Variation Reduction in Plasma Etching via Run-to-Run Process Control and Endpoint Detection

by

Minh Sy Le

Submitted to the Department of Electrical Engineering
and Computer Science

in partial fulfillment of the requirements for the degree of

Masters of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

July 11, 1997

Author .....
Department of Electrical Engineering and Computer Science
July 11, 1997

Certified by .....
Herbert H. Sawin, Professor
Department of Chemical Engineering
Thesis Supervisor

Certified by .....
Duane S. Boning, Associate Professor
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Arthur C. Smith, Chair
Department Committee on Graduate Students

Variation Reduction in Plasma Etching via Run-to-Run Process Control and Endpoint Detection

by

Minh Sy Le

Submitted to the Department of Electrical Engineering and Computer Science on July 11, 1997, in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science

Abstract

Variation in a plasma etch process is shown to be reduced through design of processing equipment for effective control. A dual coil transformer coupled plasma etcher was integrated with multiple sensor systems. Full Wafer Interferometry and optical emission spectroscopy were used for process monitoring. This information is integrated to a run-to-run model based process controller which suggests recipe modifications to reduce variation in the product. Two separate control experiments demonstrate that design of equipment for control can improve processing performance.

A new method to detect endpoint in plasma etch processes is also presented. A multivariate statistical technique that utilizes continuous broadband optical emission spectra from the plasma is employed. This technique is found to be sensitive to slight changes in the plasma emission which we believe signify endpoint. A method to make the technique both sensitive and robust is also discussed.

Herbert H. Sawin
Title: Professor, Chemical Engineering and EECS

Thesis Supervisor: Duane S. Boning
Title: Associate Professor, EECS

Acknowledgments

As I prepare to wrap up my work at MIT I found time to reflect on what my education represented. I have learned much during these past six years but am sure that my life's work is not here in Cambridge, but outside the corridors of MIT. I once read somewhere that the most important experience gained from an MIT education is to learn how to think. I truly believe in this statement. Wherever I should go, whatever task is presented to me, whatever challenge that I face, I will take that with me. MIT is not a self study course. I have learned much from my professors, colleagues, and friends and hope that I too have enriched their experiences here. Professor Herb Sawin has been my thesis advisor and mentor for two thesis thus far and deserves special recognition for supporting me and giving me the freedom to discover and the guidance to not stray too far. Professor Duane Boning provided me with the necessary vision and framework. Someone once told me that leading from behind is the sign of a true leader. I am grateful to have had two advisors that had the foresight to allow me the flexibility and resources to do the work I set out to accomplish.

Numerous other professors here at MIT and mentors in industry also have contributed to my professional development. Tim Dalton at Digital provided the mentoring and facilities that were instrumental in the last parts of this thesis. Stepahnie Butler and her colleagues at Texas Instruments were very supportive of my work and encouraged me to continue. Professor Sachs inspired me to view invention not as flashes of inspiration but rather an organized effort to provide solutions to previously unmet problems. This I will put away and draw upon when the time comes for me to inspire other people.

And to my professors at the Sloan School of Management who taught me that management is both an art and a science.

My colleagues in lab were not just other students working on their own research but friends who worked together in mutual support. I thank you for the times that we spent together in and out of the labs. Thanks to Arpan, Taber, Aaron, Jane, Sandeep, Han, YP, Scott, Brett, Heeyeop, and Brian. My gratitude to Aaron, Sandeep, and Han for helping me learn to ski and taking me down double black diamonds.

My friends and colleagues have taught me much and provided much appreciated support during these past six years. Dave, my housemate of two years was a great foil and source of digression. And to Dave's colleagues in lab for their infinite wisdom and humor: Dawn and Patricia. And to Scott, my roommate from back in my undergraduate days, who always knew when to get together and talk.

Finally I must thank those who provided the financial support that allows me to complete this thesis: The Bose Foundation, NDSEG Fellowship program, and of course my family.

Thank you MIT.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction - Variation in Plasma Etch

1.1 Problem Statement

In semiconductor manufacturing, as in any manufacturing process, the goal is not to just make one part that meets specifications, but to make many. Variation both within those parts and between parts reduces the quality and can impact the yield of the process. Though processed by the same manufacturing system, all the products will not be identical but rather can be represented by a quality distribution. The mean of this distribution may not be on target and can have a wide standard deviation. A more ideal case would be for the distribution of parts from the manufacturing system to be well centered on target and have a very small standard deviation. Feedback control can help achieve the more ideal target in many cases by correcting for systematic disturbances.

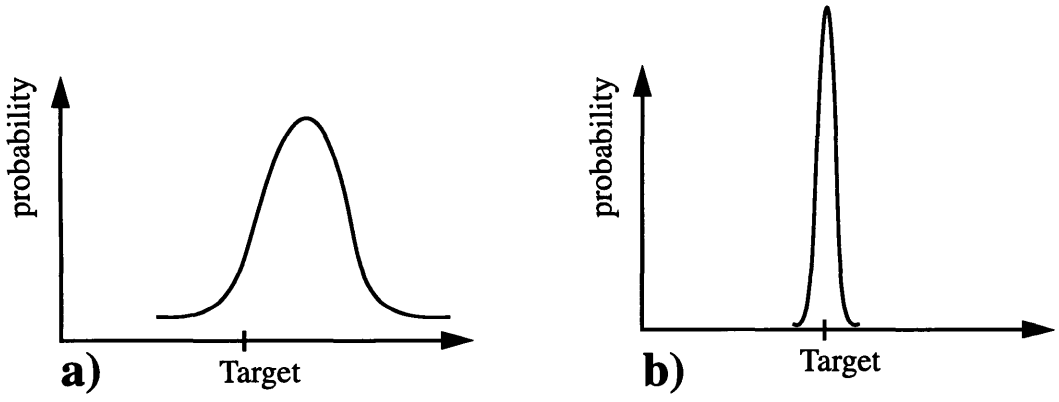

Figure 1.1: The quality of parts produced by a manufacturing system can be represented by a distribution. a) The distribution is wide and the mean is not on target. b) The distribution is narrow and on target.

Uniformity is an important quality metric in semiconductor manufacturing. Uniformity, or rather, nonuniformity exists on many different levels. There needs to be control of the quality of a product on a batch to batch basis, usually on the order of 25 wafers. On a

run to run basis, the wafers need to be processed similarly to produce similar results. Different dies within a wafer need to exhibit similar performance so within wafer uniformity is of concern. Finally, devices in different regions within a die need to have similar performance for the chip to perform properly.

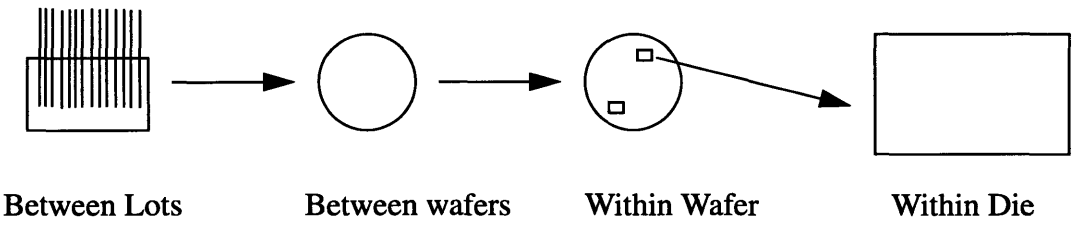

Figure 1.2: Uniformity needs to be accomplished over many scales

This work makes two key contributions to improve process uniformity on all of those levels. First is a demonstration of a dual-coil approach in plasma equipment design for improved process control and process robustness. The second contribution is an approach to detect endpoint in difficult low open area plasma etch processes.

1.2 Equipment design for process control and robustness

Robust design in semiconductor manufacturing refers to design of a process and process equipment that is robust to potential disturbances and hence improves process uniformity. In this work, robust design of process equipment takes on added meaning. The process equipment is modified and made more amenable to process control. Process control in this work will refer to feedback of information to enact process recipe changes to ensure that the product is on target. This is very different from statistical process control (SPC) that is commonly used in the semiconductor industry, which is a technique for process monitoring rather than process control. Run-to-run process control represents a significant departure from the assumption that a particular process tool is static and optimal.

This work will show the implementation of these ideas on a plasma etcher. For better control of within wafer etching uniformity, a transformer coupled plasma etcher has been modified for dual coil operation. A Full Wafer Interferometry (FWI) system provides in-situ measurements of the etch rate at many locations across the wafer. Optical Emission Spectroscopy (OES) provides information about the plasma state. Measurements from these sensors are interpreted by a run-to-run model-based process controller to suggest recipe adjustments that will keep the process specifications on target.

1.3 Plasma etch endpoint detection

A different problem exists in low open area endpoint detection. Another source of variation in plasma etch processes arises from inadequate or nonexistent detection of endpoint. Endpoint occurs when the film that is being etched is completely removed. Variation in input material thickness and process variations give rise to overetch and device damage. Existing methodologies include RF sensors that monitor the change in plasma impedance and monitoring the change in a one or two optical emission lines [Dalton]. The RF sensor can be useful if the plasma impedance changes significantly at endpoint but most modern RF power systems use an automatic matching network which make analysis of the signals difficult in all but the simplest of cases. The signal to noise ratio in monitoring one or two optical emission lines is very poor and the endpoint detection algorithms used in these cases often rely on neural network, curve fitting, or feature detection algorithms [Allen]. Many industrial etching processes are not based on any of these sensors since the signal to noise ratio is so poor and reliability is a concern. Instead, a timed etch process with an overetch ensures that the etch is complete but at the cost of variation.

This thesis will investigate the use of a wide optical emission spectrum collected at many times during the etch process to detect endpoint. Use of many optical channels (1000 or more) can improve the signal to noise ratio. Mathematical algorithms including

Hotelling's T^2 and Principle Component Analysis are applied to the collected spectra in efforts to determine endpoint. These techniques take advantage of the correlation in many spectral lines and can amplify the signal to noise ratio. The models will be made robust through use of Exponentially Weighted Moving Average and Exponentially Weighted Moving Covariance updates of spectral models of the normal plasma state.

1.4 Organization

The first part of this thesis will develop the dual coil transformer coupled plasma (TCP) etcher in the context of robust process design. Sensors that enable feedback control on the dual-coil TCP etcher are then examined in Chapter 3. FWI and OES are presented as well as the analytical tools used to interpret the data. Feedback process control will be discussed in Chapter 4. An Artificial Neural Network Exponentially Weighted Moving Average (ANN-EWMA) controller developed by Smith is examined [Smith, 1997]. Results from two separate control experiments are then discussed. First, an experiment that utilizes information from just the FWI sensor is presented in Chapter 5. Next, a multi-objective control system that uses both FWI and OES sensor information is presented in Chapter 6.

In the final part of this thesis, a new method for endpoint detection is developed and outlined in Chapter 7. The mathematical basis for process monitoring and detection based on Hotelling's T^2 statistic is discussed. The EWMA and EWMC methodology to allow the method to be robust yet sensitive is also presented. Test cases are demonstrated, first with blanket polysilicon etch and then with low open area oxide patterned wafers from a production line.

Chapter 2

The Transformer Coupled Plasma Etcher

In this chapter, we discuss equipment modifications made to a Transformer Coupled PLasma (TCP) etch tool to improve process controllability and robustness. In Section 2.1, a review of difficulties encountered in single coil TCP reactors. An approach to overcome these limitations based on axiomatic design is presented in Section 2.2. Finally, Section 2.3 presents details of the resulting dual coil TCP etcher design.

2.1 Single Coil TCP

TCP reactors have evolved to be the current dominant design in high density plasma systems. While parallel plate RF diode systems still serve a purpose in many processes, the more advanced requirements of anisotropic etching and high etch rate can more easily be met with high density plasma systems [Keller]. Most of the major semiconductor equipment makers including Applied Materials and Lam Research make variants of the TCP.

The primary method of power coupling in a TCP reactor is inductive ohmic heating. A flat spiral coil serves as the primary winding in a transformer. Radio frequency (RF) current through the coil produces a changing magnetic field which induces a curling electric field in the reactor according to Faraday's Law.

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{2.1}$$

This electric field accelerates electrons which can then transfer energy to other species in the plasma.

A side effect of inductive heating is capacitive power coupling from the coil to the plasma. Since the coil antenna is an inductor, there is a voltage drop across it when alternating current passes through. The voltage is not constant along the length of the inductor.

The peak to peak voltage is highest closest to the power supply and goes to zero at ground. The coil acts like a capacitor with different voltages along its length. As in Figure 2.1, if the power supply is connected to the outer tap of the coil antenna, more capacitive power coupling is exhibited towards the outer periphery of the plasma. Capacitive coupling heats the plasma nonuniformly. Furthermore, Ventzek et al. has shown that capacitive coupling induces nonuniform oscillations in the plasma potential [Ventzek]. There are methods to reduce capacitive coupling in TCPs. A Faraday shield was tested but it was discovered that plasma initiation is more difficult. Capacitive power coupling, though inherently non-uniform in TCPs, aids in striking a plasma.

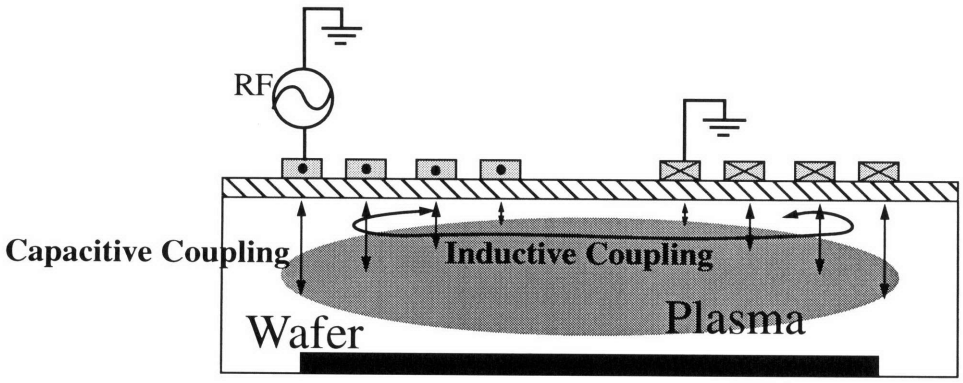

Figure 2.1: Power is coupled into the plasma both inductively and capacitively. Capacitive coupling is nonuniform since the voltage drop across the inductor is non-uniform. Inductive coupling is localized because of positive feedback. The plasma conductivity is increased with increased power.

Inductive power coupling is also not uniform across the plasma radius. The coil antenna induces a current in the plasma. Electrons are much more mobile than ions so the plasma current is carried primarily by the electrons, which through ohmic losses, heat the plasma and transfer energy to ions and neutral species. As more power is coupled to the plasma, the plasma density increases. This density increase is matched by an increase in the plasma conductivity since there are more charged species to carry current, and the species are more mobile. Current induced in the plasma will favor annular regions of high

conductivity. This leads to a situation where most of the inductive power is coupled to a donut shaped "hot" zone, as illustrated in Figure 2.2. Energy then diffuses to the rest of the plasma by collisions [Ventzek, 1994].

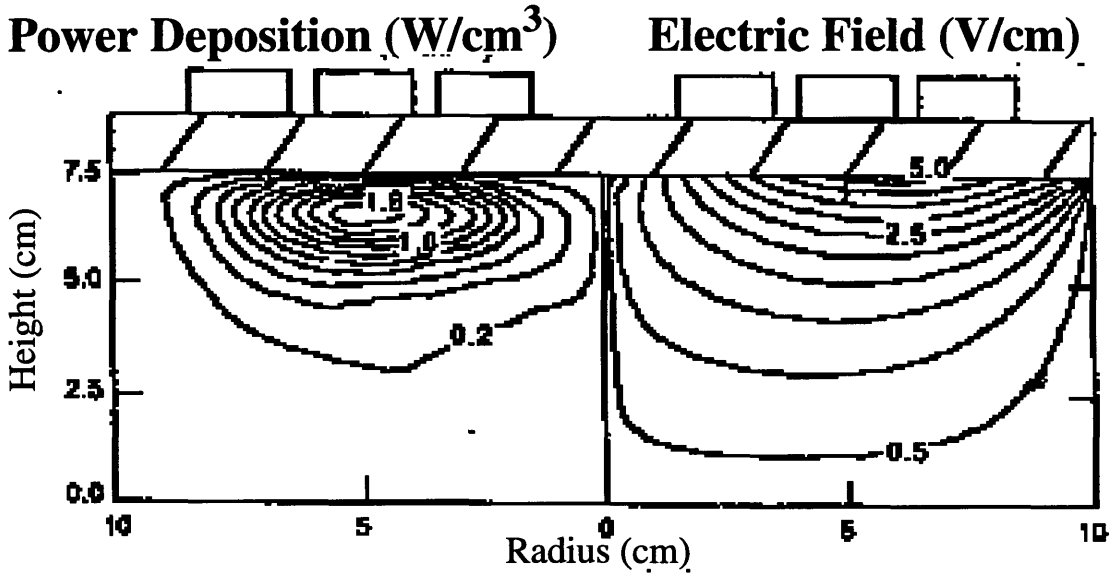

Figure 2.2: The power deposited in a TCP is localized around the spiral inductive coil antenna. A toroidal power deposition region can be seen in the above diagram [Ventzek, 1994].

2.2 Axiomatic Design

Axiomatics, as developed by Suh, presents a way of evaluating manufacturing processes [Suh]. Design Parameters (DPs) are the technical specifications for the process and Physical Variables (PVs) are the process parameters. The DPs can be written as a linear combination of the PVs. The rate and quality of a process is a function of the process parameters. In general, the following matrix notation shows the functional relationships between the process settings and the output of the process. It is not the intention of this equation to provide a linear model for a general process, but to illustrate how strongly a design parameter is affected by physical variables.

$$\begin{bmatrix} DP_1 \\ DP_2 \\ \cdots \\ DP_n \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix} \begin{bmatrix} PV_1 \\ PV_2 \\ \cdots \\ PV_n \end{bmatrix} \tag{2.2}$$

The coefficients C_{ij} are not represented by actual values but rather by X or O. O's show that there is a weak or no dependence of a Design Parameter (DP) on the Physical Variable (PV) whereas X's show a strong dependence. For a completely uncoupled process, the coefficient matrix is full of O's except for X's on the diagonal. Each Design Parameter can be satisfied by one and only one Physical Variable. A DP can be achieved without affecting other DP's. Very few systems are uncoupled. A more common case is when the matrix is very dense or fully dense. This is representative of a coupled process. A decoupled process is where the matrix is upper triangular. In this case, the process is neither coupled nor decoupled. It is possible to achieve a desired set of DP's through judicious choice of PVs. The progression of designs away from coupled processes to decoupled processes can be illustrated the following semiconductor equipment design examples.

2.2.1 Coupled Processes

Engineering is a compromise. This can be seen in manufacturing systems where there is usually a trade-off between rate and quality. In plasma etching this compromise is clear in Reactive Ion Etching (RIE) systems based on a parallel plate capacitive reactor. The density and energy of the ions striking the surface are very important factors that determine the etch rate and etch quality as represented by a number of metrics including etch selectivity and ion induced damage. RIE systems behave like diodes, as illustrated in Figure 2.3. The higher the peak to peak voltage supplied the higher the ion current. Higher current leads to higher etch rates but higher peak to peak voltages also result in reduced etch selectivity and more damage since the higher energy ions will physically sputter

material away, regardless of composition. Throughput of the RIE systems is thus strongly coupled to the quality of the process [Chapman].

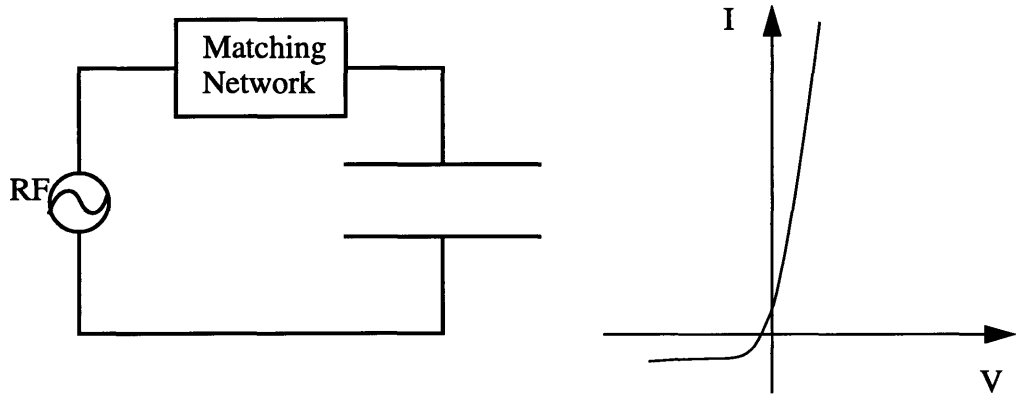

Figure 2.3: RIE plasmas exhibit a diode behavior. The ion current and ion energy (rms sheath potential) are coupled. The rate of the etch process cannot be increased without significant loss in quality.

