RUN TO RUN CONTROL IN
SEMICONDUCTOR MANUFACTURING

by

John Musacchio

Memorandum No. UCB/ERL M98/79

13 November 1998

COVER

RUN TO RUN CONTROL IN
SEMICONDUCTOR MANUFACTURING

by

John Musacchio

ELECTRONICS RESEARCH LABORATORY

College of Engineering
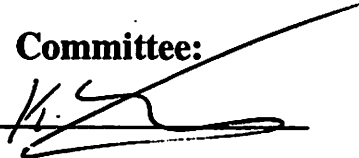University of California, Berkeley
94720

Run to Run Control in Semiconductor Manufacturing
by John Musacchio

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II.**

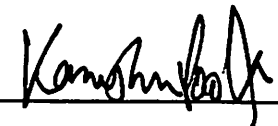Approval for the Report and Comprehensive Examination:

Committee:

Professor Costas J. Spanos

Research Advisor

11/13/98

(Date)

****

Professor Kameshwar Poolla

Second Reader

11/22/98

(Date)

Abstract

RtR Control in Semiconductor Manufacturing

by
John Musacchio

Electrical Engineering and Computer Sciences
University of California at Berkeley

Professor Costas J. Spanos, Advisor

Run to Run (RtR) control uses data from past process runs to adjust settings for the next run. By making better use of existing in-line metrology and actuation capabilities, RtR control offers the potential of reducing variability in semiconductor manufacturing with minimal capital cost.

This thesis reviews the basic techniques of RtR control, and connects these techniques to concepts in estimation theory. New RtR control techniques are developed, and the advantages of both the new and existing techniques is discussed. The thesis concludes with a case study of RtR control for DUV lithography. In this study, a RtR control architecture is developed using novel metrology, and its performance is evaluated via simulation.

1

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1. Background and Motivation

As integrated circuit producers are driven toward finer linewidths and feature sizes, there is a compelling need for the reduction of variability in semiconductor manufacturing processes.

Almost always, this need has been met by expending considerable effort in the design of processes that are very stable, by isolating environmental effects and designing processes that are insensitive to drifts in equipment performance and incoming wafer characteristics. Processes are then run with fixed recipes, and only occasionally are retuned by processing wafers with test patterns.

An alternative approach, and one that is receiving increasing attention in academia as well as industry, is the use of feedback control to reduce product variability. Various processes have been studied in this context. See for example Rapid Thermal Processing (RTP) [1], Reactive Ion Etching [2], and I-line lithography [3].

Feedback control uses measurements during processing to adjust recipe settings to counteract process drift. This requires a rudimentary process model, metrology, and actuation capability. In RtR control, recipe settings are adjusted for a given wafer based on the measurements taken from previous wafers. This approach does not require real time actuation, is minimally intrusive to the process, and can use existing in-line metrology; though some processes may require additional in-line metrology for the use of RtR control.

One of the most compelling studies of RTR control was performed at a Motorola microprocessor manufacturing facility [4]. The facility implemented RTR control in the

lithography sequence and significantly reduced CD variability. The reduction allowed Motorola to target a smaller CD without jeopardizing yield. As a result of the performance improvement from smaller linewidths, more circuits tested into the more lucrative, high speed bins.The enhanced revenue from producing faster processors offset the $600 thousand initial cost of control implementation in just a few days of operation.

Other encouraging studies of the utility of RtR control have been performed for chemical mechanical polishing (CMP) [5], and silicon epitaxy [6].

1.2. Report Objective

There are many ways one can design a RtR controller, and indeed many different types of RtR control algorithms have been developed and implemented both in industry and academia. The algorithms can be divided into three broad classes, those rooted in Estimation theory, Statistics, and Artificial Intelligence. In this work, we focus on methods based in estimation theory, partly because these methods are most amenable to a general analysis whereas methods of the other broad classes must usually be studied on a case by case basis.

In this work, we attempt to understand the theory underlying estimation theory based algorithms, explore the similarities and differences between methods, and conclude with some insights on how one might design and analyze a RtR control method for a particular problem. Our study is conducted using analysis based in estimation theory, simulation, and experiment.

1.3. Organization

We begin by reviewing the types of process models often used for RtR control, and also develop models for how the behavior of processes drift over time. Next, in Chapter 3, we survey the principles of estimating random process, as this theory is essential in understanding the connections and differences of RtR control methods. In Chapter 4, a survey of existing RtR control methods is presented, focusing on how each method is rooted in basic estimation theory. In chapter 5 we study RtR control in the context of a specific example, DUV Lithography. We examine previous work in controllers implemented on a lot to lot time scale, and then develop, using experimental results, a strategy for RtR control at a

wafer to wafer time scale. Via simulation we analyze our design. Finally Chapter 6 summarizes our findings from our simulation and experimental study, makes some remarks about the design of a RtR control strategy, and outlines directions for future research.

8

Chapter 2 Models

RtR control requires a model of how the output of process is related to inputs, which can include process settings and incoming wafer characteristics. Often it is not necessary to have an extremely accurate or detailed model. Control strategies involve making modest adjustments to input settings to counteract drifts in process behavior. Consequently, first-order sensitivies are all that is required for control.

In this section, we outline the basic form of process model that will be used in our study of RtR control. A DUV lithography model will serve as a running example to illustrate our development. Just as important as process models in RtR control, drift models are developed, and again DUV lithography will serve as an example. As will be demonstrated later in this report, the behavior of the process drift is the key factor determining the applicability of a RtR control method to a process. A schematic illustrating the structure of our process and drift models is shown in Figure 2.1

