inspection and Process Control* **1926**, 106–114.

Thompson, L.F., C.G. Willson and M.J. Bowen (1994). *Introduction to Microlithography*. Prentice-Hall.

Tung, T. L. (1997). A method for die scale simulation for cmp planarization. In: *Pro. of SISPAD Conf.*. Cambridge, MA.

Warnock, J. (1991). A two-dimensional process model for chemi-mechanical polish planarization. *Journal of the Electrochemical Society* **138**(8), 2398–2402.

Wei, S., S. Wu, I. Kao and F.P. Chiang (1998). Measurement of wafer surface using shadow moire technique with talbot effect. *Journal of Electronic Packaging* **120**(2), 166–170.