Coupled processes such as the RIE system present difficulties in achieving all of the design parameters simultaneously.

$$\begin{bmatrix} EtchRate \\ Selectivity \end{bmatrix} = \begin{bmatrix} X & X \\ X & X \end{bmatrix} \begin{bmatrix} Power \\ Pressure \end{bmatrix} \tag{2.3}$$

The above matrix relationship shows that the RF power delivered to the etcher and the pressure in the processing chamber both have strong causal relationships on not only the etch rate (throughput) but also etch selectivity (quality). This is a simplifying approximation to the capacitively coupled RIE plasma etcher but it illustrates the interdependence of etch rate and quality.

High density plasma etchers have a number of advantages over the RF diode low density plasma systems. Of significant importance is that the ion energy (voltage) is not

strongly coupled to the ion current so that the etch rate can be increased without significant increases in device damage and reduction in etch selectivity.

$$\begin{bmatrix} EtchRate \\ Selectivity \\ Damage \end{bmatrix} = \begin{bmatrix} X & X & X \\ O & X & X \\ O & O & X \end{bmatrix} \begin{bmatrix} Pressure \\ TopPower \\ BottomPower \end{bmatrix} \tag{2.4}$$

In the framework of axiomatics, this process is considered decoupled [Suh]. It should be possible to obtain a desired combination of rate and quality through judicious choice of the process variables. The Transformer Coupled Plasma etcher, according to these design parameters, is much better at achieving desired throughput and quality than the capacitively coupled RIE system. However, when within-wafer-nonuniformity (WIWNU) is also considered as a design parameter, the matrix becomes underdetermined and coupled. As discussed in Section 2.1, there is no direct way to tailor the etch rate uniformity across the wafer without also impacting other design parameters with the single coil TCP.

2.2.2 Decoupled Processes

Rapid Thermal Processing (RTP) is another example of a manufacturing processes that can be decoupled by improved equipment design. RTP is a technology that has supplanted tube furnace annealing for certain process needs (e.g. very shallow junction formation). The advantages of RTP over tube furnace processing are analogous to those arising in the replacement of RIE systems with high density inductively coupled plasma systems. RTP offers reduction in thermal ramp times and processing times over tube furnace processes. The rate can be significantly increased. However, in early RTP systems with only a single heating element, thermal nonuniformities were common. This arose because of pattern density and chamber effects which result in nonuniform emmissivities in the processing system. These problems were eventually solved by using multiple heating zones and controlling all of the heating elements independently so that the wafer is

heated uniformly. By adding the extra degrees of freedom in the physical process variable space, the RTP process was decoupled and easier to control [Schaper].

2.3 The Dual Coil Transformer Coupled Plasma Etcher

The dual coil TCP incorporates an important idea drawn from the Rapid Thermal Processing experience. Multiple heating zones allow for greater robustness and can improve both within wafer and between wafer processing uniformity. In the modified TCP, two inductive antennae couple power to the plasma, one of small radius and the other of larger radius. These two concentric planar coils allow for control of etching rate profile across the radius of the wafer as illustrated in Figure 2.4.

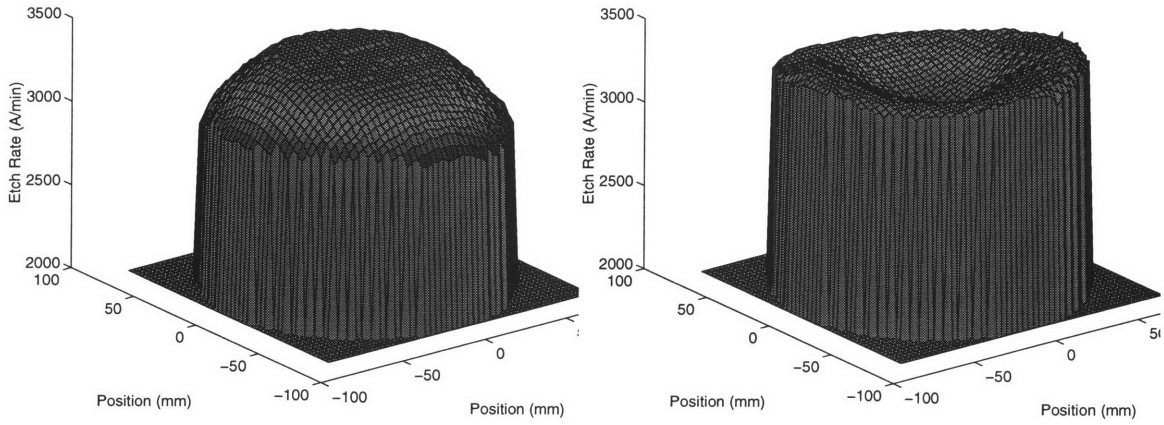

Figure 2.4: The etch rate profile can be tailored with the dual coil TCP. a) At a particular power setting, the etch rate towards the center of the wafer is higher towards the edge. b) At another power setting, the etch rate at the center is lower.

The addition of the extra coil makes the system more robust, though it is still not decoupled since there are still intermixing relationships between the DPs and PVs.

$$\begin{bmatrix} EtchRate \\ Selectivity \\ WIWNU \\ Damage \end{bmatrix} = \begin{bmatrix} X & X & X & X \\ O & X & X & X \\ O & X & X & O \\ O & O & O & X \end{bmatrix} \begin{bmatrix} Pressure \\ InnerPower \\ OuterPower \\ Bottom \end{bmatrix} \tag{2.5}$$

The idea of using multiple heating zones in plasma processes is not new. Ventzek has presented simulations of this and similar ideas for TCPs [Ventzek, 1995]. Of scientific

interest is the work dealing with more than two antenna and phasing the RF generators and frequencies to provide a beat effect that can sweep the plasma across the substrate [Yamada, 1996]. Other works have also applied the idea of multiple heating zones to microwave plasmas [Yasaka]. This thesis will show that the ideas illustrated in those simulation works can be implemented in a research grade TCP, that such equipment modifications can be effectively used in run-to-run control strategies. Specific details of our dual coil TCP implementation are described next.

2.3.1 Antennae

The planar spiral antennae are wound from 1/4" copper tubing. They each have three turns. The larger coil has an outer diameter of 9 inches while the inner antennae has a diameter of 5 inches. A diagram of the coils is shown in Figure 2.5. A key issue that this dual coil system addresses is that process uniformity is a complex function of the process recipe. Through judicious choice of power settings, the process uniformity can be optimized for different process gases and other recipe parameters. Single coil TCPs lack this very important ability to optimize the etching uniformity for different gas compositions and etch applications. Furthermore, the power settings can be adjusted on a run-to-run basis to compensate for any changes in uniformity.

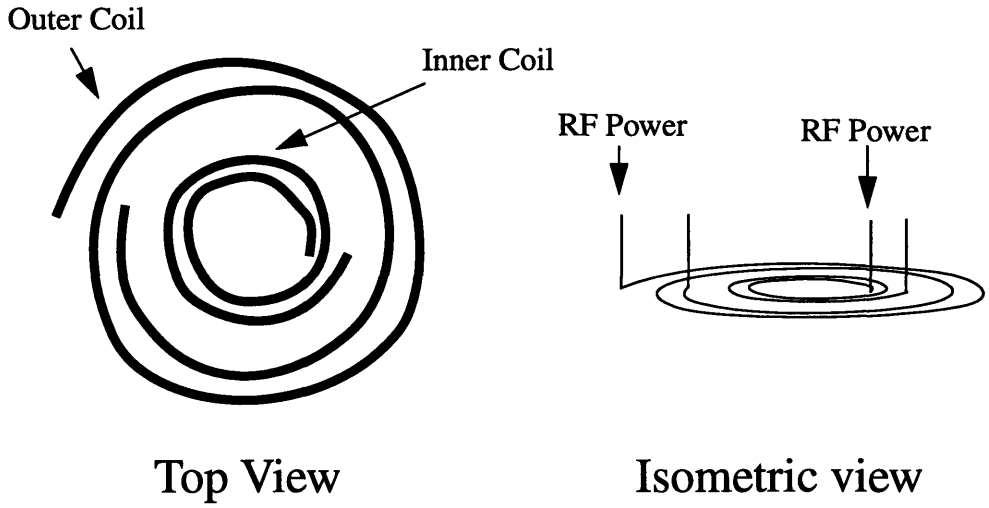

Figure 2.5: A representative diagram of copper coil antenna configuration.

2.3.2 Matching Network

The matching network allows efficient transfer of power from the generator to the plasma. The matching network is an L match AM-10 automatch from RF Power Products. The antennae coil for each of the coils in the dual coil system has an independent matching network and power supply. The coil is capacitively coupled to ground by a 50 pF ceramic capacitor to adjust the impedance of the coil system. This allows the L match to appear like a Pi match so it can properly match the impedance of the coil system to the 50 Ohm generator.

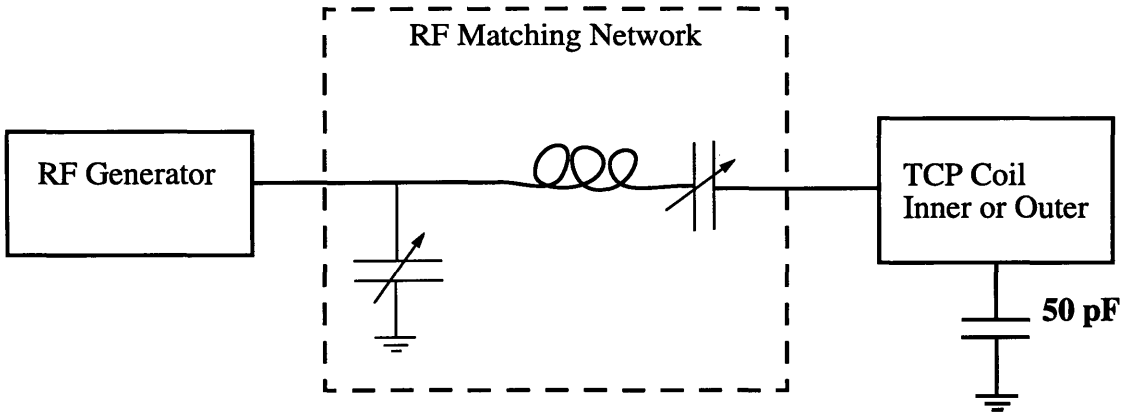

Figure 2.6: Matching network diagram

2.3.3 Generators

Two radio frequency generators are used to supply power to the inner and outer coils. The Advanced Energy 13.56 MHz generator has a maximum power output of 3 kW. The Seren IPS 11.00 MHz custom generator can supply up to 600 Watts. The generators are connected to the matching networks with type N RF cables. The matching networks are connected to the coils with high current type C RF cables.

Chapter 3

Sensors

Modifications made to the TCP allow for spatial control of the etch rate. However, without sensors to estimate the state of the process, intelligent control decisions cannot be made. Two sensor systems are used in this work. Full Wafer Interferometry and spatially resolved optical emission spectroscopy are both integrated to the dual coil TCP and provide in-situ measurement of the wafer state and plasma state respectively.

3.1 Full Wafer Interferometry

Full Wafer Interferometry provides measurements of the etch rate at multiple locations across the wafer during the etch process. This diagnostic sensor is manufactured by Low Entropy Systems (Brookline, MA). A CCD array records plasma induced emission through a narrow bandpass filter. The wavelength selected for this experiment was 486 nm. Figure 3.1 illustrated the integration of the FWI sensor system on the dual coil TCP. Other works have investigated FWI as a technique for robust endpoint detection and etch rate measurements [Dalton, 1994; Wong]. A review of the physics of interferometry arising in the etch of polysilicon on oxide over silicon is provided next.

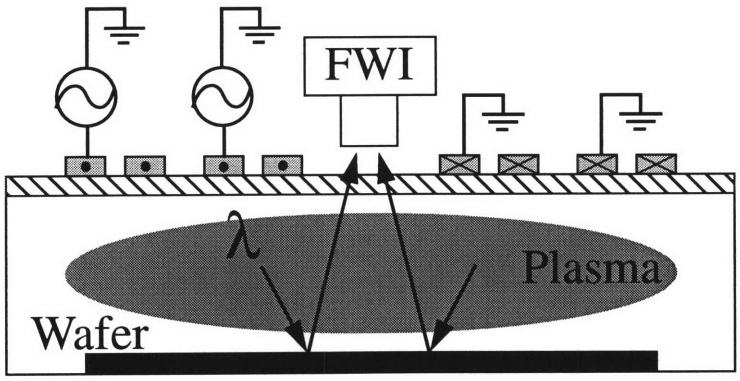

Figure 3.1: The Low Entropy Systems Full Wafer Interferometry system is installed directly above the TCP coils. A polished quartz plate coated with MgF and sapphire provides optical access to the wafer surface.

3.1.1 Physics of Interferometry

Plasma induced emission is both reflected and refracted when it shines on the polysilicon film on the surface of the wafer. The refracted beam of light follows Fermat's Principle which states that light will traverse the route with the shortest optical path length [Hecht].

For the reflected beam, the angle of incidence equals the angle of reflection. The path of the refracted beam can be described by Snell's law.

$$n_i \sin\theta_i = n_t \sin\theta_t \tag{3.1}$$

where n_i is the index of refraction for light in the film of incidence and n_t is the index of refraction in the film of transmission. The angles θ_i and θ_t are the angle of incidence and refraction respectively.

From Figure 3.2, the optical path length difference between the reflected and refracted beams is:

$$\Delta l_{opt} = 2 n_2 d \cos\theta_2 \tag{3.2}$$

where d is the thickness of the polysilicon film. Constructive interference of the plasma induced emission occurs when the optical path length difference between the reflected beam and the refracted beam is equal to an integer number of wavelengths of light in the polysilicon film.

$$\Delta l_{opt} = m n_1 \lambda, \tag{3.3}$$

where m is an integer and λ is the wavelength of the light. Destructive interference can also occur when the optical path length difference is a half integer of the wavelength. The period between signal maxima and minima corresponding to constructive and destructive interference can be expressed in terms of the thickness of the polysilicon film.

$$\Delta d_{film} = \frac{\lambda}{2n_2 \cos\theta_2} \tag{3.4}$$

The index of refraction in vacuum, n_1 is taken to be unity. As the film is etched, interference of the reflected and refracted beams modulates the intensity of the observed signal. The time between successive maxima or minima is Δt. The etch rate can be calculated based on the periodicity of the signal, e.g.:

$$EtchRate = \frac{\Delta d_{film}}{\Delta t} \tag{3.5}$$

where Δd_{film} is the thickness of film etched in one period as given in Equation 3.4.

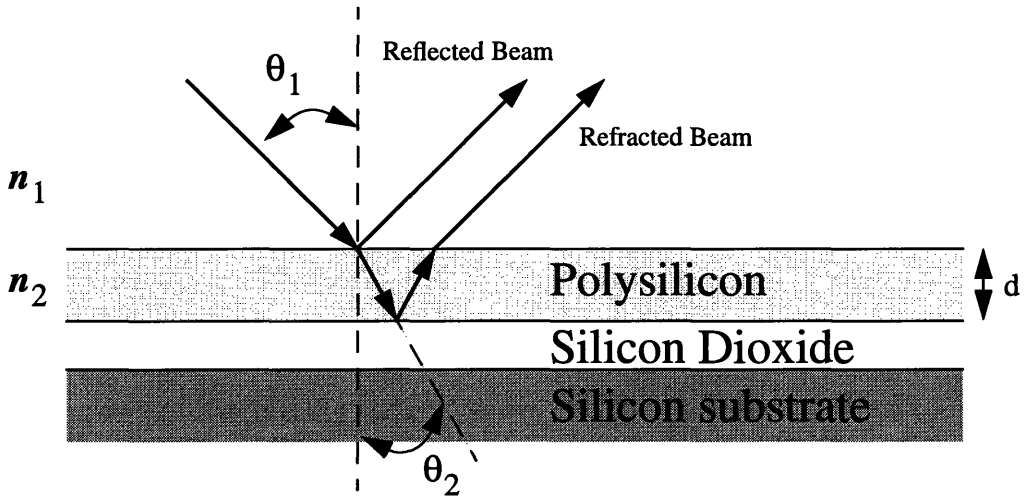

Figure 3.2: The physics of thin film interference for etching of polysilicon

Full Wafer Interferometry calculates the etch rate at many locations across the wafer as illustrated in Figure 3.3. Instead of measuring the intensity of the modulated signal with a single element photodiode, the interference signal is focused onto a CCD array to image the wafer. The above analysis is performed for each pixel element such that the etch rate at locations all across the wafer can be calculated. The Low Entropy Systems software package accounts for lens distortion and non-normal incidence. The periodicity of the interference signal is computed with a Fast Fourier Transform. Although the interferometry signal

is collected in-situ during the etch process, the Low Entropy Systems software calculates the etch rate at the end of every run.

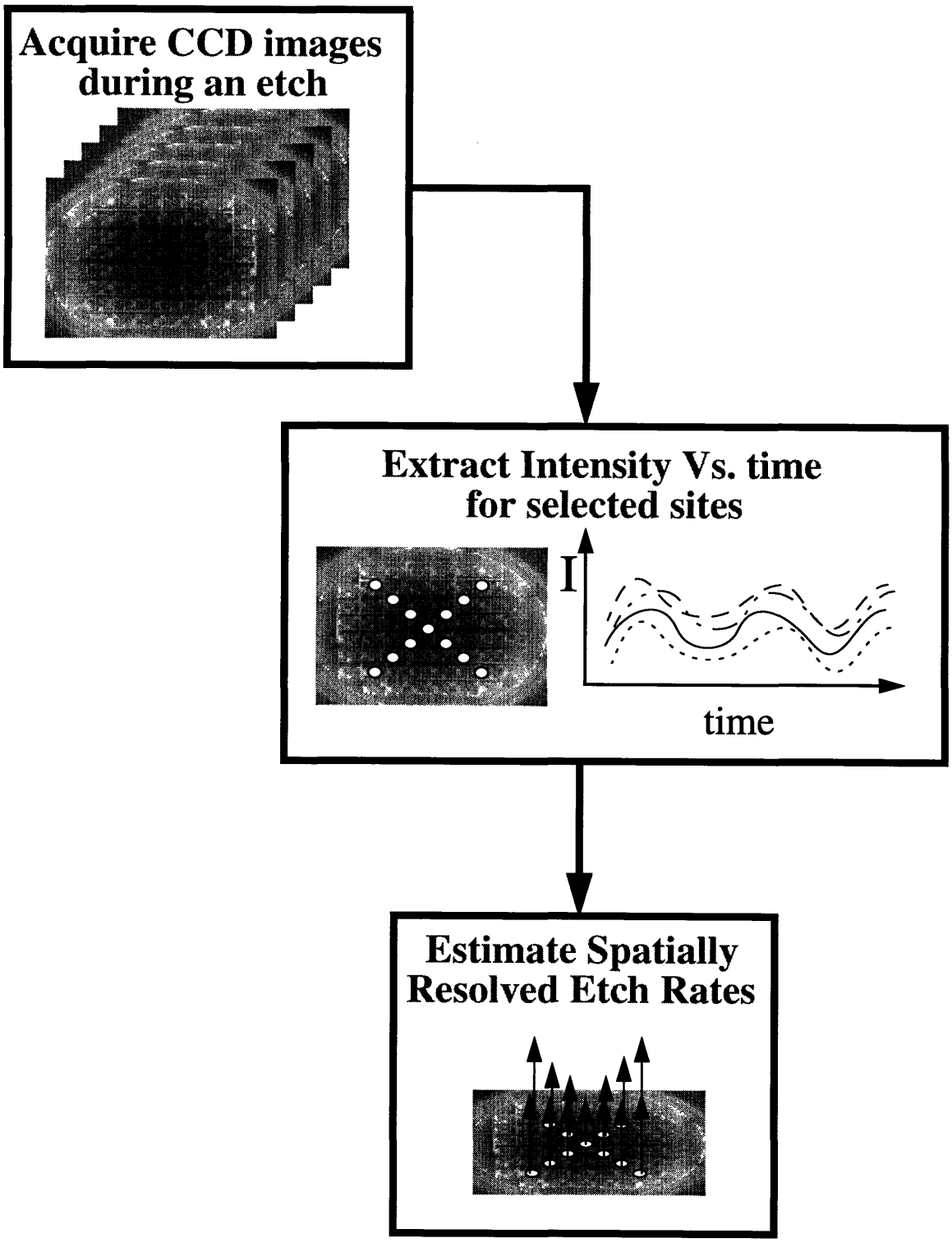

Figure 3.3: Data flow with Full Wafer Interferometry [Gower].

Analysis of the entire CCD array can be time consuming on a 90 MHz Intel Pentium class computer. Instead, a reduced number of selected sites across the wafer are analyzed. A concentric ring pattern is chosen that captures radially nonuniformity, as illustrated in Figure 3.4.