2.1. Inputs

There are two types of inputs that we will refer to in the context of RtR control. The first type of input, a *control input*, is an equipment setting that we plan to adjust either on a wafer to wafer or lot to lot basis. In DUV lithography RtR control, one might consider the exposure dose as a control input. The second type of input includes environmental factors or incoming wafer characteristics that are both measured either on a wafer to wafer or lot to lot basis, and whose effect we wish to cancel by adjustments in other inputs. For example, the reflectance of an incoming wafer may be considered as an input in DUV lithography, and its value used to adjust the exposure dose. Inputs of this second type we will refer

to as *feed-forward inputs*. The input vector containing both types of inputs refer to as simply *"the input."*
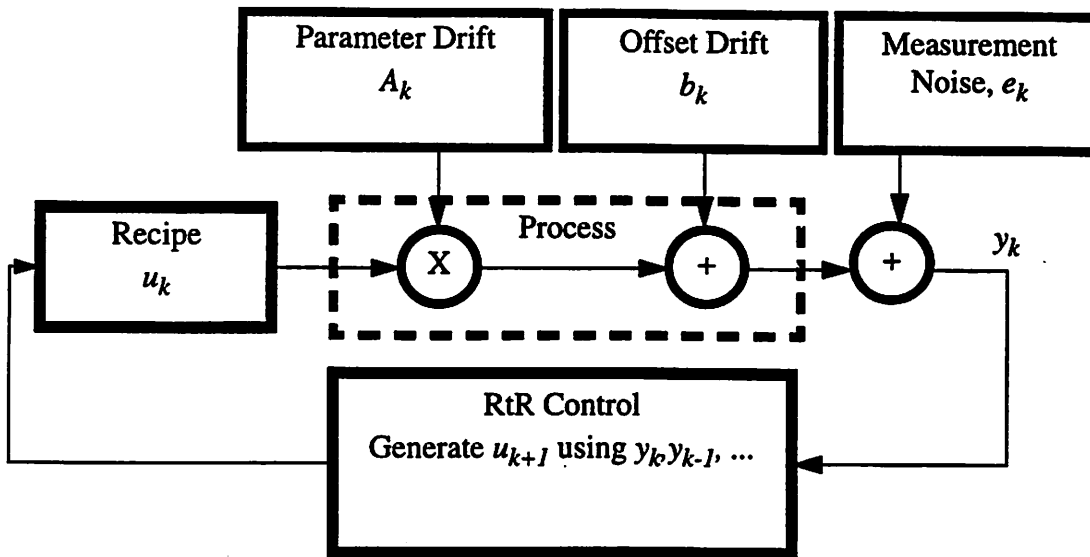
Figure 2.1. Conceptual Schematic of RtR control.

2.2. Outputs

The output of a process is the wafer characteristic(s) that we wish to regulate with RtR control. In DUV lithography, the output might be critical dimension (CD). Sometimes we will make the distinction between the *true output* -- the actual value of the wafer characteristic, and the *measured output* -- the value measured which may include some amount of measurement error.

2.3. Process Model

The process model is the relationship between the input and the output. For the purposes of RtR control, The process model for should be relatively simple and should only consider a modest number of input and outputs. For example a complicated FEM model for simulating photoresist profiles would be too complicated to use in RtR control, but perhaps could be used to help build a simpler relationship between CD and a few selected control inputs like exposure dose and focus offset.

Ideally, the form of the process model should be based on a physical understanding of the process, and the parameters of the model obtained by fitting it to experimental data. In almost all cases, the model one arrives at is affine, and can be put into the following form.

$$y_k^o = A_k u_k + b_k \tag{2.1}$$

Here k is the wafer index, u_k the process input (both adjustable and feed-forward), and y_k^o the true output. Throughout this report, lower case letters will denote vectors and uppercase letters matrices. As is the case in nonlinear regression models, u_k may contain nonlinear functions of the original inputs. When the nonlinearities are invertible, no generality is lost in using form (2.1).

The input - output relationship can vary over time. For example, the ageing of a stepper's light source may change the required exposure setting for achieving a certain linewidth. One can account for this drift in behavior by making the model parameters A_k and b_k functions of k. When the term A_k varies with k, we say the process undergoes *parameter drift*. If the b_k term varies in time, the process is undergoing *offset drift*. It is important to note that a process undergoing offset drift alone maintains a constant sensitivity to its input.

Measurement noise is an important issue in the study of RtR control. We model the noise encountered in measuring y_k^o as

$$y_k = y_k^o + e_k \tag{2.2}$$

where e_k is a random variable. Usually we will assume e_k is white noise (independent from one wafer to the next, has a Gaussian distribution, and has zero mean.)

2.4. Drift Models

As is discussed in Section 2.3, the input-output relationship of a semiconductor manufacturing process can drift over time, and the drift can be modeled by allowing A_k and b_k of Equation (2.1) to vary as a function of the time. Recall that drift in A_k is called parameter drift, while drift in b_k is called offset drift. A model for this drift should capture both the variance, as well as the amount of autocorrelation in each drifting term. The most convenient method for modeling the drift is to use a discrete stochastic process.

Obtaining a good drift model can be challenging. One can not measure every term of A_k and b_k at every run, making it difficult to compute the drift statistics. However, by performing multiple modeling experiments, and applying some physical understanding of the sources of process variability, one often can approximate the variance and autocorrelation of the drift.

2.4.1. An Example of Drift Modeling

Suppose we are modeling the drift of a DUV lithography process. There are two inputs: exposure dose, u_{1k}, and reflectance of the incoming wafer, u_{2k}, and one output, y_k, pre-etch CD. Suppose our beginning model is [7]

$$\left[A_0 \middle| b_0\right] = \left[-0.0395\frac{\mu m}{mJ} \quad 0.16\mu m \middle| \quad 0.553\,\mu m\right] \tag{2.3}$$

We will assume that the values of the A_k and b_k coefficients drift randomly, with mean equal to the their initial values defined in (2.3). From historical data, we expect the sensitivity to exposure dose, and the sensitivity to reflectance to vary with a standard deviation of $0.0002\,\frac{\mu m}{mJ}$. and $0.002\,\mu m$ respectively. The offset term, b_k we anticipate will vary with a standard deviation of $0.0005\,\mu m$. Finally, we anticipate the drift in each of these three terms to not have any cross correlations. We define a random vector of deviations,

$$\Phi_k = \left[A_k \middle| b_k\right]^T - \left[A_0 \middle| b_0\right]^T \tag{2.4}$$

which has a desired covariance of,

$$K_\Phi = \begin{bmatrix} 0.0002 & 0 & 0 \\ 0 & 0.002 & 0 \\ 0 & 0 & 0.0005 \end{bmatrix}^2 \tag{2.5}$$

Now suppose we believe that the sensitivity to exposure dose and the offset term will vary slowly -- with the correlation coefficient between successive runs being about 0.98 for each term. The autocorrelation of the sensitivity to reflectance we expect to be less, say 0.90. Using this data, we define the discrete stochastic process

$$\Phi_{k+1} = F\Phi_k + Gw_k \tag{2.6}$$

where w_k is a random vector of dimension 3 with identity covariance, and F is defined to be:

$$F = \begin{bmatrix} 0.98 & 0 & 0 \\ 0 & 0.90 & 0 \\ 0 & 0 & 0.98 \end{bmatrix} \qquad (2.7)$$

Equation (2.6) says that each drift term is 0.98 or 0.90 times the previous drift term plus a random innovation. G determines the size of the innovation at each time step. As long as F is a stable matrix (all eigenvalues inside the unit circle), The covariance of Φ_k will assymptotically approach a fixed covariance matrix. We can force this limiting covariance matrix to be our "desired" covariance K_Φ, by choosing G to be:

$$G = \sqrt{K_\Phi - FK_\Phi F^*} \qquad (2.8)$$

Finally, If we initialize the recursion (2.6) by defining Φ_0 to be a random vector with covariance K_Φ, Φ_k will have the covariance matrix K_Φ independent of k.