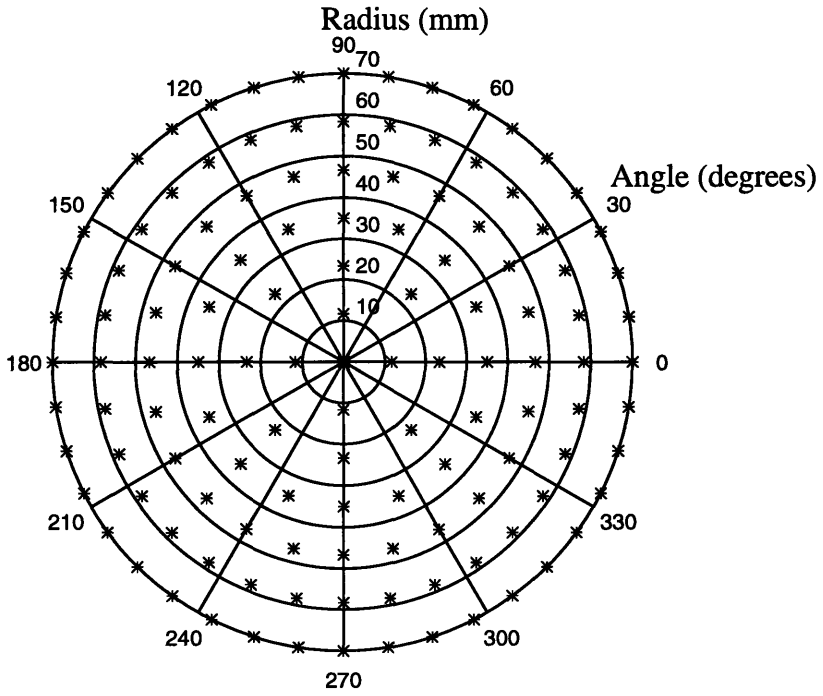

Figure 3.4: The etch rate at various sites across the wafer is measured with FWI. A circular ring pattern was chosen to capture radial nonuniformities.

3.2 Spatially Resolved Optical Emission Spectroscopy

Optical emission spectroscopy provides an abundance of data about the plasma. The presence of excited radicals are observed by distinct spectral features. Different species in the plasma are observed when an excited electron relaxes to a lower energy state. Since these species are excited by electron impact collisions, the intensity of the spectral peak is proportional to the electron density as well as the density of the species of interest. Furthermore, the emission intensity for species x depends on the cross section of electron impact ionization.

$$I^{x}_{emis} = An_e n_x \int \rho_e(E) v \sigma_x(E) dE \tag{3.6}$$

where n_e and n_x are the electron and species densities respectively. The integral is over all energies, E, of the electron energy distribution ρ_e, the electron velocity v, and the collision cross section σ_x [Dalton, 1994]. From this expression it is clear that the plasma spectrum represents a set of complex interactions between the concentrations of various species in the plasma and the electron energy distribution. The spectra is very sensitive to changes in any of the above quantities. This makes optical emission spectroscopy a good candidate to provide a sensitive measure of the plasma state.

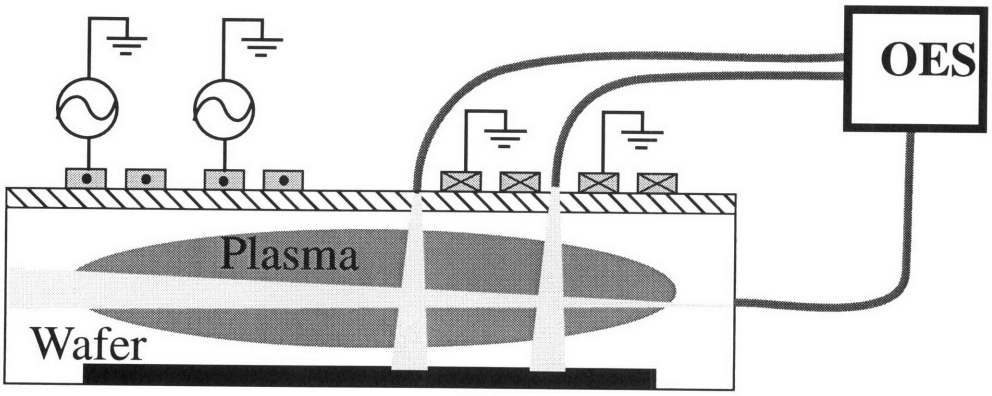

Figure 3.5: A representative diagram of the optical emission spectroscopy system shows three optical fibers located at a sideport and the top of the reactor.

A spectra is collected during the etch process every 600 milliseconds. For run-to-run control, the interest is not in the time series behavior of the spectra, but rather the plasma signature that the spectra provides. For this reason, the spectra during the main etch process can be filtered to remove time dependencies as illustrated in Figure 3.6. For the optical fiber looking through the sideport, the mean spectra is sufficient to average out time (i.e. for each spectral channel, the average over time of the collected samples). The two fibers looking down into the plasma observe interferometric signals in addition to the basic plasma emission signal, so the spectral mode was chosen. The mode for each spec-

tral channel is the midpoint between the maximum and minimum intensity for that particular channel. The result of this time filtering is a single spectra with 3000 spectral channels that is representative of the plasma.

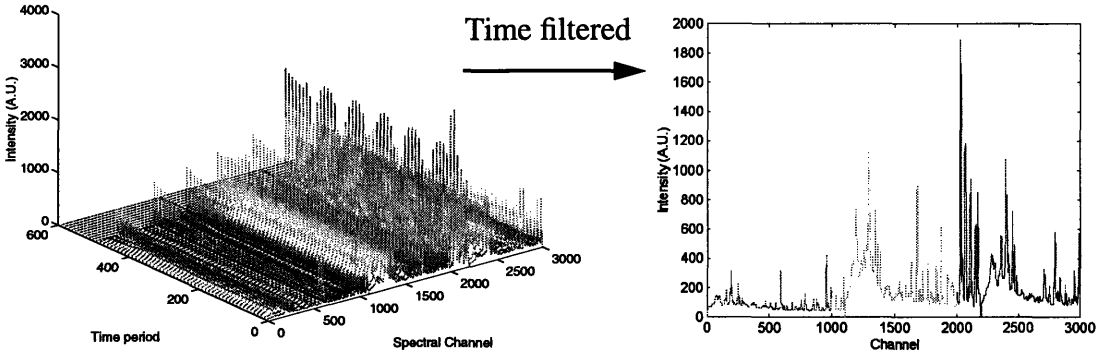

Figure 3.6: Spectra was collected in time but compressed to provide a single representative plasma signature. Three optical fibers were used.

3.2.1 Principle Component Analysis

Even after removing the time component of the spectra, there still remains 3000 spectral channels. The spectra could be modeled directly as a function of process conditions but this would be computationally intensive. Principle component analysis (PCA) is an attractive data reduction technique for this multivariate data set. By choosing a new orthogonal basis set that is a linear combination of the original spectra, the large number of correlated variables can be reduced. Principle component analysis can be seen as the projection of p-dimensional spectra onto a reduced principle component space in which functional modeling can be accomplished more easily [White, 1995].

The spectra during the bulk etch will tend to cluster around a localized region in p-dimensional spectra space. Since it is very difficult to visualize this space when p is large, principle component analysis reduce this space to a more manageable number. Plasma spectra of similar composition will cluster in principle component space as they do in p-dimensional spectra space.

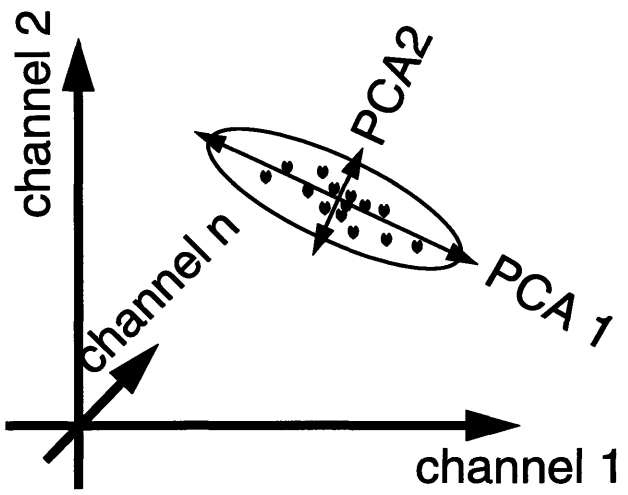

Figure 3.7: Principle component analysis can reduce the number of correlated spectra channels to a much smaller number of principle components.

The mathematics behind PCA can be found in other texts but a basic review is included below [Wise].

Let $\boldsymbol{X}$ represent a matrix with m rows and n columns. There are n spectral channels and m samples from a designed experiment. The covariance is defined as:

$$cov(\boldsymbol{X}) = \frac{\boldsymbol{X}^T\boldsymbol{X}}{m-1} \tag{3.7}$$

when $\boldsymbol{X}$ is mean centered. PCA chooses a set of eigenvectors of the covariance matrix

$$cov(\boldsymbol{X})v_i = \lambda_i v_i \tag{3.8}$$

where v_i denotes the eigenvectors and λ_i the associated eigenvalues. The eigenvectors are mutually orthogonal, that is:

$$v_i^T v_j = 0 \text{ for } i \neq j \tag{3.9}$$

$$v_i^T v_j = 1 \text{ for } i = j \tag{3.10}$$

The choice of this basis set diagonalizes the covariance matrix. The first eigenvector represents an axis in the n dimensional spectral channel space. When $\boldsymbol{X}$ is projected onto this axis, the variance along this axis is maximized. The second eigenvector is orthogonal to

the first and when the spectra are projected onto this axis, the variance along this axis is also maximized. Successive eigenvectors exhibit similar behavior, with each eigenvector capturing less and less of the variance in the collected spectra. In this manner, a full set of principle component eigenvectors can be defined. However, usually much fewer than n components are retained since the first few capture most of the variation in the dataset. Statistical tests can be used to determine the number of significant principle components [Jones].

A truncated set of k principle component vectors still retains the variance in the spectra. Any spectra z can be described as a linear combination of these principle component vectors and the coefficients represent principle component values.

$$z = \boldsymbol{PC}_1 \cdot v_1 + \boldsymbol{PC}_2 \cdot v_2 + \ldots + \boldsymbol{PC}_k \cdot v_k + \boldsymbol{e} \tag{3.11}$$

where $\boldsymbol{PC}_k$ is the k^{th} principle component value and $\boldsymbol{e}$ is the error in the fit. With this technique, instead of representing a spectra with n channels, a reduced number of k principle components can be used.

Chapter 4

Process Control

In this chapter, process control is motivated. We first examine a specific semicondcutor process where there are systematic disturbances. Process control accepts that there are those disturnbances but takes corrective action to ensure that product is within specifications. A model for manufacturing processes is then examined and motivates the need for process control. Finally, the Artificial Neural Network Exponentially Weighted Moving Average controller is discussed.

4.1 Why Process Control?

The need for process control arises from systematic variation in the rate or quality of a manufacturing process. Many processes are not stable and well behaved but rather suffer from shifts or drifts. For example, a metal sputter deposition process at Texas Instruments exhibits a steady drift in the deposition rate as wafers are processed. This drift can be quite significant. When process kits are changed during maintenance periods, the deposition rate shifts. The drifts and shifts in the process is illustrated in Figure 4.1. With feedback process control, the deposition process step duration can be changed so that the deposition thickness can be held closer to target. A good model for the deposition rate is needed for proper control [Smith, 1997].

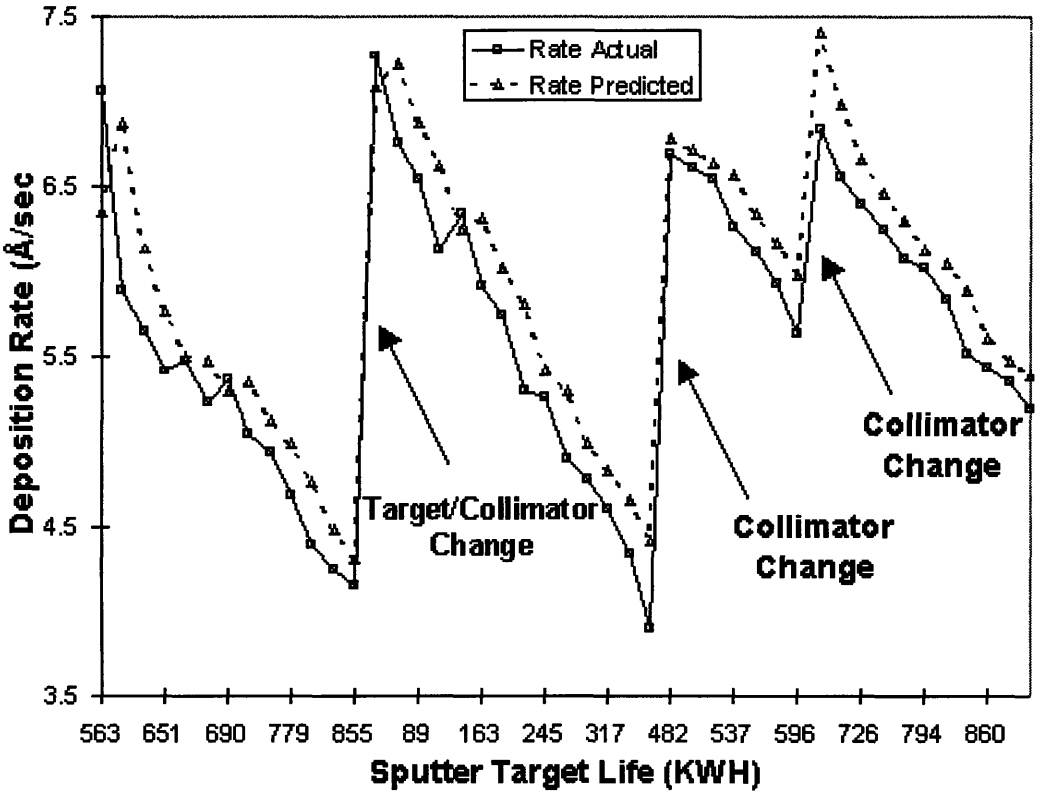

Figure 4.1: Drifts and shifts in the deposition rate are observed in an production process at Texas Instruments. An Exponentially Weighted Moving Average Model tracks the drift and shift [Smith, 1997].

Other semiconductor processes exhibit similar drifts and shifts. In Chemical Mechanical Planarization, as the polishing pad wears away, the removal rate is degraded [Boning]. In plasma etch, as polymer builds up on the chamber walls the boundary conditions are altered; the plasma changes and so does the etch rate and quality. Use of a static process recipe in these cases would result in a wide quality distribution for the parts produced.

4.2 Model for Process Control

In a manufacturing process, material is transformed from one state to another. In semiconductor manufacturing, this can involve the removal of thin films, as in etching and chemical mechanical planarization, the deposition of films, as in chemical vapor deposition and

physical vapor deposition, or the patterning of thin films, as in photolithography. These processes can be generalized as systems that requires input materials, α, and are transformed into the final product via a process recipe, $\boldsymbol{u}$. The resulting product can be characterized by a set of quality metrics such as etch rate or selectivity, summarized by $\boldsymbol{Y}$ and illustrated in Figure 4.2.

$$\boldsymbol{Y} = \Phi(\boldsymbol{u}, \alpha) \tag{4.1}$$

This equation suggests that the quality of the product is a function of the process recipe and other parameters such as material state and properties, as well as the machine state and properties. In real manufacturing systems there are disturbances that can affect the quality of the process. The disturbances in the process are observed as variation in the product, $\Delta\boldsymbol{Y}$. A Taylor expansion of Equation 4.1 shows that variation in the product is a function of variation in directly controllable inputs such as the process recipe, and disturbances in the input material and processing machine as well as the sensitivity of the process to those disturbances.

$$\Delta\boldsymbol{Y} = \frac{\partial \boldsymbol{Y}}{\partial \Delta\alpha}\Delta\alpha + \frac{\partial \boldsymbol{Y}}{\partial \Delta\boldsymbol{u}}\Delta\boldsymbol{u} \tag{4.2}$$

One of the goals of statistical process control (SPC) is to reduce those disturbances that cannot be directly controlled, that is minimize $\Delta\alpha$. Robust process design tries to minimize disturbances in the process output through operation in a region of parameter space where the process is least sensitive to those disturbances. Robust process design minimizes $\frac{\partial \boldsymbol{Y}}{\partial \Delta\alpha}$.

Statistical process control and robust process design have proved their utility in the semiconductor industry [Spanos]. However, for processes that exhibit fundamental drifts or shifts, these methods fall short in controlling variation in the output. Drifts can be due

to seasoning of the machine and shifts can be attributed to preventive maintenances or other non-idealities. Changes in the process recipe can compensate for disturbances to minimize variation in the output. In feedback process control, Δu is varied in order to minimize ΔY.

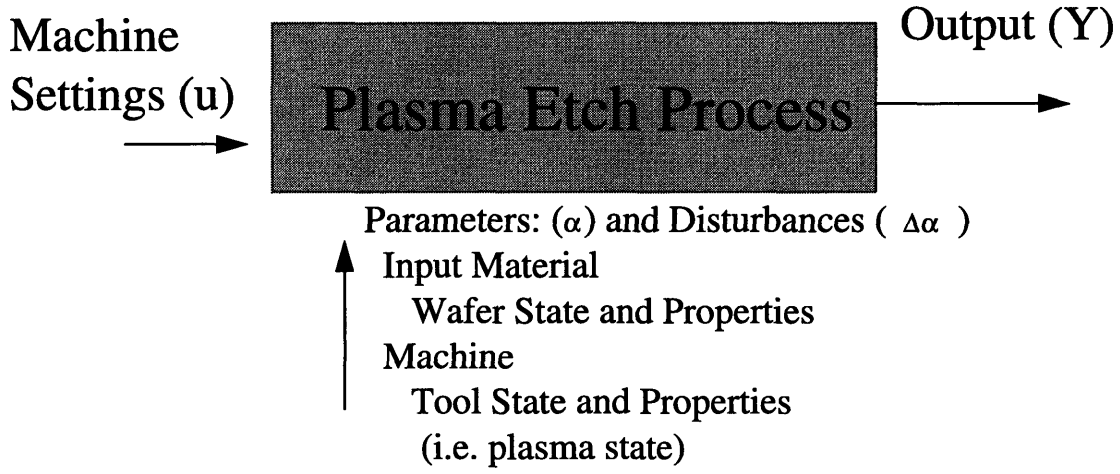

Figure 4.2: The model for manufacturing processes shows that the rate and quality of a process is a function of directly controllable parameters such as process recipe and other disturbances and parameters [Hardt].

4.3 Response Surface Models

The controller in this thesis relies on a model of the process. This model is built from a designed experiment. The empirical model provides an input-output relationship between the process recipe and the rate and quality of the process as measured by various sensor systems. In the plasma etch control demonstrated in this thesis, we utilize the Full Wafer Interferometry sensor described in Chapter 3, which can provide etching rate information at multiple locations across the wafer. A Response Surface Model (RSM) can be developed for each individual site across the wafer. From the designed experiment, a relationship between the process recipe and the etch rate at each site is obtained. All the site models are then incorporated into a model for nonuniformity. Guo and Sachs suggest that this Multiple Response Surface model for nonuniformity is more robust than modeling nonuniformity directly [Guo].

4.4 The Artificial Neural Network Exponentially Weighted Moving Average Controller

Early experiments on the dual-coil TCP reactor found significant non-linearity in the response surface model for uniformity. The Artificial Neural Network Exponentially Weighted Moving Average (ANN EWMA) controller provides the ability to incorporate such complex models with a simple and efficient EWMA model feedback mechanism [Smith, 1997; Smith, 1996]. A block diagram of the controller is shown in Figure 4.3. The controller assumes that the underlying process model curvature is not affected by drifts or shifts in the equipment. Recent measurements are used to shift the constant term in the nonlinear process model to account for machine drifts or shifts.

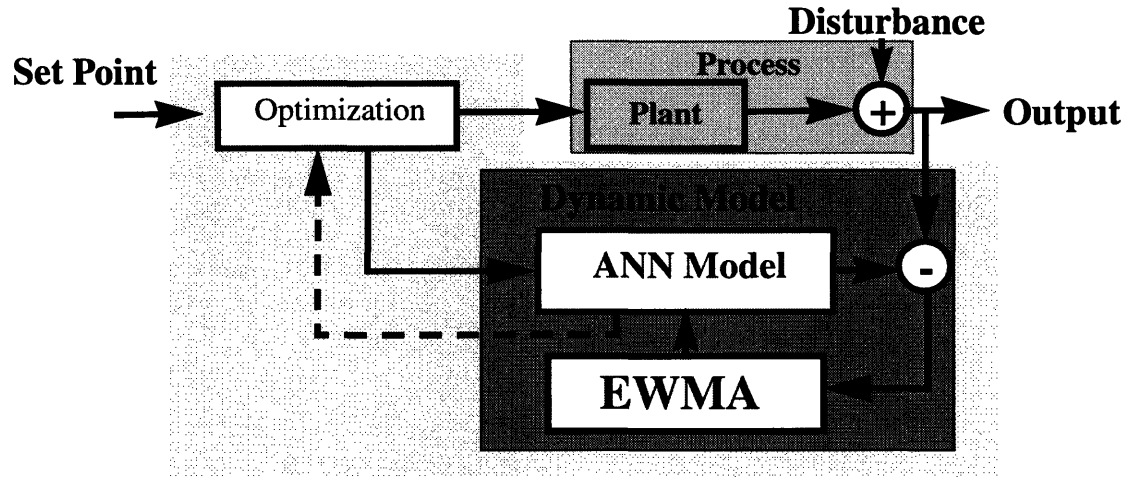

Figure 4.3: Block diagram of Artificial Neural Network Exponentially Weighted Moving Average controller. The ANN EWMA controller uses information from sensors to adapt nonlinear multiple response surface models to process disturbances. An optimizer chooses a recipe based on the most up to date model.

The controller model utilizes an approximation to the nonlinear system of the form

$$\tilde{y}[n] = \tilde{f}(x, u, r, n) + b[n], \tag{4.3}$$

where x, u, and r are vectors corresponding to the process state, process inputs, and neural network approximation parameters at discrete-time n, respectively. The underlying response is subtracted from the previous measurement of the output vector, $\hat{y}[n]$, to

obtain a measurement of the offset term

$$\hat{b}[n] = \hat{y}[n] - \tilde{f}(x, u, r, n). \tag{4.4}$$

This is then used to update the previous offset term using an exponentially weighted moving average (EWMA) of the form

$$b[n] = W(\hat{b}[n]) + (I - W)b[n-1], \text{ where } W = diag\left(\begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix}\right). \tag{4.5}$$

Such an update does not require retraining of the neural network, and thus is computationally efficient.

High weight values (w close to 1) allow the model to react quickly to drifts in the system response but run the risk of overreacting to system noise. Low weight values (close to 0) cause the controller to respond more slowly but with better rejection of noise.

4.4.1 Artificial Neural Network Model

The artificial neural network model provides a functional approximation for the multiple response surface models. For each site on the wafer, a general curve is fit to the etch rate as a function of the process parameters. The neural network model provides the basis for smooth curve fitting. One hidden neuron layer comprised of tan sigmoid functions are used in the network [Demuth]. The network is based on a Levenberg-Marquardt back-propagation technique. This Newton's Method type optimization provides more efficient training of the network than a gradient descent routine [Demuth]. A diagram representative of the neural network is shown in Figure 4.4.

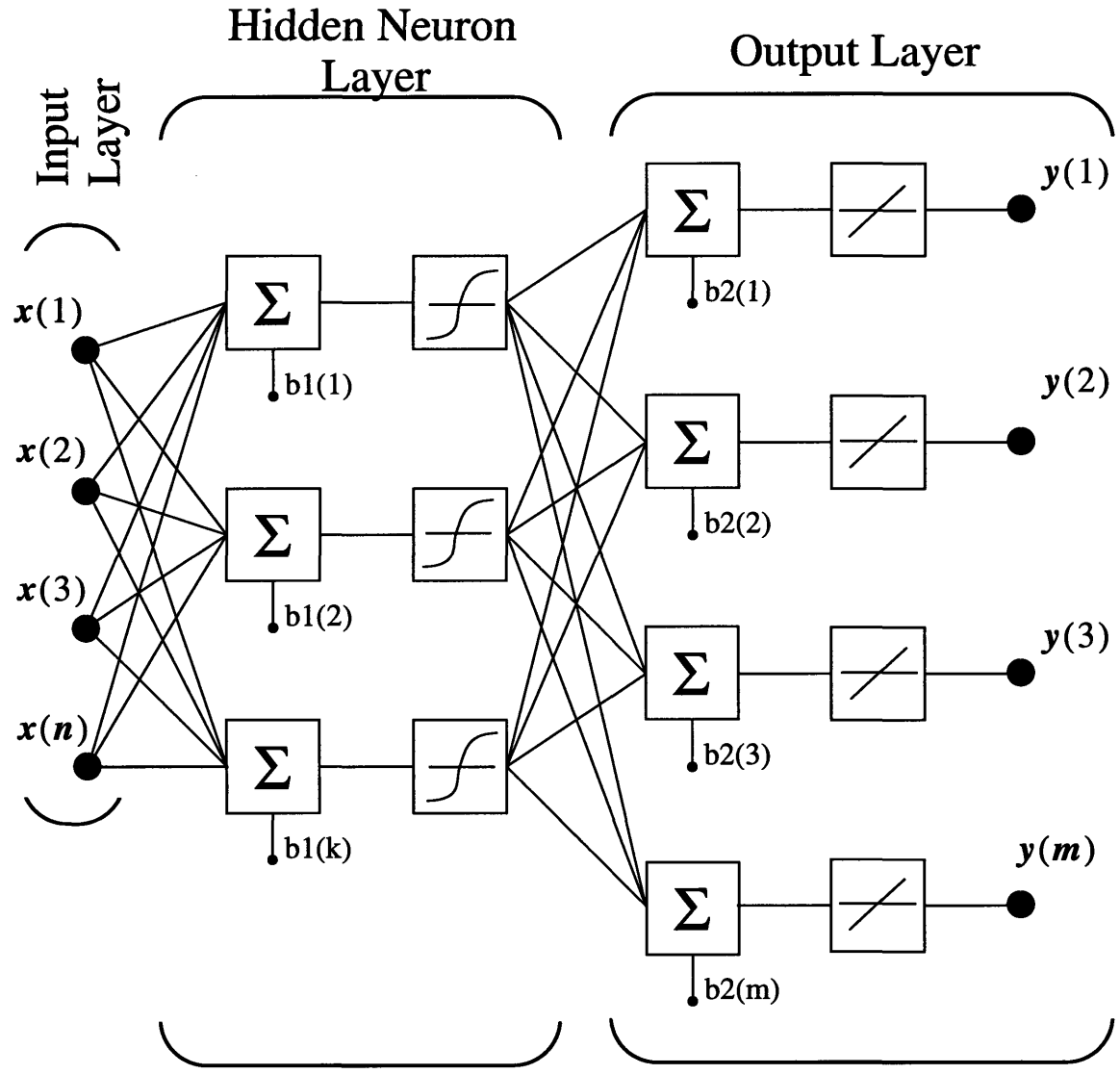

Figure 4.4: The Neural Network model is comprised of n input nodes, k nodes in the hidden layer, and m nodes in the output layer. The hidden layer uses tan sigmoidal functions while the output uses linear transfer functions [Demuth].

A concern with any form of data fitting is underfitting and overfitting. This is especially true when using neural networks as functional approximations where the physical basis for functional dependence is not present. Figures 4.5a and 4.5b show two examples of underfitting and overfitting.

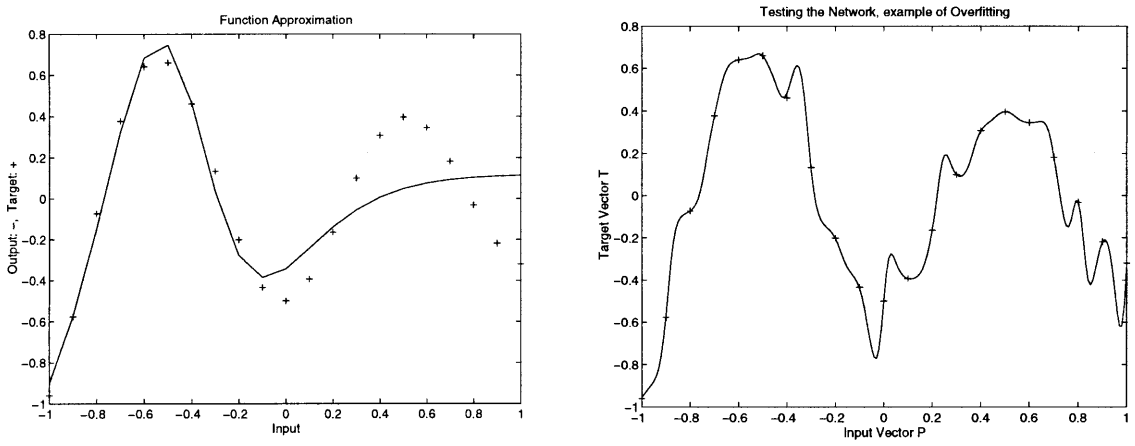

Figure 4.5: Examples of underfitting and overfitting a neural network model to discrete data points. a) The model does not fit the experimental points well. b) The model overfits the data points. There are high frequency features in the model that does not appear in the data [Demuth].

The neural network models for the etch rate as a function of the process recipe was observed for each site to ensure that underfitting and overfitting was avoided. Underfitting is representative of too few neurons in the hidden layer and likewise, overfitting is indicative of too many.

4.4.2 Exponentially Weighted Moving Average

The bias layer in the neural network model is well suited for update since it involves constant offsets to the model. An Exponentially Weighted Moving Average is performed on the bias layer as described by Equation 4.5. The EWMA shifts each response surface model but retains the functional relationship between the inputs and outputs of the model.

The EWMA update can be visualized with the aid of Figures 4.7 and 4.8. In the first example, a linear response surface model is used. The original model is developed through a series of experiments that determine the slope and offset of the model. A target process output is chosen and a process recipe is calculated from the linear model. Over the period

of many runs, the EWMA update adapts the model to the drifting processes. The slope of the model does not change, only the offset is updated. Many experiments are run over a wide region of parameter space to develop the process model. During operation, a process operates in a much more narrow region of parameter space. If the slope of the linear model is changed, there is a very strong risk of adapting the model to process noise and overfitting to the local region. This is avoided by shifting the offset instead of the slope in the process mode.

A new process recipe can be calculated with the most up to date process model. The process target defines the output and a linear solution can be found.

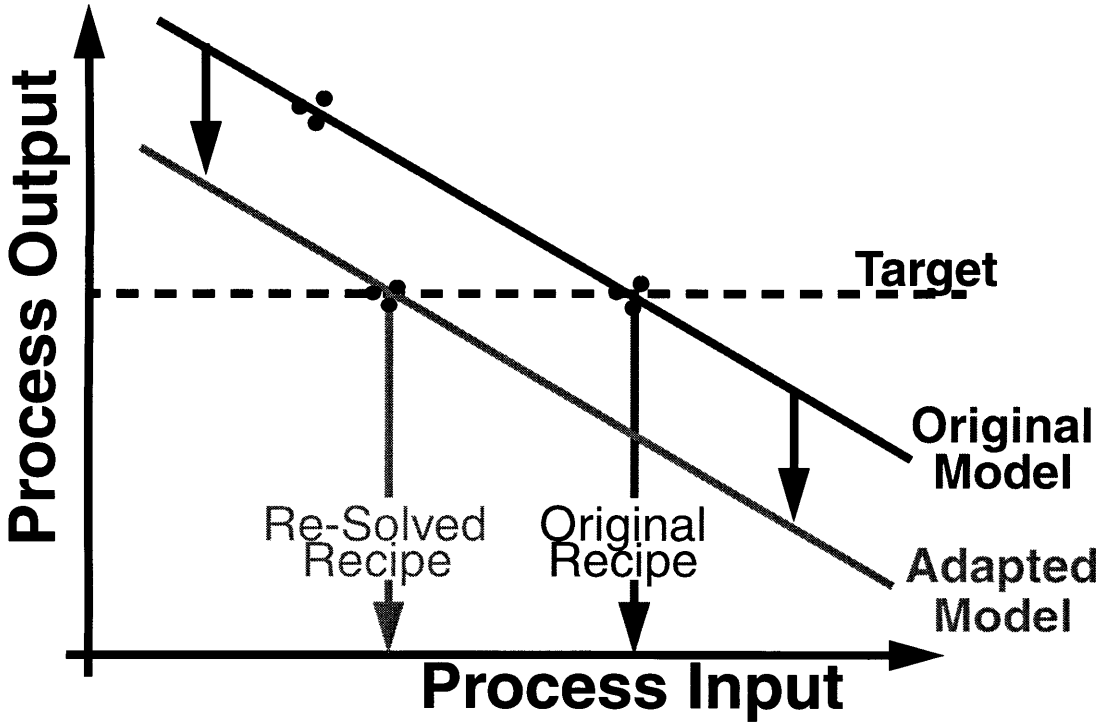

Figure 4.6: EWMA of the offset in a linear model based controller. The model offset is updated and a new recipe is calculated.

The linear EWMA controller has been successfully applied to many semiconductor processes including LPCVD and CMP [Sachs; Boning]. In these cases, in-situ measurements were not available whereas in this work, FWI and OES affords an opportunity to

use sensor data on a wafer to wafer basis to make corrective action if necessary. Sachs, Hu, and others explored run-to-run plasma etch control in parallel plate and magnetically confined plasmas [Hu]. In those experiments, pre- and post-measurement of the film thicknesses were required to estimate etch rate; thus practical application of the control was not possible (since the thin film is completely removed during actual etch processes). In this work we integrate both in-situ metrology and equipment modifications to achieve effective and practical plasma etch control.

The diagram for the linear EWMA controller can be extended to incorporate nonlinear response surface models. In the case of the ANN-EWMA controller, the nonlinear model is represented by the artificial neural network. The model is again shifted through update of the offset terms as shown in Figure 4.7. In the linear EWMA controller, a new process recipe is solved through inversion of a linear model. The artificial neural network model cannot be directly inverted. The process recipe is instead calculated with an optimizer that will be described in the Section 4.4.3.

With nonlinear models, a single solution is not guaranteed. There may be two or more process recipes that result in the same process output, or there may even be no solution to the model. These situations can be avoided through judicious choice of quality metrics (process outputs) and optimal cost function. By ensuring that the objective cost function is concave, there is identically one solution to the minimization problem.

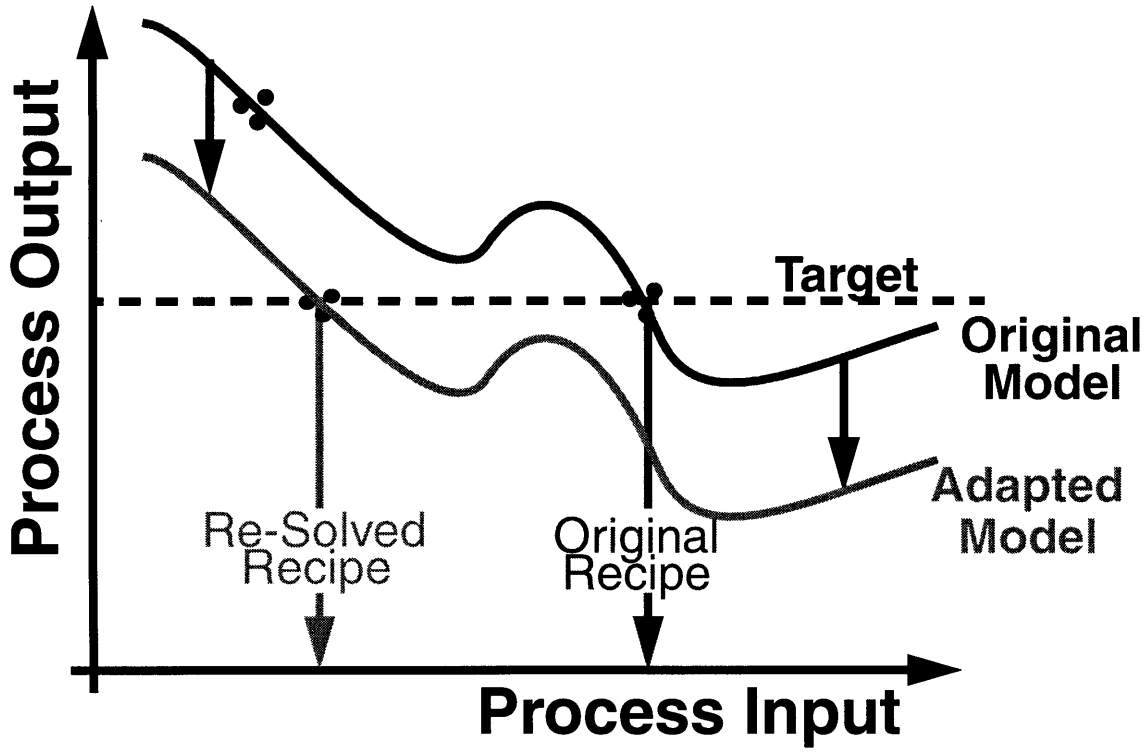

Figure 4.7: EWMA applied to a generalized nonlinear process model.

4.4.3 Optimizer

The best recipe to run next, based on the most up to date model, is chosen with an optimizer that seeks to minimize some overall cost function. The cost function is calculated from the multiple response surface models in the space of the process recipe. For example, for a two input space, the cost function is a surface in that space and the optimizer chooses the lowest point on that surface. This optimization is nonlinear because the response surfaces are based on neural networks which have nonlinear process parameter dependencies. Furthermore, the optimization is constrained because the recipe suggestion should not be outside of the input space of the designed experiment. Sequential Quadratic Programming is used to achieve this nonlinear constrained optimization of the cost function. In the applications of Chapters 5 and 6, the cost functions were well behaved and concave so that

locating a minima was computationally efficient. The previous recipe is used as the start point for the optimization since the optimal recipe should not change much on a run to run basis. Specific details of the cost functions developed for different plasma etch control scenarios will be presented in Chapters 5 and 6.

One concern with optimization is that the optimal recipe that is calculated may not be the best recipe in the input space but rather represent a local minimum in the cost function. Simulations of our plasma etch cost functions show that this is not a significant concern, however, this issue could be addressed by performing multiple optimizations of the cost function with random start points. After all of the local minimas are discovered, the lowest cost is chosen. By using many start points, the chance of falling into a local minimum instead of a global minima is reduced. Though this method has been implemented in other optimization codes, it was not used in this work but can be readily incorporated in future controllers.

Chapter 5

Process Control on the Dual Coil TCP with Full Wafer Interferometry

5.1 Research Objective

A key goal of this research is to demonstrate the ability for a model based controller to optimize etching rate uniformity across the wafer and maintain that uniformity between wafers. The first experiments involve the manipulation of only two process parameters, the power delivered to the inner and outer coils. Full Wafer Interferometry is used to measure the quality of the etching process. Gower has implemented a similar control system on an AME 5000 plasma etcher [Gower]. In that tool, major equipment modifications were not made as radial etching uniformity was well controlled by adjustment of the rotating magnetic field intensity. In this work, we apply the same concept to the dual coil TCP that was specifically modified to enable effective process control.

5.2 Process Conditions

Power to the inner and outer coils are the only parameters in the process recipe that are manipulated. Pressure was held constant at 15 mTorr, as was gas flow (120 sccm) and composition (5:1 of HBr:Cl_2). 50 Watts of power to the lower DC bias electrode was delivered by a 4 MHz RF generator. 150 mm wafers with 5000 Angstroms of blanket polysilicon over 1000 Angstroms of silicon dioxide were etched. Table 5.1 summarizes the process conditions.

Gas	HBr:Cl2 (5:1) 120 sccm
Pressure	15 mTorr
RF Bias Power	50 Watts
Electrode temperature	50° C
Inner Coil Power	400-600 Watts
Outer Coil Power	100-200 Watts

Table 5.1: Process Parameters for control with FWI

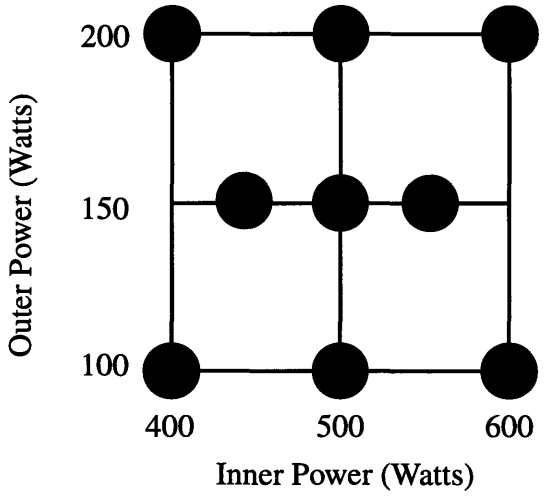

Figure 5.1: A two input variable designed experiment was performed to model the etch rate at specific sites across the wafer.

Results of a two variable experiment were used to train an artificial neural network. The power to the inner coil ranged between 400 and 600 watts at 11.00 MHz while power to the outer coil changed between 100 and 200 watts at 13.56 MHz. The optimum etch rate uniformity, in this operation space, is about 1.5% measured at 1 sigma. The experimental design is summarized in Figure 5.1 and Table 5.2.

Run	Inner Power (W)	Outer Power (W)
DEC3	600	200
DEC4	500	300
DEC5	500	100
DEC6	600	100
DEC7	500	200
DEC8	400	200
DEC9	500	150
DEC10	550	150
DEC11	450	150
DEC12	500	150
DEC13	400	100

Table 5.2: Process parameters and results

5.3 Linear Multiple Response Surface Models

A linear model was trained to the designed experiment to gauge the goodness of fit. This model had the form:

$$EtchRate_i = a + b \cdot x_1 + c \cdot x_2 \tag{5.1}$$

where $EtchRate_i$ is the etch rate for site i on the wafer. x_1 and x_2 are the process settings and a, b, and c are the fitting parameters. A model for the etch rate at the center point was found to have an R^2 value of 0.885. This result indicates that it may be possible to use a linear model for control. Higher order models were evaluated to test the significance of higher order terms.

	Sum of Squares	Degrees of Freedom	Mean Square	F-Ratio	P value
Regression	1.649	2	0.824	27.064	0.001
Residual	0.213	7	0.030		
Total	1.862	9			

Table 5.3: ANOVA table for a linear site model [Goodlin]

5.4 Quadratic Multiple Response Surface Models

A quadratic model was fit to the etch rate at a single site on the wafer. This model had the form:

$$EtchRate_i = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_1 x_2 + e \cdot x_1^2 + f \cdot x_2^2 \qquad (5.2)$$

where the coefficients and process parameters are as defined above. The results of the fit indicate that the R^2 was 0.997, a very good fit. The ANOVA table below summarizes the results.