Though it is impossible to create a stochastic model that perfectly describes the drift in a real process, a stochastic drift model that captures at least the first order statistics of the drift will prove to be extremely useful in RtR control design. Stochastic drift models will also be vital in understanding the theory underlying RtR control methods, and for testing designs in simulation.

2.4.2. Exclusively Offset Drift Models

As we saw in the previous example, developing a process drift model that includes both offset and parameter drift can involve a lot of guess work. The problem of drift modeling becomes significantly easier to base on historical data when one models all of the process drift as a change in the offset term alone. Though the process drift may not truly be restricted to the offset term, we will show in Chapter 4 that under certain conditions on the process and controller, the RtR control performance is insensitive to whether one assigns process drifts to the b_k or A_k terms.

Suppose we have a historical data record containing a series of process inputs $\{u_k\}$ and process outputs $\{y_k\}$. Then to build the drift model, simply compute the residuals $b_k = y_k - A\, u_k$ where A is from our process model and is fixed. Now the problem is to model the

13

signal b_k as coming from a linear system driven by white noise. This is a standard problem in signal processing and system identification and there are a wide range of techniques [8] [9]. One of the most straight forward methods of building the drift model would be to find an nth order Auto Regressive (AR) model by regression on the b_k data.

Chapter 3 Estimation Principles

Almost all RtR control algorithms recursively estimate parameters in a process model, and then invert this model to find a recipe that should produce the desired output. Therefore, estimation theory is essential for a complete understanding of RtR control. In this chapter, we review the principles of estimation.

3.1. Linear Least Squares Estimation

To avoid confusion between a random variable and a particular sample value, we denote random variables with bold face letters. This distinction is suspended in subsequent chapters.

Suppose we have two correlated random vectors x and y. We observe a sample value of y, denoted y, and with this information we would like to make an estimate of x, which we will call $\hat{x}(y)$. In linear least squares estimation, $\hat{x}(y)$ is restricted to be an affine function of y, which has form:

$$\hat{x}(y) = Hy + g \tag{3.1}$$

The values of H and g are determined by first defining a cost function to be the expected Euclidean distance between the estimate $\hat{x}(y)$ and x.

$$COST = E[|\hat{x}(y) - x|^2] \tag{3.2}$$

Note that in (3.2) we regard the estimate itself to be a random variable, because it is a function of the random variable y. The values of H and g used are those that minimize the above cost function, and can be found by differentiation of (3.2) [11]. After one carys out this procedure, the linear least squares estimate (LLSE) is found to be

$$\hat{x}(y) = K_{xy}K_y^{-1}(y - m_y) + m_x \qquad (3.3)$$

where K_y is the covariance of y, m_x & m_y are the means of x and y, and K_{xy} is the cross-covariance of x and y, $E[(x - m_y)(y - m_y)^*]$.

It can be shown that when x and y are jointly gaussian, [3.3] is the best estimator in terms of cost function (3.2) among all forms of estimators, affine or otherwise [12]. In this case, the estimate $\hat{x}(y)$ is equal to the conditional expectation, $E[x \mid y]$.

Whether x and y are jointly gaussian or not, the covariance of the estimation error, $E[|\hat{x}(y) - x|^2]$ has the value:

$$K_{xx} - K_{xy}K_y^{-1}K_{yx} \qquad (3.4)$$

When x, and y are jointly gaussian, the covariance of the estimation error conditioned on a particular observed sample value, is the same as the unconditioned estimation error covariance.

$$E[|E[x|y] - x|^2 \big| y = y] = K_{xx} - K_{xy}K_y^{-1}K_{yx} \qquad (3.5)$$

However, in the general case where when x and y are not jointly gaussian, the covariance of the estimation error conditioned on sample value y will depend on which value of y is observed.

3.2. Wiener Filtering

The problem formulation of the Wiener filter begins by assuming that we have two jointly wide sense stationary random processes $\{y_k\}$ and $\{x_k\}$. Recall that wide sense stationary processes have means that are independent of time, and autocorrelations that only depend on the time lag. A pair of jointly wide sense stationary processes are individually wide sense stationary and have a crosscorrelation function that depends only on the time lag. The Wiener Filtering problem is to find an LTI filter with impulse response $\{h_k\}$, that minimizes the cost function

$$E\left[\left|\left(\sum_{n=-\infty}^{\infty} h_n y_{k-n}\right) - x_k\right|^2\right] \qquad (3.6)$$

The solution is obtained by solving the Wiener-Hopf equations

$$r_{xy}(i) = \sum_{k=-\infty}^{\infty} h_k r_y(i-k) \tag{3.7}$$

where r_y is the autocorrelation of the process $\{y_k\}$ and $r_{xy}(m)$ is the crosscorrelation of $\{x_k\}$ with $\{y_k\}$, defined as $E[x(n)\,y(n\text{-}m)^*]$. (3.7) can be derived by taking complex gradients of the cost function and setting them to zero. Obviously a filter used for on-line estimation must be causal, so we must add the restriction that h_k vanishes for $k < 0$. With this restriction, the Wiener-Hopf equations become

$$r_{xy}(i) = \sum_{k=0}^{\infty} h_k r_y(i-k) \tag{3.8}$$

For many problems, (3.8) has no closed form solution. However for the problems we will be applying (3.8) to, we will be able to find solutions using z-transform techniques.

3.2.1. Causal Wiener Filtering based on Power Spectra

Suppose the power spectrum of $\{y_k\}$ can be expressed as a rational z-transform $Sy(z)$. Then the $Sy(z)$ can be factored into the following form [13]:

$$S_y(z) = \sigma^2_v G(z) G^*\left(\frac{1}{z^*}\right) \tag{3.9}$$

where $G(z)$ is causal, stable, minimum phase (has all zeros in the unit circle) and monic (has impulse response with $g(0) = 1$). $G(z)$ has the interpretation of being a synthesis filter, because the process $\{y_k\}$ may be created by passing white noise with variance σ^2_v through $G(z)$. The white noise used as the input to the synthesis filter is called the innovations process, because each sample of the innovations process represents the new "information" about the $\{y_k\}$ process that the synthesis filter could not infer from past information. Thus, σ^2_v is called the power of the innovations process.