	Sum of Squares	Degrees of Freedom	Mean Square	F-Ratio	P value
Regression	1.858	5	0.372	314.838	0.000
Residual	0.005	4	0.001		
Total	1.862	9			

Table 5.4: ANOVA table for a quadratic site model [Goodlin]

These results indicate that there are higher order effects that the second order model captures that the linear model does not. A more general nonlinear model could replace the second order model.

5.5 Neural Network Models

A neural network model is trained to the results of the factorial experiment. The etch rate

is measured at 81 locations across the wafer in a concentric circular array. The model relates the process recipes to the etch rate at each of the 81 locations on the wafer. This represents a multiple response surface model for the system response. There is a functional relationship between the etch rate at each of the 81 locations and the process recipe. A block diagram showing the input and output nodes of the neural network model is illustrated in Figure 5.2. The neural network model was chosen for its generality. It can be an approximation to an arbitrary response.

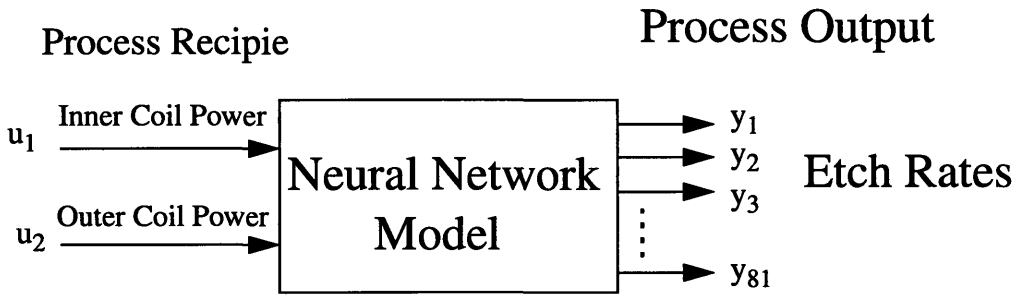

Figure 5.2: Process model for the etch rate across the wafer

The neural network model had one hidden layer comprised of 6 tan sigmoial neurons. There were two nodes in the input layer and 81 nodes in the output layer. The response surface from the neural network model for a single wafer site is shown in Figure 5.3. The goal is to have the response surface for all wafer sites to cross at the same point in process recipe space; for that particular process recipe, the etch rate at all measured points would then be equal. Because the system is underdetermined, (we have many fewer control inputs than sites to be affected), the more realistic goal is to choose a recipe such that the variation in etch rate is minimized.

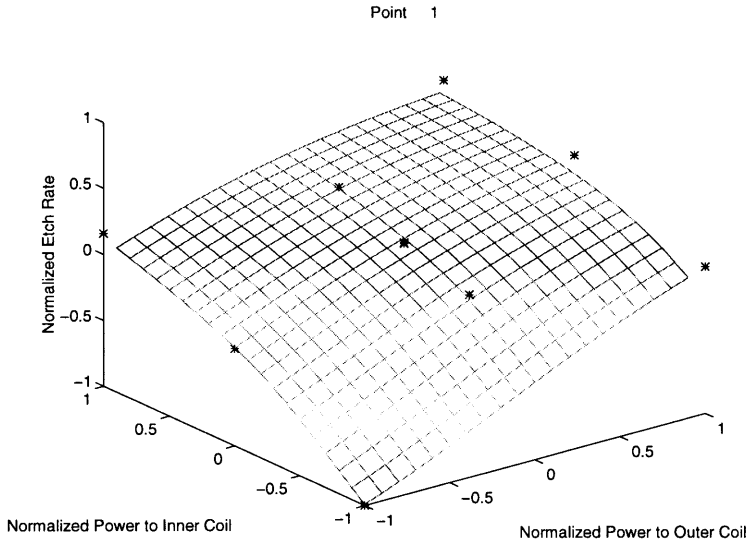

Figure 5.3: NN Model for etch rate at site 1 (center of the wafer). As the power to the inner and outer coils are increased, the etch rate increases.

5.6 The ANN-EWMA controller with FWI sensor information

The Artificial Neural Network Exponentially Weighted Moving Average controller incorporates the multiple response surface model constructed using Full Wafer Interferometry data. Each of the site models is allowed to evolve by means of the EWMA of the offset term as described in Section 4.4. A diagram of the integrated controller is shown in Figure 5.4 illustrating the data flow between the sensor and the controller. Simulations of the dual-coil TCP suggest a small amount of noise, and in order to test the response of the controller, high EWMA weights, 0.75, were chosen. Using the most up to date model for the process, the controller chooses a recipe perturbation that optimizes an objective defined by a specified cost function.

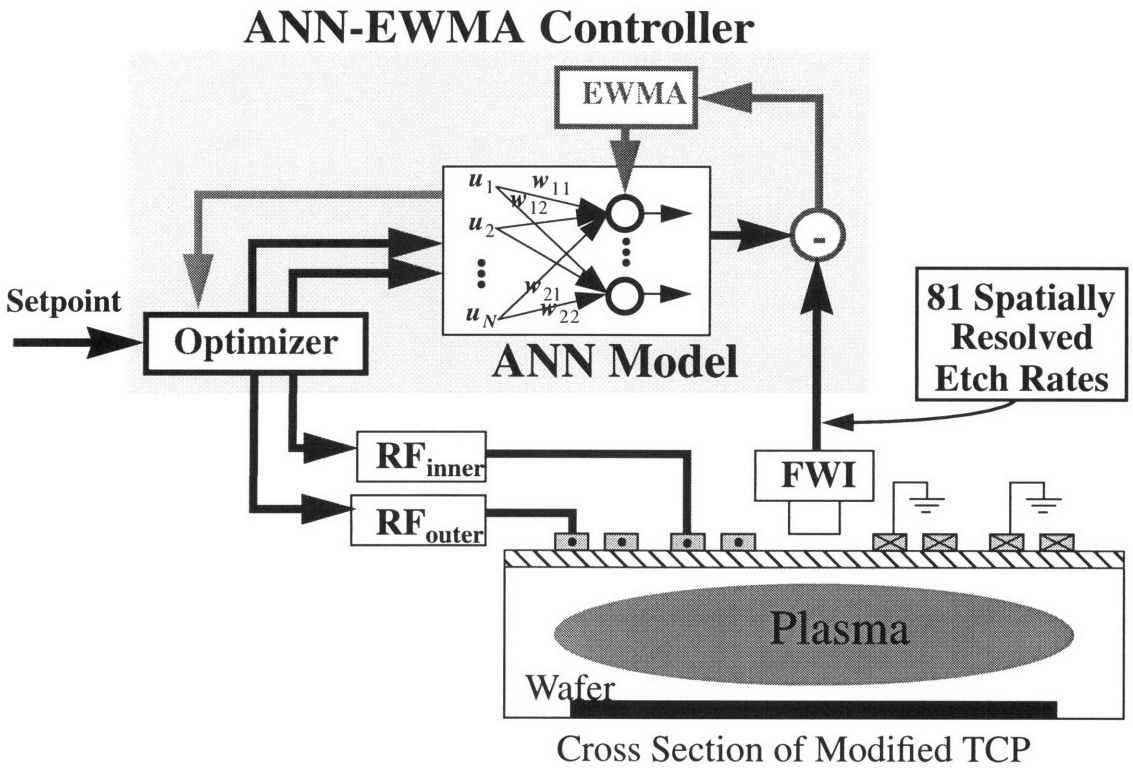

Figure 5.4: The integrated controller and sensor shows the data flow between the sensor and various components of the ANN EWMA controller.

5.7 Objective function

To test the ability of the controller to optimize etch rate uniformity, the objective cost function chosen was:

$$min\left(\beta \cdot \frac{std(\hat{y}[n])}{mean(\hat{y}[n])} + (1-\beta) \cdot \|u[n] - u[0]\|\right), \tag{5.3}$$

where all parameters were mean centered and normalized. There are two terms in the cost function. The first involves the normalized standard deviation of the etch rates. This term pushes the controller towards maximizing the etch rate uniformity at 81 locations on the wafer, as measured by FWI, while also seeking to increase the mean etch rate. The second term involves a minimum recipe change from a setpoint. This ensures that the recipe proposed by the controller does not deviate too far from the setpoint. The weighting factor, β,

of the two terms was chosen to be 0.99, based on simulations of the ANN EWMA controller, and takes into account the relative magnitudes of the two terms.

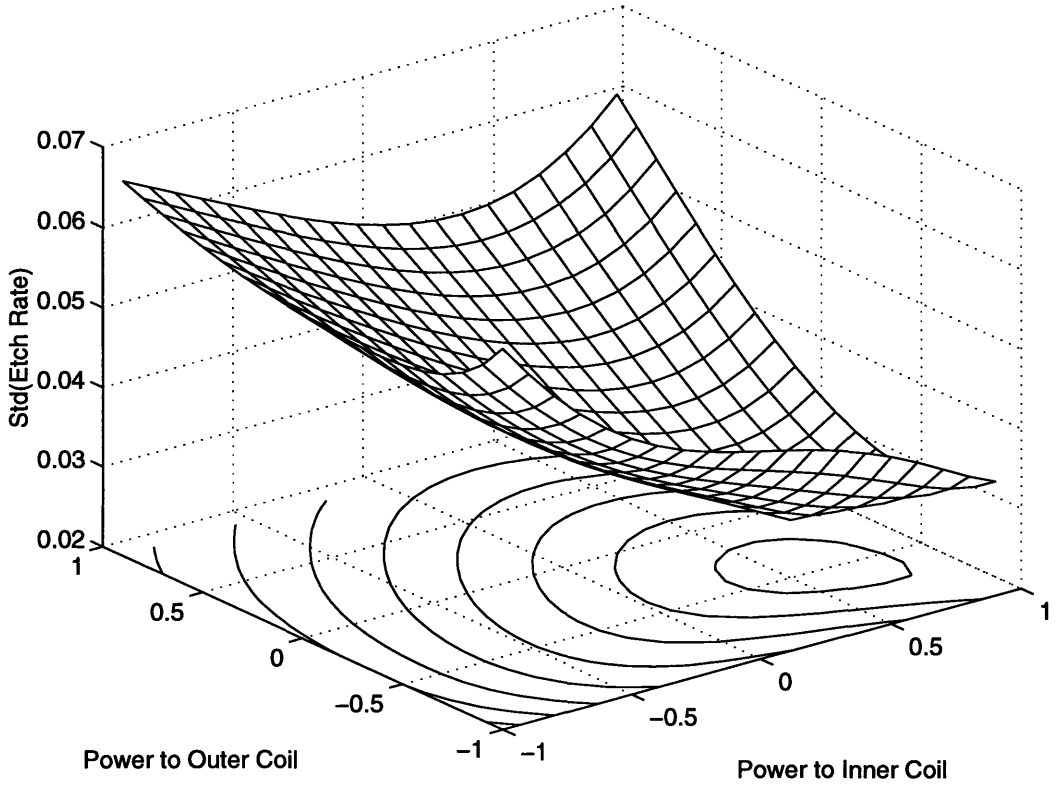

Figure 5.5: The standard deviation of the etch rate across the wafer is lowest at some process recipe. This is incorporated into the objective cost function which like the above hyperplane is concave.

After a wafer is etched, the most recent measurement of the etch rate at each individual site updates the model. This can be visualized by imagining the response surface in Figure 5.3 shifting up or down. Since each of the 81 response surfaces shifts, the overall objective function defined by Equation 5.1 changes shape and the minima of that curve is the optimal recipe for the next wafer. For processes with only two inputs, this is easy to visualize, however as the length of the input vector increases, visualization becomes more difficult. An extension to higher dimensions involves hyperplanes but mathematically, the procedure is no different.

5.8 Experimental results

A 50% step disturbance (50 Watts) in the outer coil was introduced during the 6th run in order to observe the controller response. This offset caused a disruption in the etching uniformity. With spatial etching rate information from FWI, the controller was able to observe that the edge of the wafer etched slower and made a process recipe correction accordingly, as shown in Figure 5.6. After three wafers, the ANN EWMA controller was able to bring the etching uniformity back to roughly 1.5%. By the end of the experiment, the uniformity improved to about 1%.

From Figure 5.6 we observe that the mean etch rate was held close to about 2500 Angstroms per minute. The dashed lines show the model prediction of the system response given no control action. If the controller was not allowed to adjust the process recipe given the disturbance introduced, the etch rate would have remained below 2200 Angstroms per minute and the uniformity would have remained poor.

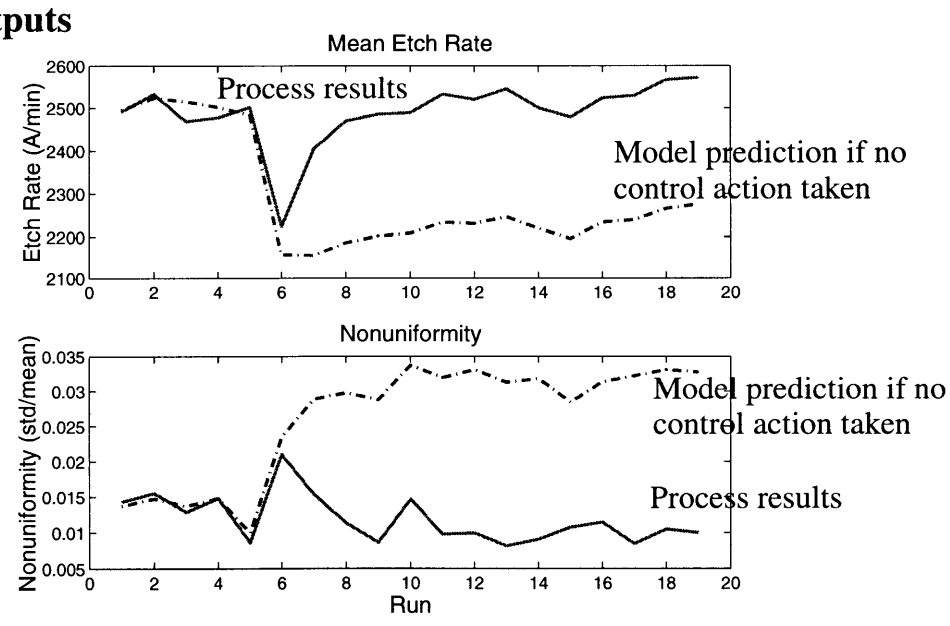

Figure 5.6: The results above are for control of the etching rate uniformity with just FWI sensor information. A step disturbance was introduced at wafer 6 and shifted the mean etch rate and nonuniformity. After a few runs, the controller response brings the mean etch rate and nonuniformity back on target.

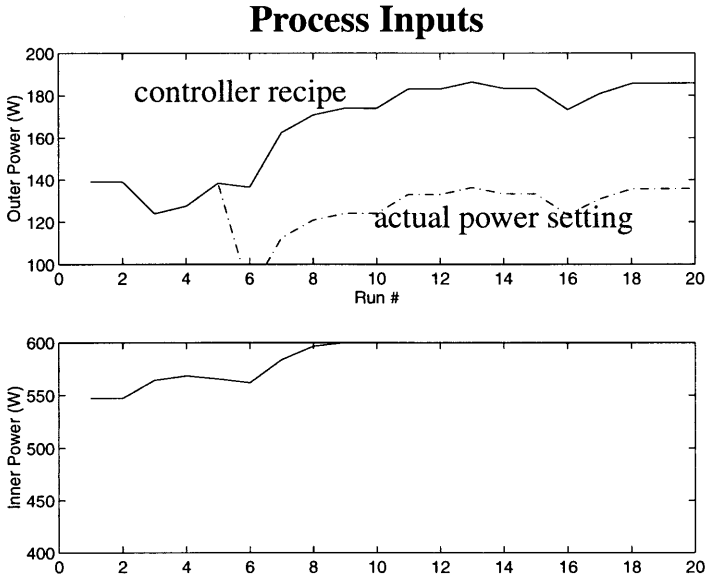

Figure 5.7: The process recipe included two variables that were allowed to change, power to the inner coil and power to the outer coil.

The process shift introduced at wafer #6 was a 50 watt shift in the outer coil power. After the disturbance, the controller did not return the power exactly 50 watts back but instead made control actions that changed both the inner coil power and outer coil power. The control action actually improved the nonuniformity down to about 1% as demonstrated in Figure 5.6.

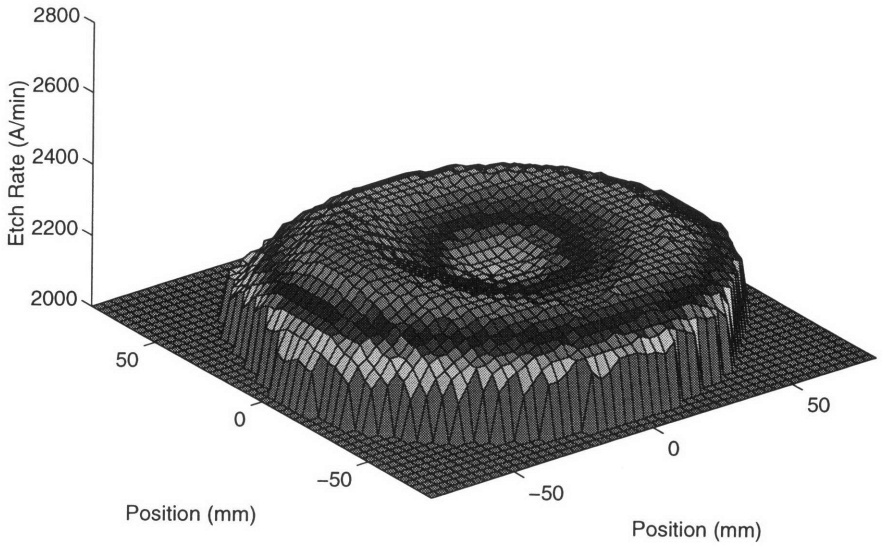

Figure 5.8: The process disturbance was introduced at wafer #6. The mean etch rate drops and the etch rate uniformity is poor. The distinct bull's-eye pattern represents radial etching rate nonuniformities.

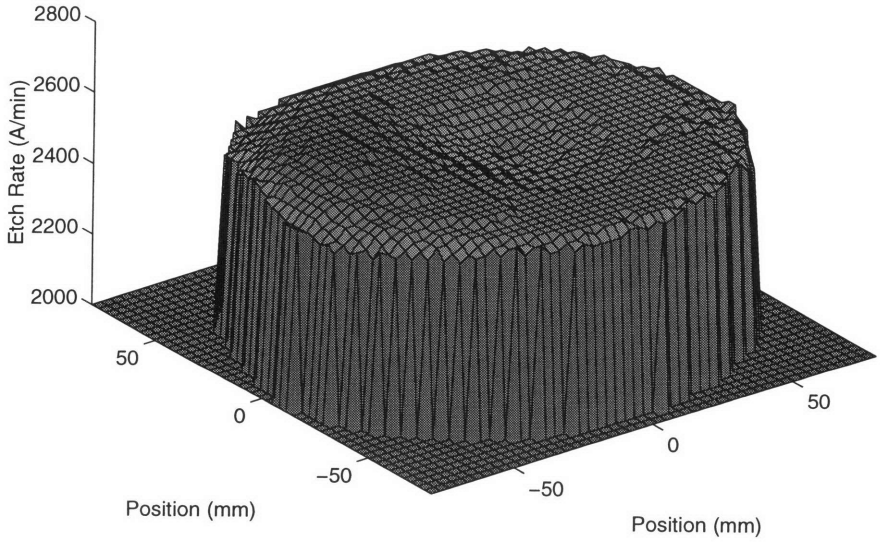

Figure 5.9: After many wafers were etched, the process recipe perturbations suggested by the controller was able to bring the mean etch rate and the etch rate uniformity back to the process targets. The uniformity has improved to about 1% and the mean etch rate up to about 2500 Angstroms per minute.

These results indicate that the run-to-run neural network model based process controller was able to respond to a step disturbance in the process and bring the process quality metric back on target. The quality metric used in this experiment was the etching rate non-uniformity as measured by the in-situ FWI system.

Chapter 6

Process Control on a Dual Coil TCP with FWI and OES

6.1 Multi-Objective Control with Multiple Sensors

The experiment presented above incorporates a single sensor that provides information about the wafer state. Other types of sensors can provide different information that can complement Full Wafer Interferometry. We choose to use optical emission spectroscopy to provide a measure of the plasma state. This sensor is also incorporated into the process control strategy.

The ANN-EWMA controller is readily adapted to take advantage of multiple sensors. A new model for OES is added to the ANN-EWMA controller presented in Section 4.4. Instead of just a single model, the optimizer uses two models to arrive at an optimal recipe as shown in Figure 6.1. The ANN-EWMA controller with two sensors can clearly be further extended to include multiple sensors and multiple objectives by adding more models.

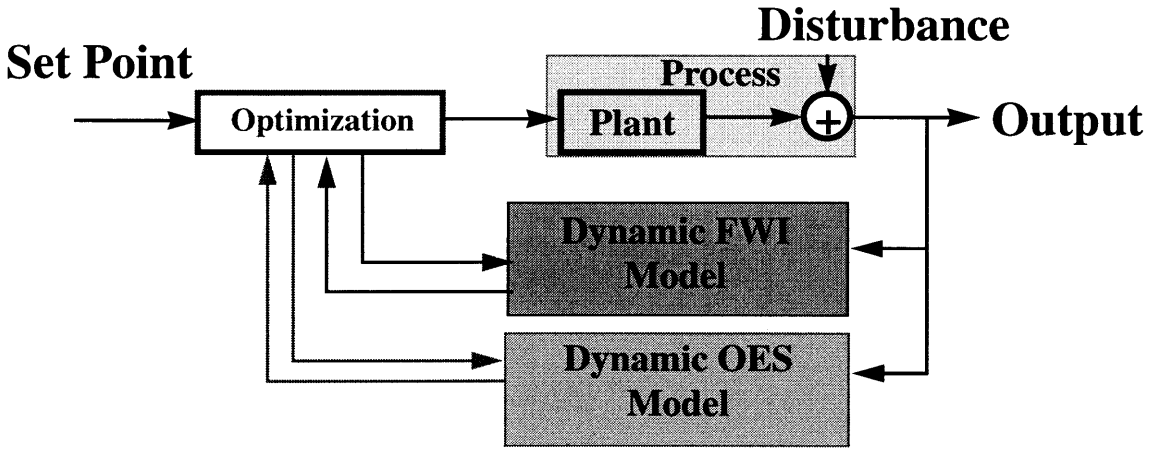

Figure 6.1: Integrated FWI/OES controller. Two separate models are used by the optimizer to arrive at a recipe perturbation.

6.2 Process Conditions

A three variable, three level full factorial experiment was performed to build new models for both the FWI sensor data and the OES sensor data. The design of experiment (DOE) is presented below in Table 6.1.

Run	Inner Power (W)	Outer Power (W)	Pressure
LL001	500	300	15
LL002	600	200	15
LL003	500	400	24
LL004	500	200	15
LL005	600	300	24
LL006	500	200	24
LL007	400	300	24
LL008	400	200	15
LL009	600	400	20
LL010	400	300	15
LL011	400	300	20
LL012	600	300	15
LL013	400	400	15
LL014	400	400	20
LL015	400	300	20
LL016	600	400	24
LL017	500	300	20
LL018	500	400	20
LL019	500	300	15
LL020	600	300	20
LL021	600	200	24
LL022	500	300	24
LL023	600	300	20
LL024	400	400	24
LL025	500	200	20
LL026	600	400	15
LL027	400	200	24
LL028	500	400	15

Table 6.1: Three variable three level full factorial experiment

In these experiments, the generator supplying power to the inner coil was the 13.56 MHz Advanced Energy power supply while the outer coil was driven by the 11.0 MHz Seren IPS system. This is opposite from the previous experiment to allow more power to the inner coil since it was observed that the inner coil power was limited by the Seren IPS generator. With the swap, the power settings are no longer constrained by the maximum output of the generator.

The other process parameters such as bottom electrode power and gas flow are kept constant as they were in the previous experiment. A summary of the process conditions is shown in Table 6.2.

Gas	HBr:Cl2 (5:1) 120 sccm
Pressure	15-24 mTorr
RF Bias Power	50 Watts
Electrode temperature	50° C
Inner Coil Power	400-600 Watts
Outer Coil Power	200-400 Watts

Table 6.2: Process Parameters for control with FWI and OES sensors

6.3 Response surface models

Two separate neural network models were developed for this multi-sensor control experiment. The first model relates the etch rate across the wafer to the process recipe and is similar to the previous case. In this case instead of measurements at only 81 sites on the wafer, 121 sites were chosen to more fully utilize FWI's capability. Whereas in the previous experiment with only two input variables and 81 outputs, only 6 neurons were required in the hidden layer of the neural network. In this experiment with 3 inputs and 121 outputs, on the other hand, 15 neurons are needed to properly model the data.

Optical emission spectroscopy was the second sensor integrated to the dual coil TCP the second model involves this spatially resolved OES sensor. Since the OES sensor provides spectral intensity signals for 3000 optical channels, this would overwhelm the neural network. Instead, the spectra is reduced to principle components (PCs) as discussed in Section 4.2.1. The Matlab Chemometrics Toolbox provides tools to decide on the number of PCs to retain [Kramer.] With **revpca** it was found that 15 principle components were statistically significant as illustrated in Figure 6.2. The reduced number of PCs are then modeled with the neural network as shown in Figure 6.3.

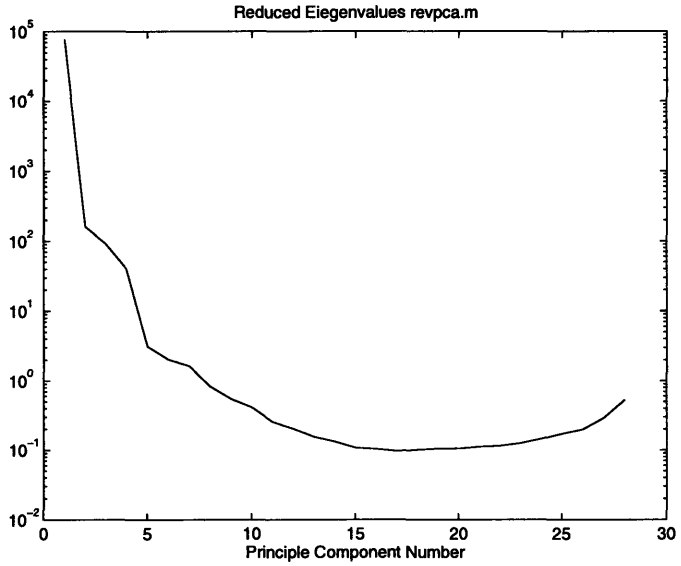

Figure 6.2: The minima in the reduced eigenvalue shown above is 16. Therefore 15 principle components are statistically significant.

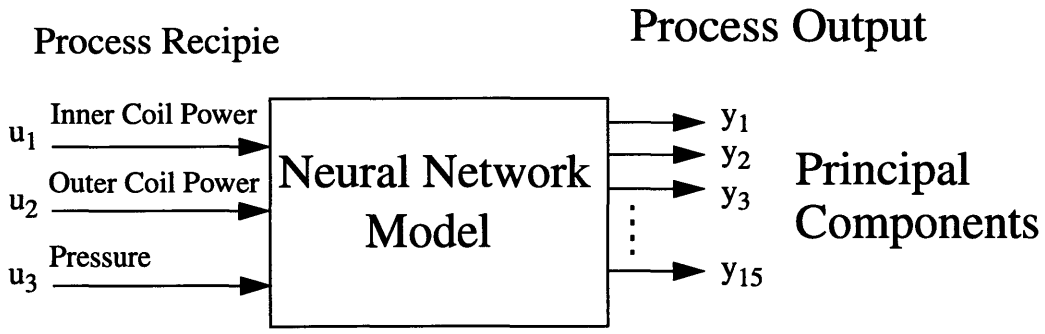

Figure 6.3: The OES model involves in process inputs and the principal components which represent the optical emission spectra.

Though the tests for the significance of the principle components indicate that 15 are statistically significant, the first 3 components captured more than 99.5% of the variance in the optical emission spectra so only those were used in the objective cost function model. The dominance of the first 3 principal components is illustrated in Figure 6.4.

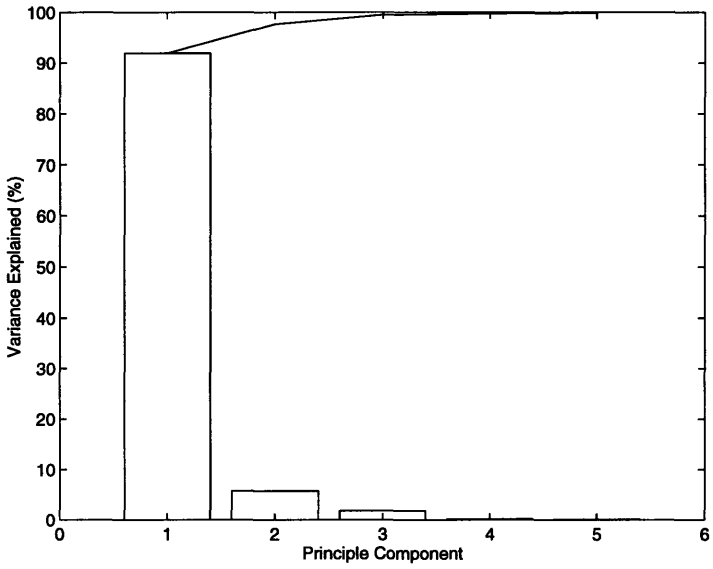

Figure 6.4: The first three principle components capture more that 99.5% of the variance in the spectra.

6.4 Cost function

The multi-objective cost function used for the control of both plasma chemistry and etch uniformity was:

$$min\left(\alpha \cdot \|\hat{z}[n] - \hat{z}[0]\| + \beta \cdot \frac{std(\hat{y}[n])}{(mean(\hat{y}[n]))^2} + \gamma \cdot \|u[n] - u[n-1]\|\right) \tag{6.1}$$

The first term measures the distance that the OES principle components are from the target (first wafer) values. This term ensures that the plasma chemistry is similar to the initial chemistry, as measured by the OES spectra. The second term has been modified from the previous example to weigh the variation in the etch rate variance by the time it takes to clear the wafer (since the variance of the final film thickness is the variance of the etch rate times the duration to clear). The last term has also been modified from the prior example. The cost from the initial setpoint is no longer necessary since the plasma state is

now controlled by the OES system. Therefore, a cost penalty for the incremental change in the process recipe is added. The multi-objective cost function involves signals from two different sensors, possibly in conflict. Appropriate weighting of the three terms in the cost function becomes extremely important. α, β, and γ were chosen from simulations to be 0.4, 5×10^3, and 0.005 respectively, to ensure the magnitude of the OES and FWI components were weighted heavily with an order of magnitude less weight on recipe change penalties.

6.5 Experimental results

The results of a 15 wafer experimental run are shown in Figure 6.5. A disturbance was introduced at wafer number 6. This time the power to the inner coil was reduced by 100 Watts, again a 50% shift in its allowable range. The OES spectra shows a significant shift as represented in a large increase in the OES cost which is shown in Figure 6.5. With this process disturbance, the FWI cost does not significantly change. The controller quickly responds with a combination of adjustments to all the process inputs in order to rapidly reduce the OES cost, and return the process to the controlled state. Figures 6.5 and 6.6 summarizes the results of the experiment.

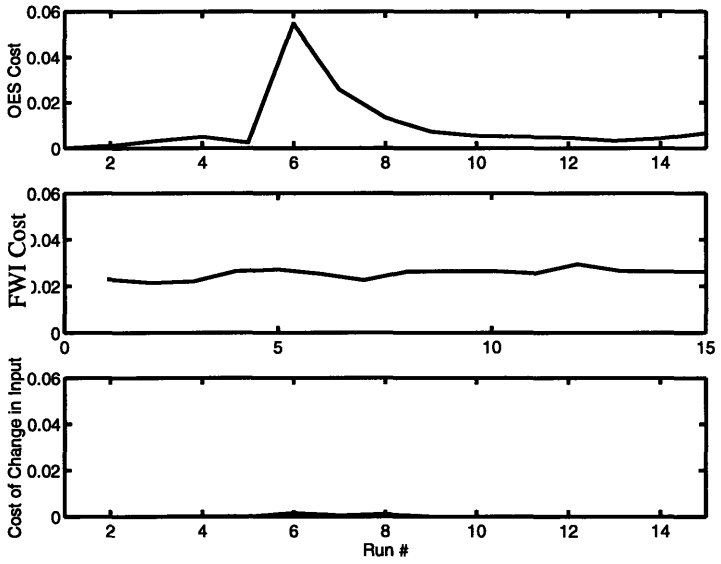

Figure 6.5: Process Outputs show that the OES cost was reduced after the disturbance was introduced. The FWI cost and the cost associated with a change in process recipe were not greatly affected.

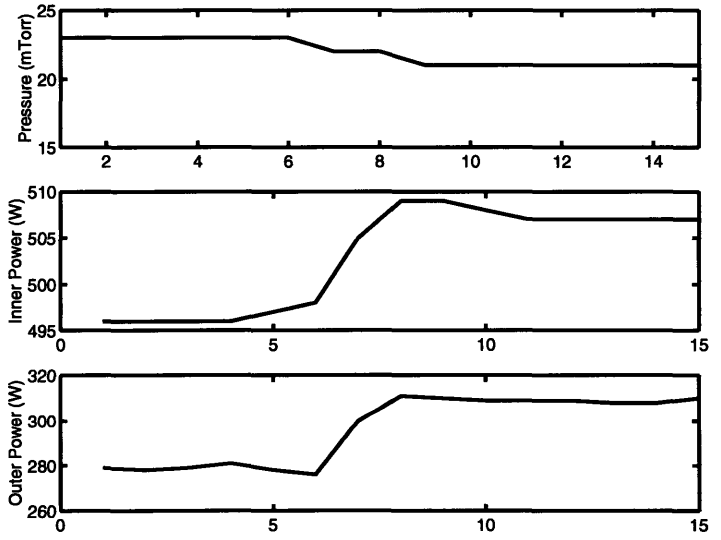

Figure 6.6: Process Inputs show that the controller responded the process disturbance with a recipe modification that was a combination of changes to the pressure, inner power, and outer power.

The results indicate that the controller did reduce the overall cost function after the step disturbance. The models indicated a minima in the cost function at the suggested recipes. The OES cost, however, seems to dominate the recipe decisions as shown in Figure

6.5. The weighting factors between the OES cost and the FWI cost in Equation 6.1 were chosen so that the costs would be on the same order of magnitude. However, the results above indicate that rather it should be the sensitivity of the costs to disturbances that should be equal. When the disturbance was introduced, the OES cost increased much more than the FWI cost. The optimizer found that the best way to reduce the optimal cost function was to minimize the OES cost. This, however, is perhaps not the best way to achieve a desired plasma state and wafer state. More weighting on the FWI cost would solve this problem. The controller performed as expected with the given models and goals and illustrates that the controller is only as good as the models and goals that it is based on.

In addition, the FWI cost did not properly represent the desired results. It was a goal of this experiment to achieve uniform removal of the underlying film, while the first control experiment sought to achieve uniform etching rate. It was believed that since the removal amount is equal to the product of etch time and etch rate, that the proper FWI cost should be:

$$\frac{std(\hat{y})}{(mean(\hat{y}))^2} \tag{6.2}$$

where y, is the etch rate at each site. The extra mean etch rate in the denominator more strongly penalizes slow etch rates.

The total removal thickness can be expressed as shown below:

$$RemovalThickness = EtchTime \times y \tag{6.3}$$

so that the variance in the removal thickness can be expressed as a function of the variance of the etch rate.

$$V(RemovalThickness) = (EtchTime)^2 V(y) \tag{6.4}$$

therefore,

$$V(RemovalThickness) \propto \frac{V(y)}{(mean(y))^2} \tag{6.5}$$

However, since the standard deviation of the etch rate is the square root of the variance,

$$std(RemovalThickness) \propto \frac{std(y)}{mean(y)}, \tag{6.6}$$

thus the mean etch rate in the denominator should only be to the first power, not squared as in Equation 6.1. The extra term in Equation 6.1 reduced the importance of the FWI cost in the overall objective cost function.

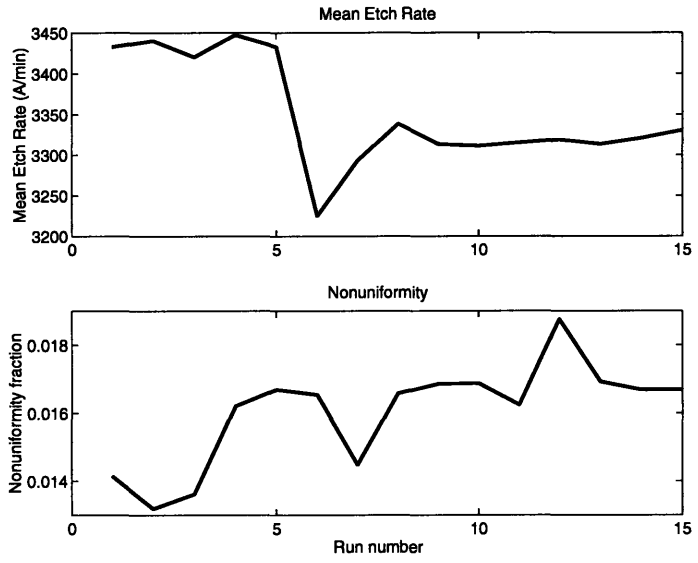

Figure 6.7: The mean etch rate and the etch rate nonuniformity does not recover to initial values after the step disturbance.

The mean etch rate and nonuniformity can be extracted from the control experiment. Minimization of the OES cost did cause the mean etch rate to recover from the step disturbance but the recovery was not complete as shown in Figure 6.7. The nonuniformity as measured by the standard deviation of the etch rate divided by the mean etch rate does not appear to be controlled. It is important to remember that these metrics were not included in the objective cost function so that control of these parameters is not expected.

Despite these concerns, the ANN-EWMA controller did respond to the disturbance based on the models provided. The OES cost was reduced, however, since the FWI cost was so weakly weighted, it provided little guidance to the controller.

Chapter 7

Multivariate analysis of Optical Emission Spectra

A new technique that may be useful in detecting endpoint in plasma etch process is presented in this chapter. This technique is based on Hotelling's T^2 statistic and makes use of the high degree of correlation in the plasma spectra to increase detection sensitivity. The endpoint detection algorithm is demonstrated on two different test cases. The first case involves blanket polysilicon etch and the second much more challenging case is a low open area oxide etch example. The last section in this chapter discusses a methodology to make this endpoint detection algorithm both sensitive and robust.

7.1 Endpoint Detection

Narrow bandpass optical sensors have traditionally been used to detect endpoint in plasma etch. Some detection systems use up to two spectral wavelengths for this purpose. The signal to noise ratio decreases significantly when the layer that is etched is not greatly exposed, i.e. low open area. Multivariate data analysis techniques has been applied to data rich environments in the semiconductor industry [Barna]. These techniques offer potential improvements in the signal to noise ratio. Instead of just a few spectral channels, we propose to use p channels where p is large, on the order of 1000 or more.

In recent years, statistical process control has been applied to real-time tool data and certain sensors such as RF power monitors in an effort to rapidly detect possible faults [Spanos, 1992; Chen, 1995]. In this work, we show that multivariate statistical process control can be applied to OES and used to detect endpoint during the etching process.

7.2 Optical Emission Spectroscopy in Polysilicon Endpoint Detection

The first set of experimental data was collected on the Lam Research TCP modified for dual-coil operation. The Ocean Optics SQ2000 optical emission spectrometer has the

capability for multiple fibers so spatial resolution is possible; only one fiber (the horizontal beam in Figure 7.1) was used for this analysis for simplicity. Polysilicon was etched with an HBr:Cl_2 chemistry at a ratio of 5:1. RF bias power to the lower electrode was 50 Watts and power to the inner and outer inductive coils was 524 W and 302 W respectively. The mean etch rate was on the order of 3200 Angstroms per minute.

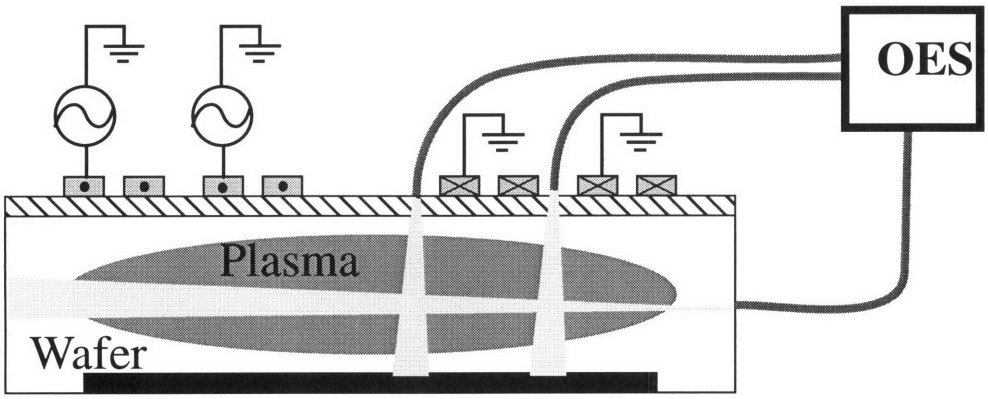

Figure 7.1: A representative diagram of an optical emission spectroscopy system. The data used in this work was collected on a Applied Materials High Density Plasma Oxide etcher as well as a modified Lam TCP Poly etcher shown above.

Spectra were collected during the etch process on the fiber looking in through the side-port. One thousand spectral channels were used and were sampled every 600 milliseconds. A time evolution of the spectra is shown in Figure 7.2. Initially there is a transient due to matching network tuning and bottom power initiation. After a few time slices, the main etch step occurs. As the polysilicon film is etched away and less polysilicon remains, the chemistry of the plasma changes.