Now suppose we are given $\{y_k\}$ and wish to find the innovations process, which we call $\{v_k\}$. This may be accomplished simply by passing $\{y_k\}$ through the filter, $G^{-1}(z)$, which we denote as $Q(z)$. Though $\{v_k\}$ is white noise, $\{v_k\}$ contains all of the information in $\{y_k\}$ because we can always recover $\{y_k\}$ by applying a linear filter to $\{v_k\}$. Suppose now

that instead of building a Wiener filter to estimate $\{x_k\}$ given the process $\{y_k\}$, we build a filter to estimate $\{x_k\}$ using $\{v_k\}$. To do this, we need to solve (3.8) but replacing y with v. Since $\{v_k\}$ is white it has the autocorrelation function, $\sigma^2_v \delta[k]$, which makes finding the solution to (3.8) very easy.

$$h(n) = \frac{r_{xv}(n)}{\sigma_v^2} u(n) \tag{3.10}$$

where u(n) is the unit step function. Taking the z transform of (3.10) cannot, in general, be done in closed form, so we define the notation $[R_{xv}(z)]^+$ to denote the z transform of $r_{xv}(n)u(n)$. Thus, we obtain

$$H(z) = \frac{[R_{xv}(z)]^+}{\sigma_v^2} \tag{3.11}$$

With a few algebraic manipulations, we obtain the causal Wiener Filter in terms of the statistical properties of y_n.

$$H(z) = \frac{1}{\sigma_v^2} Q(z) \left[Q^*\left(\frac{1}{z^*}\right) R_{xy}(z) \right]^+ \tag{3.12}$$

3.2.2. FIR Wiener Filtering

Finally, suppose we wish to restrict h_n to be a a M tap FIR filter. Then the Wiener-Hopf equations can be written in terms of the correlation matrix and crosscorrelation matrix of $\{x_k\}$ with $\{y_k\}$ [13]:

$$R_y = \begin{bmatrix} r_y(0) & r_y(1) & \dots & r_y(M-1) \\ r_y(-1) & r_y(0) & \dots & r_y(M-2) \\ \vdots & \vdots & & \vdots \\ r_y(1-M) & r_y(2-M) & \dots & r_y(0) \end{bmatrix} \quad R_{xy} = \begin{bmatrix} r_{xy}(0) \\ r_{xy}(1) \\ \vdots \\ r_{xy}(M-1) \end{bmatrix} \tag{3.13}$$

$$R_y^T \begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(M-1) \end{bmatrix} = R_{xy} \tag{3.14}$$

18

One case where we might want to use a FIR wiener filter is in one step ahead prediction of the random process $\{y_k\}$. In this case $\{x_k\}$, the process we are trying to match, is just $\{y_k\}$ advanced by one time unit. The Wiener filter reduces to:

$$\hat{y}_{k+1} = \begin{bmatrix} r_y(1) & \dots & r_y(M) \end{bmatrix} R_y^{-1} \begin{bmatrix} y_k \\ \vdots \\ y_{k-M+1} \end{bmatrix} \tag{3.15}$$

Note how similar (3.15) is to the basic formula for linear least squares estimation (3.3). In fact, (3.15) can also be derived by applying (3.3).

3.3. Kalman Filtering

Like Wiener Filtering, Kalman Filtering is an extension of Linear Least Squares estimation applied to stochastic processes. In the development of Wiener Filtering, we assumed the desired signal is wide sense stationary, and used spectral functions such as the power spectral density to derive the optimal filter. In contrast, the Kalman Filter development does not assume stationarity of the desired signal, but is developed from a state space description of the desired and measured signals. A thorough treatment of the Kalman Filter can be found in [10], which much of our development parallels.

The Kalman Filtering problem begins with the following signal model:

$$x_{k+1} = F_k x_k + G_k w_k \tag{3.16}$$

$$z_k = H_k^* x_k + v_k \tag{3.17}$$

We assume v_k, w_k are both zero mean, independent of each other, and have covariances $E[v_k v_l^*] = R_k \delta[k-l]$, $E[w_k w_l^*] = Q_k \delta[k-l]$. The initial state x_0 is random with mean $\bar{x}_0$, covariance P_0, and is independent of v_k, w_k. Let Z_k denote the set of observations $\{z_0, z_1, \dots, z_k\}$, and define the estimator error covariance matrix as

$$\Sigma_{k+1|k} = E[(x_{k+1} - \hat{x}_{k+1}(Z_k))(x_{k+1} - \hat{x}_{k+1}(Z_k))^*] \tag{3.18}$$

where $x_{k+1|k} = \hat{x}_{k+1}(Z_k)$ is our estimate of x_{k+1} based on the data Z_k. The problem is to find the function $\hat{x}_{k+1}(Z_k)$ that minimizes $\Sigma_{k+1|k}$. It is shown in [10] that the Kalman Filter, which is affine function of Z_k, achieves a estimator error covariance, $\Sigma_{k+1|k}$ which

is less than or equal to the estimator error covariance of any other affine estimator. Because $\Sigma_{k+1|k}$ is a matrix, "less than" is meant in the standard linear algebra sense. The Kalman Filter is defined by the recursion relations:

$$\hat{x}_{k+1|k} = (F_k - K_k H_k^*)\hat{x}_{k|k-1} + K_k z_k, \quad \hat{x}_{0|-1} = \bar{x}_0 \tag{3.19}$$

$$K_k = F_k \Sigma_{k|k-1} H_k [H_k^* \Sigma_{k|k-1} H_k + R_k]^{-1} \tag{3.20}$$

The estimation error covariance used in (3.20) to compute the Kalman gain, K_k, is given by a discrete time Riccati equation:

$$\Sigma_{k+1|k} = F_k [\Sigma_{k|k-1} - \Sigma_{k|k-1} H_k (H_k^* \Sigma_{k|k-1} H_k + R_k)^{-1} H_k^* \Sigma_{k|k-1}] F_k^* + G_k Q_k G_k^* \tag{3.21}$$

where $\Sigma_{0|-1} = P_0$.