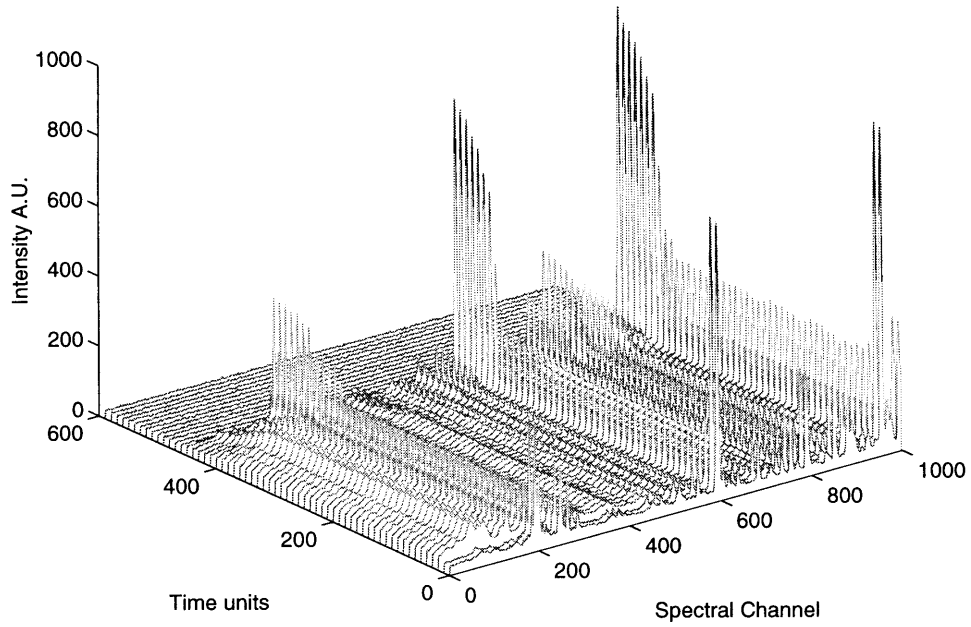

Figure 7.2: Spectra taken during polysilicon etch process. At around t=350, the polysilicon film begins to clear and the plasma chemistry changes as monitored by the change in the spectra.

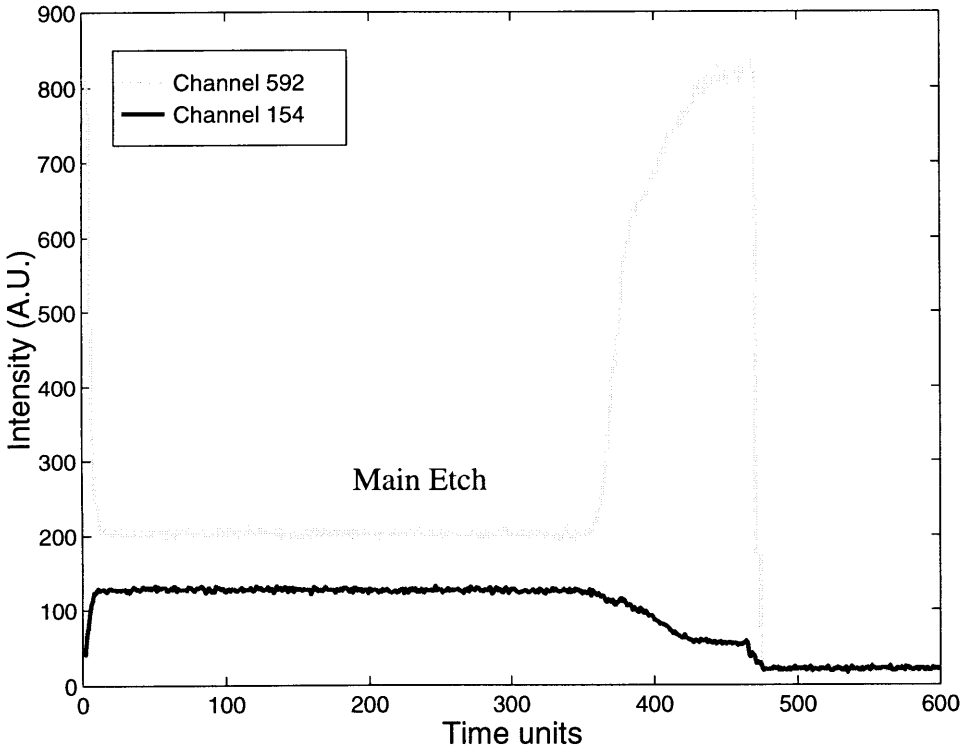

Figure 7.3: Two representative spectral channels exhibit different behavior as endpoint occurs. Channel 592 increases in intensity and channel 154 decreases.

After about t=350 time units, the plasma spectra exhibits a significant change as can be seen in Figures 7.2 and 7.3. This can be attributed to the start of endpoint. Atomic and molecular spectral lines corresponding to by-product species would decrease in intensity when there is no longer any silicon to consume. Likewise, channels corresponding to reactant species lines should exhibit an increase in intensity since they are not longer consumed. Figure 7.3 show two different spectral channels, one increases in intensity while the other deceases at endpoint. Two representative reactions are given below that can explain this phenomena.

$$Si + Cl_2 \rightarrow SiCl_2 \tag{7.1}$$

$$Si + 2Cl_2 \rightarrow SiCl_4 \tag{7.2}$$

In this case, the chlorine lines should increase and the $SiCl_x$ lines decrease. Furthermore, as the density of the various components in the plasma change, interactions between those components will cause more complex changes such as changes in the electron temperature and plasma density. We can take advantage of the high degree of correlation in the spectral channels and the significant changes in the entire spectra when endpoint occurs by applying multivariate analysis techniques to the spectra.

7.3 Hotelling's T^2 Statistic

In Section 7.2, it was demonstrated that there is a high degree of correlation between different spectral lines. During endpoint, some lines increase in intensity while others decrease. Hotelling's T^2 statistic can take advantage of this correlation and produce a single statistic that identifies whether the observed spectra is similar to a model spectra built with historical data.

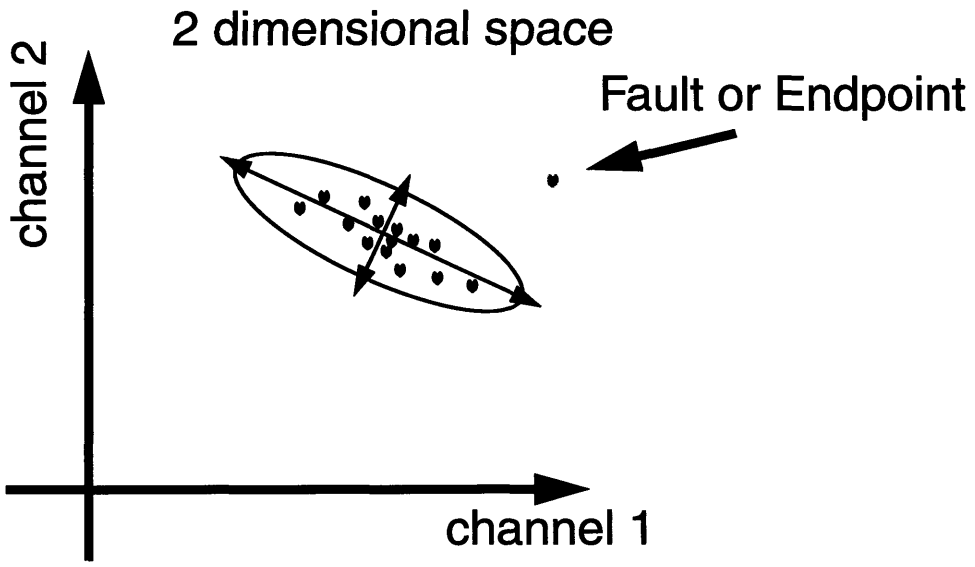

Figure 7.4: The spectra during the main etch step clusters around a local region in *p*-dimensional space where *p* is large. Shown above for 2 dimensional space, the spectra can be bounded by an ellipsoid. Spectra outside this ellipsoid identifies a fault or endpoint.

Since there are *p* spectral channels of data for each time step, we can visualize this as a single vector in *p* dimensional space. The spectra during the main etch step will cluster around a region in this *p*-dimensional space. A hyper-ellipsoid will bound the points. A spectrum that lie outside this hyper-ellipsoid is deemed to be statistically significantly different from the main etch plasma. The Hotelling's T^2 statistic is defined below:

$$T^2 = (\vec{x} - \vec{\bar{x}})S^{-1}(\vec{x} - \vec{\bar{x}})^T \tag{7.3}$$

where $\vec{x}$ represents the p-dimensional spectra. The mean spectra, $\vec{\bar{x}}$, and the covariance matrix, S, are determined from historical data. From the above equation, T^2 is found to be a single value for each time step. It represents a weighted generalized distance from the process mean. The Upper Control Limit (UCL) can be calculated from the F statistic.

$$UCL = \frac{(m-1)(m+1)}{m(m-p)} F_{(1-\alpha, p, m-p)} \tag{7.4}$$

where m is the number of samples in the historical data, and p is the number of spectral channels. The confidence level, α determines the size of the hyperellipsoid. With a large number of spectra collected from historical data, the F distribution can be approximated by a chi-squared distribution.

$$UCL = \chi^2_{\alpha, p} \tag{7.5}$$

For 1000 spectral channels and a confidence level of 99%, the UCL is set to 1107. A T^2 statistic below the UCL is believed to be the main etch while the start of endpoint occurs when the T^2 statistic jumps above the UCL.

7.3.1 Validity of using T^2 statistic

Use of the T^2 statistic assumes that the intensity of each spectral channel during the main etch step is independent and identically distributed about a mean value. This is the IIND assumption. The spectra cannot have any time series behavior during the main etch. Figures 7.5 and 7.6 validate this assumption. The intensity of the spectra has a stable mean. A histogram of the intensity of the spectra during this etch step also shows that it seems to be normally distributed. The autocorrelation plot in Figure 7.6 conclusively shows that the spectral channel does not exhibit time dependent behavior during the main etch step.

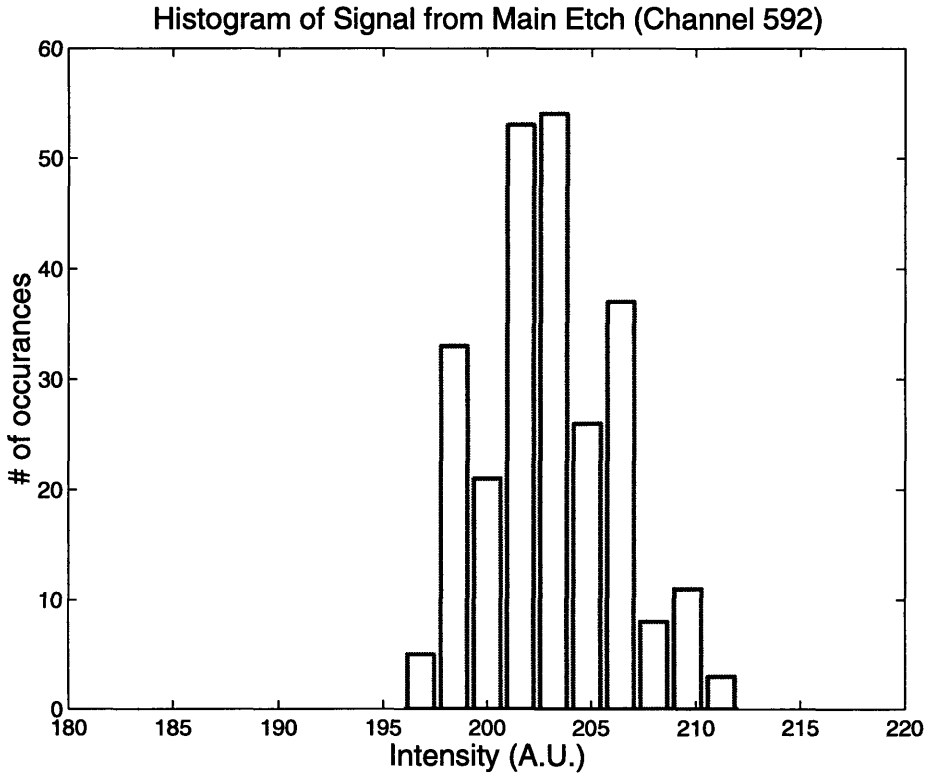

Figure 7.5: Histogram of the intensity of channel 592 during the main etch step. The data follows a Gaussian distribution.

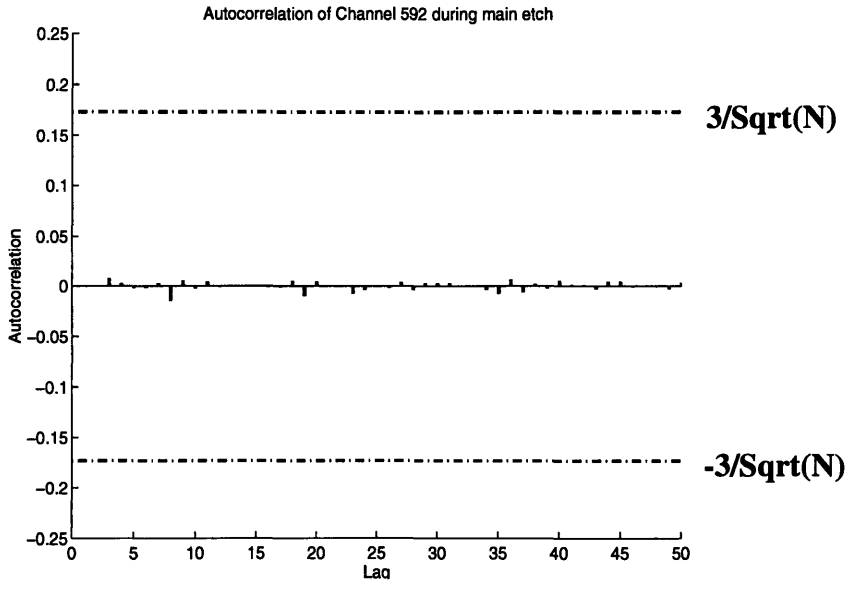

Figure 7.6: The autocorrelation plot above shows that the intensity of the spectral channel does not exhibit time series behavior for the blanket polysilicon etch process. The spectra during the bulk etch is composed of white noise. The 3 sigma limits are not exceeded.

7.4 Results for blanket poly etch

The analysis technique described above was applied to spectra obtained from the etching of blanket polysilicon films. The bulk etch of the first wafer was used to build the mean spectra, $\vec{\bar{x}}$, and the covariance matrix, S. The T^2 statistic was obtained for subsequent spectra. The initial transient that is observed is due to upper and lower matching network tuning and so the initial 50 time units in the spectra are discarded. During the steady state bulk etch, the T^2 statistic is very small and less than the upper control limit as shown in Figure 7.7 and 7.8. However, at the onset of endpoint, the statistic increases greatly and passes above the UCL as illustrated clearly in Figure 7.8.

An estimate of the signal to noise ratio can be made. The signal is taken to be the difference in the T^2 statistic during the etch process and after much of the poly film clears away. The noise is considered to be the standard deviation of the statistic during the bulk etch. With this metric, the S/N ratio is approximately 3000. If only a single optical wavelength was used for endpoint detection, from the trace in Figure 7.3 the signal to noise ratio is estimated to be about 150. The signal to noise ratio has increased by a factor of 20 through use of Hotelling's T^2 with the entire spectra.

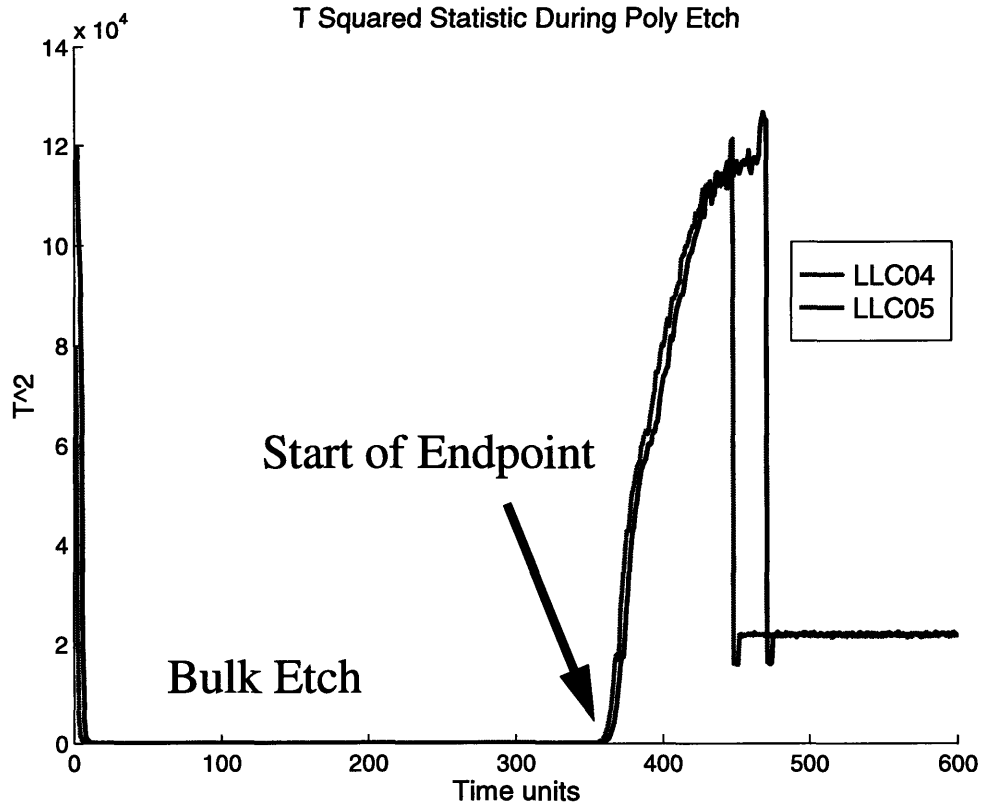

Figure 7.7: The T^2 score is low during the bulk etch step but increases many orders of magnitude at endpoint. The analysis was performed for two wafers.

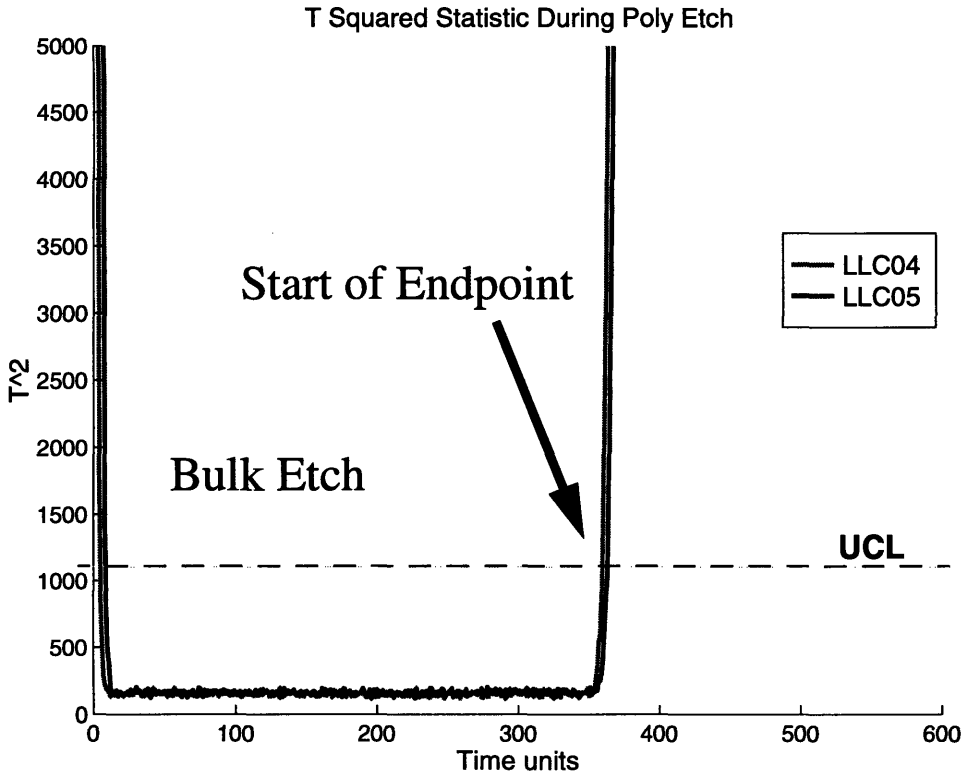

Figure 7.8: A close-up view of the T^2 statistic for blanket poly etch. The dashed line is the Upper Control Limit. The spectra is a very sensitive measure of endpoint.

7.5 Low open area oxide etch

Spectra were collected from an oxide etch process at Digital Semiconductor. The wafers were patterned for contact etch with between 1% and 1.5% open area. Contact is to be made to the gate, source, and drain. The wafers were patterned with a deep UV photoresist. As with the results shown for polysilicon etch above, the low open area oxide etch process was performed on an inductively coupled plasma system, in this case an Applied Materials HDP oxide etcher. The etch was performed in a multistep recipe; Figure 7.9 shows that different plasma chemistries were used. Initially an Argon chemistry was used to strike the plasma. That chemistry was then changed to etch the antireflective coating. The main etch was performed with a C_2F_6 plasma that then changed to an oxygen containing post etch treatment. It is the C_2F_6 plasma that is of interest. The current Digital Semi-

conductor process on this etcher does not use an endpoint detection system. Though many were evaluated on the tool, none met the extreme requirements of reliably detecting endpoint [Dalton, 1997]. This system was especially difficult to diagnose endpoint in because the process kit for the etcher is made of quartz, and the interior chamber wall is thus reactive to the process plasma. The challenge is to detect a change in the plasma emission spectrum when the oxide layer finishes etching even though the chamber liner is etching. The effective signal is reduced by an amount proportional to the amount of the chamber wall that is etched.

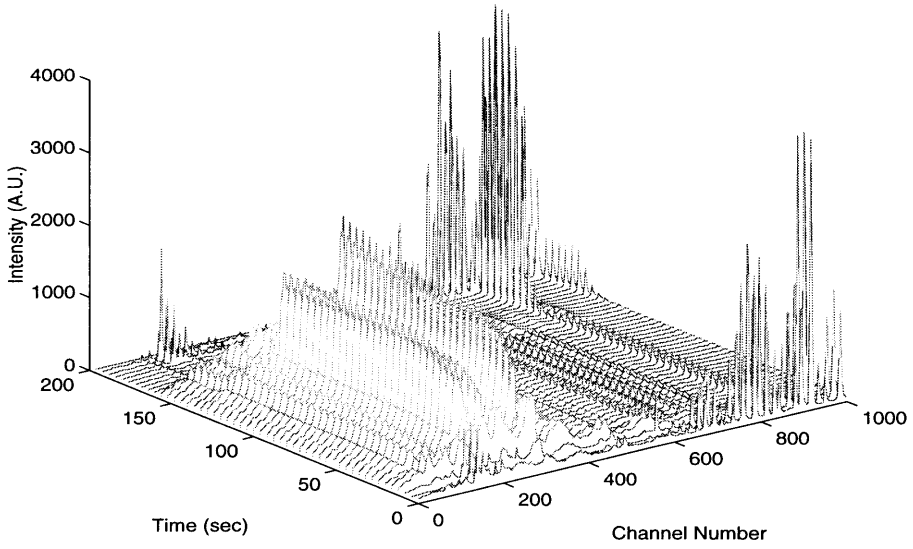

Figure 7.9: The spectra collected during a low open area oxide contact etch does not exhibit the same dramatic change at the start of endpoint as in the polysilicon etch example presented in Section 7.4. There are no discernible features in the spectra that indicate an obvious start of endpoint.

7.5.1 Application of Hotelling's T^2 to Low Open Area Oxide Etch

The same technique applied to the blanket polysilicon etch can also be applied to the optical emission spectrum for contact etch. Using the Ocean Optics SQ2000 spectrometer, spectra were collected at 5 Hertz with an integration time of 15 milliseconds. The spectra in Figure 7.9 shows that there are many channels that have very little information about the plasma. Molecular and atomic emission lines do not lie within these channels. Instead

of the more than 1000 spectral channels, only a reduced number of channels were retained for further analysis. There were 759 channels that exhibited a mean intensity of more than 100 (A.U.) that can be considered more significant than noise.

The covariance matrix was trained with two wafers. The spectra during part of the main etch step was used to build the model. Subsequent spectra were projected onto this model to determine the T^2 statistic.

The results for a set of etches performed on the Applied Materials HDP is presented in Figure 7.10. From Equation 7.5, the 99% confidence interval for the Upper Control Limit (UCL) is a T^2 score of 853. During the bulk of the main etch step, the T^2 score is lower than the UCL. Towards the end of the etch step, the statistic begins a characteristic increase and crosses decisively above the UCL. This transition above the UCL is common in the collected spectra.

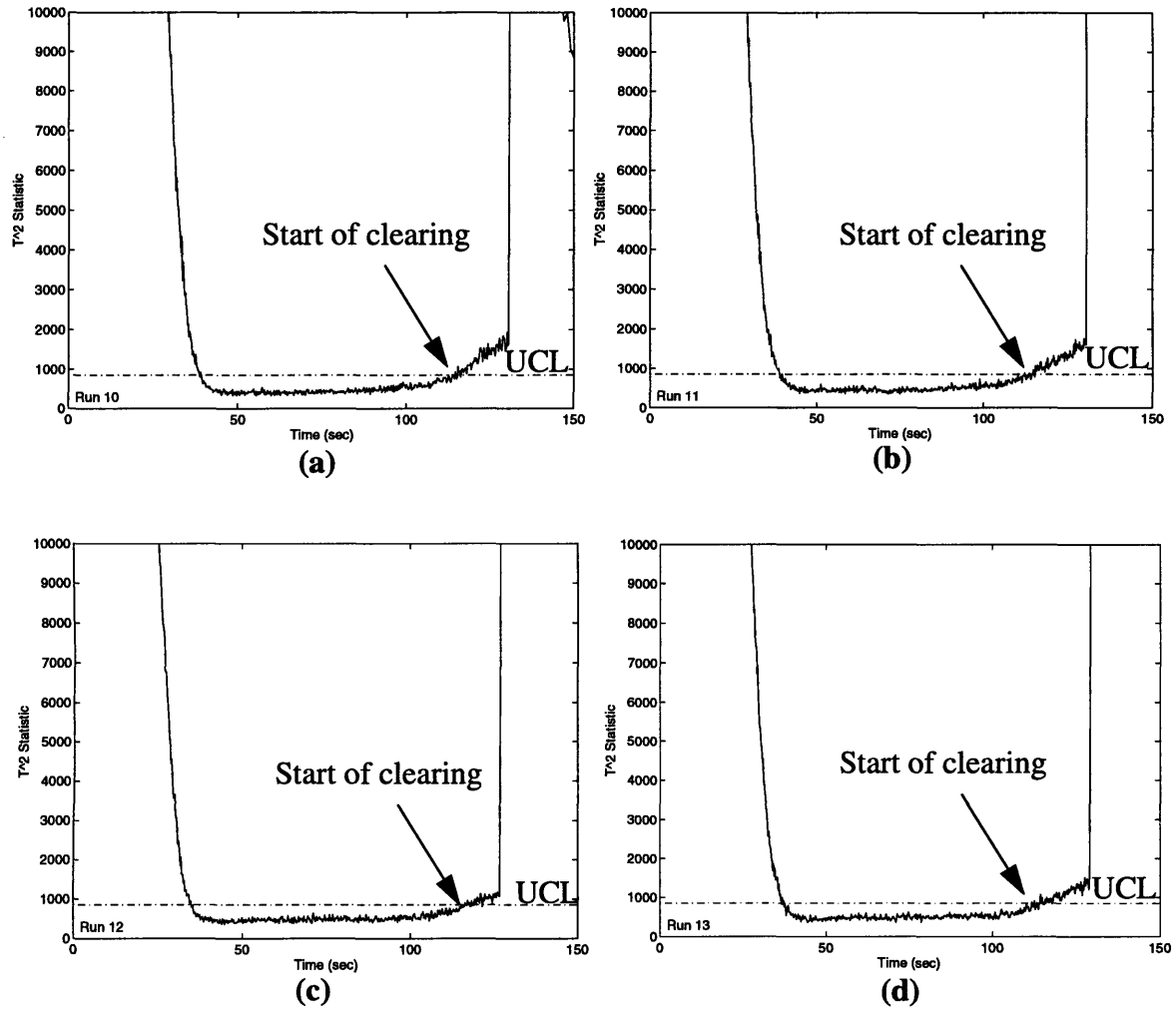

Figure 7.10: The T^2 statistic of the spectra for four different etches shows a transition above the UCL. The statistic is also stable and does not change significantly until towards the end when the start of endpoint is believed to occur.

The results above indicate that a signal that represents a significant change in the plasma chemistry is observed with the aid of Hotelling's T^2 statistic. It is however not conclusive that the change is associated with endpoint. Experiments need to be performed to verify this assumption.

7.6 EWMA and EWMC

Polymer buildup on the window will affect the intensity of the optical emission signal that

is collected by the optical fiber. Furthermore, the chamber seasoning may also cause drifts in the process itself that can manifest as a slow drift in the plasma spectra. In the *p*-dimensional spectra space, the spectra from one run to the next does not cluster right on top of each other. Furthermore, it is also believed that the polymer buildup does not simply act as a neutral density filter [Barna, 1997]. If this were the case, the entire spectra could be scaled by a single value. Since each spectral channel may drift, the interaction between them will change also, so that the covariance matrix is also changing over the time scale of many wafers. A diagram that shows how the spectra mean and covariance is changing over many wafers is shown below in Figure 7.11.

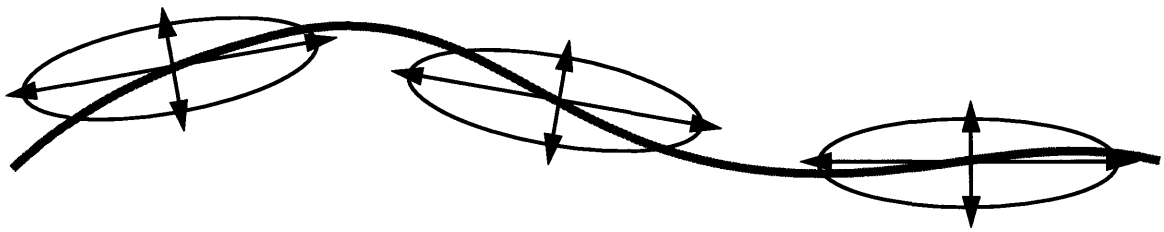

Figure 7.11: The spectra is a very sensitive measure of the processing system. Buildup of material on the optical window will affect the mean signal intensity as represented by the line. The covariance of the spectra will also change over time as represented by the ellipses [White, 1997].

To allow the spectra mean and covariance matrix to evolve in time and track the drifts, it has been suggested to apply an exponentially weighted moving average to the T^2 algorithm [White, 1997]. The mean spectra is updated with the EWMA and the covariance with an Exponentially Weighted Moving Covariance (EWMC).

$$\hat{\bar{x}}[n] = W(\hat{\bar{x}}[n]) + (1 - W)\hat{\bar{x}}[n-1] \tag{7.6}$$

$$W = diag\left(\left[w_1 \ \ldots \ w_m\right]\right) \tag{7.7}$$

where w, are the weights. The weights in the recursive equation allow the mean spectra to change over time. The covariance matrix can be updated in a similar fashion to the

EWMC.

$$S[n] = W(S[n]) + (1 - W)S[n-1] \tag{7.8}$$

and W is again the weights but now is a constant or can be a dense matrix filled with weight values. This technique can be applied to the spectra collected during the etch process to track the slowly changing signals due to chamber seasoning. The model can be rapidly reset after a preventive maintenance to account for process kit changes.

Experiments were not carried out over a long time period so the EWMA and EWMC techniques outlined above were not applied to the spectra. It is assumed that during such a short period of a few minutes, the chamber does not degrade significantly so the spectra is expected to be reproducible.

7.7 Discussion of results

The results presented in this section for the polysilicon etch process are very compelling. The use of the multivariate statistical technique improved the signal to noise ratio in the simple test case. The same techniques were also applied to a system where other endpoint detection systems have failed to detect the faint signal that comes from a low open area oxide etch process. The algorithm presented above shows that a statistically significant change in the plasma spectra occurs towards the end of the process and it is believed that this plasma chemistry change is associated with endpoint.

The algorithm is shown to be sensitive to changes in the spectra. This is possible because it makes use of the high degree of correlation between the many atomic and molecular emission lines. There is a fine balance between sensitivity and robustness. The endpoint detection algorithm needs to be sensitive to endpoint but insensitive to other forms of disturbances. An Exponentially Weighted Moving Average (EWMA) and Expo-

nentially Weighted Moving Covariance (EWMC) allows the model of the spectra to slowly evolve to account for chamber seasoning and other sources of drifts or shifts.

Chapter 8

Summary

8.1 Conclusions

Two key contributions were presented in this thesis to reduce process variation: (1) equipment design for control, and (2) improved endpoint detection.

Process control was introduced as a mechanism to reduce process variation. The primary platform for the control experiments outlined in this thesis was a modified Lam Research TCP. The extensive modifications included installation of two inductive coils instead of a single coil antennae. With the dual coil TCP, the etch rate profile could be effectively controlled. Diagnostic systems were also installed on the dual coil TCP and included an in situ Full Wafer Interferometry sensor and a spatially resolved Optical Emission Spectroscopy system. These two systems provided information about the wafer state and plasma state respectively. This information was interpreted by a run to run model based process control algorithm that made necessary perturbations to the process recipe to achieve the desired objective of maintaining a stable high performance process. The results obtained indicate that the process control strategy that was implemented can be used to make process corrections when disturbances are measured by the sensors.

The second method presented to potentially reduce etch process variation was endpoint detection. A sensitive and robust method to detect the slight change in plasma chemistry that occurs at endpoint is demonstrated in two extreme cases. The first trivial case of blanket polysilicon etch shows that the Hotelling's T^2 statistic applied to a full spectrum can improve the signal to noise ratio by a factor of 20 when compared to using just a single spectral channel. The technique was also applied to a more realistic industrial oxide etch process in which low open area production wafers were processed. In this case, end-

point is much more difficult to detect but the algorithm was able to signal a significant change in the plasma chemistry that is believed to be associated with endpoint.

8.2 Future Work

As with most scientific endeavors, this thesis opens up more questions to be answered. The dual coil work thus far has only been performed on a blanket polysilicon wafers. Other issues will arise when patterned wafers are examined. Of primary importance will be the signal integrity of the various sensor systems. Another question that must be investigated is the effect of the various process recipes on the microscopic etching characteristics. The goal of process control is not to just ensure that the etching rate across the wafer as measured by FWI is uniform, but also to ensure that the various features within a die etch uniformly as well. The question that must be answered is whether FWI information is a good proxy for the etching behavior, both macroscopic and microscopic. The OES system merits further work also. The experiments were performed over a brief period of time, usually no longer than 20 wafers at a time. Chamber seasoning is not expected to significantly affect the OES signal on a run-to-run basis in these well controlled experiments. However, in more realistic industrial settings where a sensor is active for extended periods of time, mechanisms need to be investigated to allow the OES model to adapt slowly to optical window fogging effects yet be sensitive to other forms of disturbances. In other words, the issue of sensitivity and robustness arises again.

The work on endpoint detection presents some exciting opportunities. However, there are still some issues that need to be resolved. The signal that represents a statistically significant change in plasma chemistry is just that. We infer that this change is associated with endpoint but cannot prove this unless further experiments are performed. By using a split lot with thinner oxide layers, it is possible to show that there is correlation between endpoint and the signal observed with the endpoint detection algorithm. Cross sectional

Scanning Electron Micrographs can also show that the signal correctly identifies the start of endpoint. The issues of sensitivity and robustness are important here as they are elsewhere: the detection algorithm needs to be sensitive and robust at the same time. Testing the algorithm over extended periods in a production environment would help verify the utility of the EWMA and EWMC. Finally, time series behavior of the spectral signals needs to be investigated to examine the statistical validity of applying the algorithm.

References

Barna, Gabe, Private communication, May 1997.

Boning, A. Hurwitz, J. Moyne, W. Moyne, S. Shellman, T. Smith, J. Taylor, and R. Telfeyan, "Run by Run Control of Chemical Mechanical Polishing," IEEE Trans. on Components, Packaging, and Manufacturing Technology, Vol. 19, No. 4, pp. 307-314, Oct. 1996.

Butler, S.W., J. Stefani, M. Sullivan, S. Maung, G. Barna, and S. Henck, "Intelligent model-based control system employing *in situ* ellipsometry," *J. Vac. Sci. Technol. A,* 12(4) Jul./Aug. 1994.

Chapman, Glow Discharge Processes: Sputtering and Plasma Etching, John Wiley & Sons, New York, 1980.

Chen, R., "Real-Time SPC for Plasma Etching Using Optical Emission Spectroscopy," UC Berkeley course EE290W S95 project, 1995.

Dalton, T., "Pattern Dependencies in the Plasma Etching of Polysilicon," Ph.D. Thesis, Massachusetts Institute of Technology, Dept. of Chemical Engineering, 1994.

Dalton, T., Private communication, June 1997.

Demuth, H., and M. Beale, "Neural Network Toolbox: For Use with MATLAB," The Mathworks Inc., 1995.

Grace, A., "Optimization Toolbox: For Use with MATLAB," The Mathworks Inc. 1994.

Goodlin, B., ANOVA analysis performed by Goodlin, June 1997.

Gower, A., "An architecture for flexible distributed experimentation and control with an AME 5000 plasma etcher," S.M. Thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 1996.

Guo, R.S., and E. Sachs, "Modeling, Optimization and Control of Spatial Uniformity in Manufacturing Processes," *IEEE Trans. Semi. Manuf.*, Vol. 6, No. 1, Feb. 1993.

Hardt, D., "Manufacturing Processes and Process Control," Notes for MIT Course 2.830, Control of Manufacturing, Spring Semester 1994.

Hecht, E., *Optics*, Addison-Wesley Publishing Company, Massachusetts, 1990.

Ingolfsson, A., and E. Sachs, "Stability and Sensitivity of an EWMA Controller," *Journal of Quality Technology*, Vol. 25, No. 4, Oct. 1993, p. 271.

Le, M., T. Smith, D. Boning, and H. Sawin, "Run to Run Model Based Process Control on a Dual Coil Transformer Coupled Plasma Etcher," The 191st Meeting of the Electrochemical Society, Montreal, Canada, May 5th, 1997.

Le, M., T. Smith, D. Boning, and H. Sawin, "Run by Run Uniformity Control on a Dual Coil Transformer Coupled Plasma Reactor with Full Wafer Interferometry," 43rd National Symposium of the American Vacuum Society, Philadelphia, PA, Oct. 1996.

Litvak, H., "Endpoint control via optical emission spectroscopy," *J. Vac. Sci. Technol. B,* 14(1) Jan./Feb. 1996.

Munsat, T., W.M. Hooke, S.P. Bozeman, and S. Washburn, "Two new planar coil designs for a high pressure radio frequency plasma source," *Appl. Phys. Lett.,* 68(11) March 1996.

Paranjpe, A.P., "Modeling an inductively coupled plasma source," *J. Vac. Sci. Technol. A,* 12(4) Jul./Aug. 1994.

Rietman, E.A., "Neural networks in plasma processing," *J. Vac. Sci. Technol. B,* 14(1) Jan./ Feb. 1996.

Sachs, E., A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Trans. Semi. Manuf.*, Vol. 8, No. 1, Feb. 1995.

Schaper, C., M. Moslehi, K. Saraswat, and T. Kailath, "Control of MMST RTP: repeatability, Uniformity, and Integration for Flexible Manufacturing," *IEEE Trans. Semi. Manuf.*, Vol. 7, No. 2, May 1994.

Smith, T., and D. Boning, "An Artificial Neural Network EWMA Controller for Semiconductor Processes," 43rd National Symposium of the American Vacuum Society, Philadelphia, PA, Oct. 1996.

Smith, T., J. Stefani, D. Boning, and S. Butler, "Run By Run Advanced Process Control of Metal Sputter Deposition," The 191st Meeting of the Electrochemical Society, Montreal, Canada, May 5th, 1997.

Spanos, C.J., "Statistical Process Control in Semiconductor Manufacturing," *Proceedings of the IEEE,* Vol. 80, No. 6, June 1992.

Spanos, C.J., H.F. Guo, A. Miller, and J. Levine-Parrill, "Real-time Statistical Process Control Using Tool Data," IEEE Trans. Semi. Manuf., Vol. 5, No. 4, Nov. 1992, pp. 308-318.

Suh, N., *The Principles of Design,* Oxford University Press, New York, 1990.

Ventzek, P.L.G., N. Yamada, Y. Sakai, and H. Tagashira, "Simulations of real-time control of two-dimensional features in inductively coupled plasma sources for etching applications," *J. Vac. Sci. Technol. A,* 13(5) Sept./Oct. 1995.

Ventzek, P.L.G., R.J. Hoekstra, and M.J. Kushner, "Two-dimensional modeling of high density inductively coupled sources for materials processing," *J. Vac. Sci. Technol. B,* 12(1) Jan./Feb. 1994.

White, D.A., "In-situ wafer uniformity estimation using principal component analysis and function approximation methods," Masters Thesis, Massachusetts Institute of technology, Dept. of Electrical Engineering and Computer Science, 1995.

White, D., G.G. Barna, S.W. Butler, B. Wise, and N. Gallagher, "Methodology for Robust and Sensitive Fault Detection," 191st Meeting of the Electrochemical Society, Montreal, Canada May 1997.

Wong, K.S., "Real-time analysis and control of plasma etching via full wafer interferometry," Ph.D. Thesis, Massachusetts Institute of Technology, Dept. of

Mechanical Engineering, 1996.

Yamada, N., P.L.G. Ventzek, H. Date, Y. Sakai, and H. Tagashira, Model for large area multi-frequency multiplanar coil inductively coupled plasma source," *J. Vac. Sci. Technol. A*, 14(5), Sept./Oct. 1996.

Yasaka, Y. and T. Nakamura, "Control of process uniformity by using electron cyclotron resonance plasma produced by multiannular antenna," *Appl. Phys. Lett.,* 68(11) March 1996.