When v_k, w_k, and x_0 are Gaussian, much more can be said about the Kalman Filter. $\hat{x}_{k+1}(Z_k)$ is actually the conditional mean $E[x_{k+1}|Z_k]$, and $\Sigma_{k+1|k}$ is the conditional covariance $E[(x_{k+1} - \bar{x}_{k+1})(x_{k+1} - \bar{x}_{k+1})^*|Z_k]$. Thus the Kalman Filter equations become a mechanism for updating the entire conditional probability density of x_k [10].

3.3.1. Asymptotic Behavior of the Kalman Filter

Often the signal model that best represents our desired signal is time invariant, and for computational reasons we would like to estimate the signal with a time invariant filter. As we have seen, the Kalman filter is in general a time varying filter, but there are conditions under which it asymptotically becomes time invariant, and is truly time invariant with a proper choice of P_0. Our signal model is now:

$$x_{k+1} = Fx_k + Gw_k \tag{3.22}$$

$$z_k = H^* x_k + v_k \tag{3.23}$$

Again, we assume v_k, w_k are both zero mean, independent of each other, and have covariances $E[v_k v_l^*] = R_k \delta[k-l]$, $E[w_k w_l^*] = Q_k \delta[k-l]$. Then if the spectrum of F is contained in the unit circle ($|\lambda_i(F)| < 1 \,\forall i$), then for any nonnegative symmetric initial matrix P_0, we have [10]:

$$\lim_{k \to \infty} \Sigma_{k+1|k} = \bar{\Sigma} \qquad (3.24)$$

Where $\bar{\Sigma}$ satisfies the Ricatti equation:

$$\bar{\Sigma} = F[\bar{\Sigma} - \bar{\Sigma} H (H^* \bar{\Sigma} H + R)^{-1} H^* \bar{\Sigma}] F^* + GQG^* \qquad (3.25)$$

As a result, the Kalman gain approaches a limiting value of

$$\bar{K} = F\bar{\Sigma} H [H^* \bar{\Sigma} H + R]^{-1} \qquad (3.26)$$

It should also be clear that if P_0 is selected to be $\bar{\Sigma}$, the Kalman Filter will be truly time invariant, not just asymptotically time invariant.

These results still hold when the spectrum of F is not contained in the unit circle, but instead the pair $[H, F]$ is detectable and $[F, G\sqrt{Q}]$ is stabilizable. For a review of the concepts of stabilizability and detectability refer to [15].

Finally, when the spectrum of F is contained in the unit circle, the signal model (3.22) leads to a $\{x_k\}$ that is asymptotically wide sense stationary. Because of this one can also build a Wiener filter to estimate x_k from Z_k. One might expect that in this case, the Kalman filter asymptotically approaches the Wiener filter estimator. This is indeed the case, and is proved in detail in [10].

Chapter 4 RtR Control Methods

In Chapter 2 we defined two classes of process drift, offset drift and parameter drift. These two ways of modeling process drift lead to two distinct classes of RtR control: *offset drift cancellation* and *parameter adaptive* approaches.

4.1. Offset Drift Cancellation Approaches

Here process variation is assumed to be entirely in the offset term (the parameter drift is absent). Consequently, the input sensitivies are assumed to be fixed and known. The idea is simply to estimate the current offset term b_k and select an input setting to compensate for the offset.

4.1.1. Exponentially-Weighted Moving Average

This is one of the most straightforward methods, and is studied extensively in [14]. Gradual mode EWMA, as it is termed in [14], assumes a process model of form

$$y_k^o = Au_k + b_k \tag{4.1}$$

$$y_k = y_k^o + e_k \tag{4.2}$$

where y_k^o is the true process output (assumed to be a scalar), e_k is the measurement noise, and y_k is the measured process output. An estimate for the term b_k is computed recursively as

$$\hat{b}_{k+1} = (1 - \omega)\hat{b}_k + \omega(y_k - Au_k) \tag{4.3}$$

The estimator (4.3) is intuitively appealing. It says that our estimate of b_k is a weighted average of the current residual and the value of our previous estimate. By carying out the

recursion in (4.3) one sees that $\hat{b}_k$ is a weighted average of all of the past residuals with form

$$\hat{b}_{k+1} = \sum_{i=-\infty}^{k} \omega(1-\omega)^{k-i}(y_i - Au_i)$$

(4.4)

This relationship is precisely why this type of filtering is called Exponentially Weighted Moving Average.

Having obtained an estimate of the drift, the input setting is selected by finding a u_k that will meet the target output T by cancelling the estimated drift:

$$T = Au_{k+1} + \hat{b}_{k+1}$$

(4.5)

When there are multiple inputs, the choice of u_{k+1} that satisfy (4.3) is not unique. Considerations such as which input requires the least effort to change, or which input has a corresponding sensitivity in the process model with the least modeling error can be used in determining which component of u_{k+1} to actuate the most. One method suggested in [14] is to find the value of u_{k+1} that satisfies (4.5) and has the least euclidean distance from the previous input u_k. This can be calculated easily as

$$u_{k+1} = u_k + A^*(AA^*)^{-1}(T - Au_{k-1} - \hat{b}_k)$$

(4.6)

Equation (4.6) is not independent of the units used to measure the inputs. Because of this, one may wish to use a different cost function in the minimization besides euclidean distance, such as $(u_{k+1} - u_k)^* W (u_{k+1} - u_k)$ where W is positive definite and hermetian. Then u_k simply becomes

$$u_{k+1} = u_k + W^{-1}A^*(AW^{-1}A^*)^{-1}(T - Au_{k-1} - \hat{b}_k)$$

(4.7)

Of course, when some components of u_k are feed-forward, (4.5) must be achieved by actuating the control inputs.

4.1.1.1. Conditions for Optimality of EWMA Filtering.

Using the estimation theory reviewed in chapter 2, we can show that the EWMA filter used in obtaining the estimate $\hat{b}_k$ is optimal for a class of drift models, provided that ω is chosen properly, though in most practical situations the choice of ω is usually ad-hoc.

Suppose that the process b_k is governed by the following simple Autoregressive (AR) model:

$$b_{k+1} = b_k + w_k \tag{4.8}$$

where w_k is a white random process (meaning it has zero mean and autocorrelation of $E[w_k w_l^*] = \sigma_w^2 \delta[k-l]$. Similarly we assume the measurement noise is white, with variance σ_e^2. Then the Kalman filter recursions for estimating b_k can be found by substitution of (4.8), (4.1), and (4.2) into (3.19), (3.20), and (3.21):

$$\hat{b}_k = (1 - K_k)\hat{b}_k + K_k(y_k - A_k u_k) \tag{4.9}$$

$$K_k = \sigma_{k|k-1}^2 [\sigma_{k|k-1}^2 + \sigma_e^2]^{-1} \tag{4.10}$$

$$\sigma_{k+1|k}^2 = \sigma_{k|k-1}^2 - \sigma_{k|k-1}^2 (\sigma_{k|k-1}^2 + \sigma_e^2)^{-1} \sigma_{k|k-1}^2 + \sigma_w^2 \tag{4.11}$$

Where $\sigma_{k|k-1}^2$ is the estimator error variance (taking the place of $\Sigma_{k|k-1}$) and $(y_k - A_k u_k)$ is the measured variable (taking the place of z_k).

In the signal model (4.8), the F matrix is simply 1. Because of this, b_k is not asymptotically wide sense stationary. However, we learned in the previous chapter that when the pair $[H, F]$ is detectable and $[F, G\sqrt{Q}]$ is stabilizable, the Kalman filter asymptotically approaches a time invariant filter, even when F is unstable. This is indeed the case in this example because $H = F = G\sqrt{Q} = 1$. Using (3.24), (3.25), and (3.26) we find that the Kalman filter gain converges to:

$$\bar{K} = \bar{\sigma}^2 [\bar{\sigma}^2 + \sigma_e^2]^{-1} \tag{4.12}$$

Where $\bar{\sigma}$ satisfies:

$$\frac{1}{\bar{\sigma}^2 - \sigma_w^2} = \frac{1}{\bar{\sigma}^2} + \frac{1}{\sigma_e^2} \tag{4.13}$$

Thus as $k \to \infty$, the Kalman Filter, which is the optimal linear time varying predictor of b_{k+1} from $\{(y_k - Au_k) \dots (y_{k0} - Au_{k0})\}$, approaches a linear time invariant filter. Equivalently, if we consider the case where the initial time, k_0, approaches $-\infty$, then the optimal predictor at any finite time k, is the linear time invariant filter characterized by (4.12) and (4.13).

When the Kalman gain K_k is fixed to be $\bar{K}$, (4.9) becomes the EWMA filtering relationship (4.3) whith the weight ω equal to $\bar{K}$. Thus, EWMA is optimal when we have exclusively offset drift, described by a simple random walk model of form (4.8), and the weight ω is selected to be the $\bar{K}$. Finally, we must assume that the drifting and filtering processes began in the infinite past.

This final point may seem a bit technical, but it can be justified quite intuitively by a simple example. Suppose the offset drift begins at time $k = 0$, but at that time we know the value of b_0 with certainty. Then at time $k = 1$, we should expect that the optimal linear estimate would rely greatly on the initial estimate b_0, and very little on the measurement $(y_1 - Au_1)$. Gradually as the process drifts and the uncertainty in b_k grows, the optimal estimator will depend more an more on the most recent measurement. Only after the process has drifted for a fairly long time, will the optimal filter converge to fixed weightings on the most recent measurement and the previous estimate. To demonstrate this, we can compute the optimal estimator of b_1 using the basic linear least squares estimator result (3.3), we find:

$$\hat{b}_{1opt} = \left(1 - \frac{\sigma_w^2}{\sigma_w^2 + \sigma_e^2}\right)\hat{b}_0 + \frac{\sigma_w^2}{\sigma_w^2 + \sigma_e^2}(y_1 - Au_1) \tag{4.14}$$

And indeed, one can show that $\sigma_w^2/(\sigma_w^2 + \sigma_e^2)$ is less than the value K of (4.12).

4.1.1.2. The Certainty Equivalence Assumption in EWMA Control.

In the EWMA scheme, we recursively estimate the parameter b_k and select an input as if our estimate were certain. However, one cannot assume that in general the optimal adaptive controller consists of two decoupled blocks, one that estimates the parameters and the other that selects inputs as if the estimated parameters were known with certainty. The optimal controller might be one that introduces noise into the input in order to achieve

better parameter estimates, or perhaps scales down the input that a certainty equivalence law would use because of low confidence in the estimated parameters. Fortunately, under the assumption of having exclusively offset drift, the certainty equivalence rule is optimal for EWMA control.

Result: For a process undergoing offset drift alone, a controller consisting of:

- A linear least squares estimate of b_{k+1} from data y_k thru y_{k_0}, denoted $\hat{b}_{k+1}(y_k, ..., y_{k_0})$

- A certainty equivalence input selector that chooses u_{k+1} to satisfy the equation:

$$T = Au_{k+1} + \hat{b}_{k+1} \tag{4.15}$$

results in an output mean squared error that is lower or as low as the mean squared error achieved by any other linear controller. By "linear controller," we mean any control that computes the input u_{k+1} as an affine function of data: y_k thru y_{k_0}

Proof: Define $\hat{b}_{k+1}(y_k, ..., y_{k_0})$ to be the linear least squares estimate of b_{k+1} using the measured data. And suppose it achieves an estimation mean squared error of:

$$\hat{\sigma}^2 = E[|b_{k+1} - \hat{b}_{k+1}|^2]$$

Then if we use the certainty equivalence control:

$$u_{k+1} = -A^*(AA^*)^{-1}(\hat{b}_{k+1}(y_k, ..., y_{k_0}))$$

The output variance, $E[|y_{k+1}|^2]$, will be equal to the estimation error $\hat{\sigma}^2$.

Now suppose there exists a different linear control $\tilde{u}_{k+1}(y_k, ..., y_{k_0})$ which results in an output variance, $E[|y_{k+1}|^2]$ equal to $\tilde{\sigma}^2$, where $\tilde{\sigma}^2$ is less than the variance achieved by the certainty equivalence control, $\hat{\sigma}^2$.

Then we could build a new linear estimator of b_{k+1}, $\tilde{b}_{k+1}$ defined as:

$$\tilde{b}_{k+1} = T - A\tilde{u}_{k+1}(y_k, ..., y_{k_0})$$

Where the estimation mean squared error achieved by this estimator is:

$$\tilde{\sigma}^2 = E[|b_{k+1} - \tilde{b}_{k+1}|^2]$$

Because $\tilde{\sigma}^2 < \hat{\sigma}^2$, this estimator achieves a smaller error variance than the estimator $\hat{b}_{k+1}(y_k, ..., y_{k_0})$, which was defined to be the linear least squares estimator. This is a contradiction. Thus, no other linear control achieves a better mean squared ouput error than the certainty equivalence control.

4.1.1.3. Practical issues in EWMA control

The attractiveness of the EWMA scheme lies in its simplicity, and intuitive form. However, even though we have showed conditions under which EWMA is optimal, these conditions are almost never true. If the offset drift of a process did obey (4.8), its variance would grow unbounded as time elapsed, which is unrealistic. In addition, choosing the weight ω can be difficult in most situations, and most often is chosen ad-hoc. Though we computed the optimal weight ω in the preceding section, the result is of limited utility because it only holds when the offset drift behaves according to (4.8).

EWMA control schemes have been successfully deployed in processes such as Chemical Mechanical Polishing [5], and I line lithography [4] (on a lot to lot basis rather than wafer to wafer).

4.1.2. Robust Drift Cancellation

We have seen that EWMA control is a simple, intuitive strategy but is optimal under a very limited class of drift models. We present here a variation of FIR Wiener filtering which we term Robust Drift Cancellation (RDC) that achieves optimality for a larger range of problems because it is designed directly from historical data.

As in EWMA, we assume a process model of the form:

$$y_k^o = Au_k + b_k \tag{4.16}$$

$$y_k = y_k^o + e_k \tag{4.17}$$

where y_k^o is the true process output (assumed to be a scalar), e_k is the measurement noise, and y_k is the measured process output. We assume b_k drifts according to some unknown random process, but with a known mean $\bar{b}$. Suppose we have m data records, and within each data record are process recipes and outputs for L wafers. Using residuals from the mth data record, we can compute an estimate of the (n x n) correlation matrix of $(b_k - \bar{b} + e_k)$, which we denote z_{mk}, where the m in the subscript emphasizes that the value is from the mth data record.

$$R_Z^m = \frac{1}{L-n+1} \sum_{k=n}^{L} \mathbf{z}_{mk} \mathbf{z}^*_{mk} \tag{4.18}$$

Where,

$$\mathbf{z}_{mk} = [z_{mk}, z_{m(k-1)}, \ldots, z_{m(k-n+1)}]^T \tag{4.19}$$

Similarly, the time average autocorrelation can be computed as:

$$C_Z^m = \frac{1}{L-n} \sum_{k=n}^{L-1} z_{m(k+1)} \mathbf{z}_{mk}^* \tag{4.20}$$

Using these matrices, we may compute an n-tap FIR Wiener filter for estimating z_k using (3.15):

$$\hat{z}_{(k+1)} = (C_Z^m)^T (R_Z^m)^{-1} \begin{bmatrix} (y_k - Au_k - \bar{b}) \\ \vdots \\ (y_{k-n+1} - Au_{k-n+1} - \bar{b}) \end{bmatrix} \tag{4.21}$$

Finally if we assume that the process $\{e_k\}$ is white, then $\hat{z}_{m(k+1)} = \hat{b}_{m(k+1)}$. Thus, a FIR Wiener filter based control law uses any u_k satisfying:

$$Au_{k+1} = T - \bar{b} - (C_Z^m)^T (R_Z^m)^{-1} \begin{bmatrix} (y_k - Au_k - \bar{b}) \\ \vdots \\ (y_{k-n+1} - Au_{k-n+1} - \bar{b}) \end{bmatrix} \tag{4.22}$$

The above control law will reduce the variance of the output from σ_y^2 to $\sigma_y^2 - (C_Z^m)^T (R_Z^m)^{-1} (C_Z^m)^T$, assuming that the when this controller is in operation, the statis-

tics of b_k and e_k are identical to how they were when the mth data record was recorded. However, in any realistic process the statistics of b_k and e_k are nonstationary. If the statistics worsen too much, a controller designed using a single past data record maybe too aggressive and have poor performance. This is the motivation for developing a design methodology that uses "worst case" bounds on drift statistics obtained from analyzing all m data records [17].

We begin by finding a positive definite, hermetian, matrix upper bound, $\bar{R}_Z$ that satisfies:

$$\bar{R}_Z > R_Z^i \quad \forall i \in [1,m] \tag{4.23}$$

Next, find a vector, $\underline{C}_z$ such that

$$\underline{C}_z^T \bar{R}_Z^{-1} \underline{C}_z < (C_z^m)^T \bar{R}_Z^{-1} C_z^m \quad \forall i \in [1,m] \tag{4.24}$$

Then, the robust drift cancellation law is to use the control u_k satisfying:

$$Au_k = T - \bar{b} - \underline{C}_Z^T \bar{R}_Z^{-1} \begin{bmatrix} (y_k - Au_k - \bar{b}) \\ \vdots \\ (y_{k-n+1} - Au_{k-n+1} - \bar{b}) \end{bmatrix} \tag{4.25}$$

To study the performance of the above control law, we make two modest assumptions. One is that the sample correlation matrix of z_k for any set of L wafers in the future satisfies:

$$\frac{1}{L-n+1} \sum_{k=n}^{L} \hat{z}_k \hat{z}^*_k < \bar{R}_Z \tag{4.26}$$

Similarly, the sample autocorrelation vector for any set of L wafers in the future satisfies:

$$\left(\frac{1}{L-n} \sum_{k=n}^{L-1} z_{(k+1)} \hat{z}_k^* \right)^T \bar{R}_Z^{-1} \left(\frac{1}{L-n} \sum_{k=n}^{L-1} z_{(k+1)} \hat{z}_k^* \right) < (C_z^m)^T \bar{R}_Z^{-1} C_z^m \tag{4.27}$$

In other words, we assume the statistical bounds we computed on the m data records are still valid in the future, but we do not assume any sort of statistical model for the drift, nor

do we assume the processes is wide sense stationary. Then the sample mean squared error of the measured output using the RDC law can be computed as follows:

$$\frac{1}{L-n}\sum_{k=n}^{L-1}(y_{k+1}-T)^2 = \frac{1}{L-n}\sum_{k=n}^{L-1}\left(z_{k+1}-\underline{C}_Z^T\bar{R}_Z^{-1}\begin{bmatrix}(y_k-Au_k-\bar{b})\\\vdots\\(y_{k-n+1}-Au_{k-n+1}-\bar{b})\end{bmatrix}\right)^2$$

$$= \frac{1}{L-n}\sum_{k=n}^{L-1}(z_k^2+\underline{C}_Z^T\bar{R}_Z^{-1}\underline{z}_k\underline{z}_k^T\bar{R}_Z^{-1}\underline{C}_Z-2\underline{C}_Z^T\bar{R}_Z^{-1}\underline{z}_k z_{k+1})$$

$$< \frac{1}{L-n}\sum_{k=n}^{L-1}z_k^2-\underline{C}_Z^T\bar{R}_Z^{-1}\underline{C}_Z \tag{4.28}$$

Observe that the quantity:

$$\frac{1}{L-n}\sum_{k=n}^{L-1}z_k^2 \tag{4.29}$$

is what the sample output mean squared error would have been had control not been used. Therefore, the sample mean squared output error is guaranteed to be reduced by $\underline{C}_Z^T\bar{R}_Z^{-1}\underline{C}_Z$ across a set of L wafers. The RDC design procedure is summarized in Figure 4.1.

The advantages of the RDC approach are clear. It can be easily designed from historical data. It is robust to non-stationary in the drift statistics. And finally, one can easily compute a conservative, a priori estimate of how much RDC will reduce process variance.

Select number of taps: n

From m historical data records, compute sample correlation matrix and autocoreelation vectors:

$$R_Z^m, \quad C_Z^m$$

Compute Matrix upper bound using numerical optimization

$$\bar{R}_Z > R_Z^i \quad \forall i$$

Compute lower bound vector satisfying:

$$\underline{C}_z^T \bar{R}_Z^{-1} \underline{C}_z < (C_z^m)^T \bar{R}_Z^{-1} C_z^m \quad \forall i$$

Build the Control Law:

$$A u_{k+1} = T - \bar{b} - \underline{C}_Z^T \bar{R}_Z^{-1} \begin{bmatrix} (y_k - A u_k - \bar{b}) \\ \vdots \\ (y_{k-n+1} - A u_{k-n+1} - \bar{b}) \end{bmatrix}$$

Using the Control Law, output mean squared error reduced by:

$$\underline{C}_Z^T \bar{R}_Z^{-1} \underline{C}_Z$$

Figure 4.1. Robust Drift Cancellation (RDC) Design Procedure

4.1.2.1. Robust Drift Cancellation Design Issues

In the preceding discussion, it was not mentioned how to choose the number of taps n. Computational limitations may dictate a maximum value for n. The control designer may wish to try many different values of n and find the n which is the best trade-off between computational complexity and potential variance reduction as computed by $\underline{C}_Z^T \bar{R}_Z^{-1} \underline{C}_Z$.

The design of the RDC law required the computation of a matrix upper bound in equation (4.23). Computation of this upper bound is nontrivial, because one would like to find a least upper bound; an upper bound that is too large would lead to an overly conservative controller with a low value for $\underline{C}_Z^T \bar{R}_Z^{-1} \underline{C}_Z$. Fortunately, the set of matrices that satisfy (4.23) form a convex set, making this problem amenable to numerical optimization. For a discussion of appropriate optimization techniques refer to [16].

4.2. Parameter Adaptive Approaches

Here we assume that the observed process drift is due to both parameter and offset drift. The general strategy is to update the process parameters A_k and b_k using the available data. The recipe u_k is generated using the latest estimates of A_k and b_k and the target output T.

4.2.1. Kalman Filter Methods

Kalman filtering RtR control is studied in [18] for the application of photoresist spin coating. In this approach we assume a process model of form:

$$y_k^o = \begin{bmatrix} u_k & 1 \end{bmatrix} \begin{bmatrix} A_k^T \\ b_k \end{bmatrix} \tag{4.30}$$

$$y_k = y_k^o + e_k \tag{4.31}$$

where y_k^o is the true process output (assumed to be a scalar), e_k is the measurement noise, and y_k is the measured process output. We assume e_k to be a white random process with variance σ_e^2. Furthermore, we assume that the deviation vector of A_k and b_k, defined as

$$x_k = \begin{bmatrix} A_k & b_k \end{bmatrix}^T - \begin{bmatrix} \bar{A} & \bar{b} \end{bmatrix}^T \tag{4.32}$$

is a Gauss Markov process, obeying the stochastic difference equation:

$$x_{k+1} = Fx_k + Gw_k \tag{4.33}$$

where w_k is a white random process with covariance matrix I. Also, we assume x_0 is $\begin{bmatrix} \bar{A} & \bar{b} \end{bmatrix}^T$.

The Kalman filter control approach is to use Kalman filtering equations outlined in Chapter 3 to estimate A_k and b_k. Referring to (3.19) - (3.21), substituting $\begin{bmatrix} u_k & 1 \end{bmatrix}$ for H^*_k and e_k for v_k we have:

$$\hat{x}_{k+1|k} = \left(F_k - K_k \begin{bmatrix} u_k & 1 \end{bmatrix} \right) \hat{x}_{k|k-1} + K_k \left(y_k - \begin{bmatrix} u_k & 1 \end{bmatrix} \begin{bmatrix} \bar{A} \\ \bar{b} \end{bmatrix} \right) \tag{4.34}$$

$$K_k = F_k \Sigma_{k|k-1} \begin{bmatrix} u_k \\ 1 \end{bmatrix} \left[\begin{bmatrix} u_k & 1 \end{bmatrix} \Sigma_{k|k-1} \begin{bmatrix} u_k \\ 1 \end{bmatrix} + \sigma_e^2 \right]^{-1} \tag{4.35}$$

$$\Sigma_{k+1|k} = F_k \left[\Sigma_{k|k-1} - \Sigma_{k|k-1} \begin{bmatrix} u_k \\ 1 \end{bmatrix} \left(\begin{bmatrix} u_k & 1 \end{bmatrix} \Sigma_{k|k-1} \begin{bmatrix} u_k \\ 1 \end{bmatrix} + \sigma_e^2 \right)^{-1} \begin{bmatrix} u_k & 1 \end{bmatrix} \Sigma_{k|k-1} \right] F^*_k$$

$$+ G_k G^*_k \tag{4.36}$$

The input used is any u_{k+1} satisfying:

$$T = \begin{bmatrix} u_{k+1} & 1 \end{bmatrix} \left(\begin{bmatrix} \bar{A}^T \\ \bar{b} \end{bmatrix} + \hat{x}_{k+1|k} \right) \tag{4.37}$$

where T is the target output.

4.2.1.1. Certainty Equivalence in Kalman Filtering Control

In the Kalman filtering control scheme, we recursively estimate the parameter x_k and subsequently select an input as if our estimate were certain. Though we showed this approach is optimal in offset drift situations, it is not optimal for the parameter drift case. Depending on the particular problem, the optimal control (in terms of mean squared output error) might introduce extra excitation into the input at times to achieve better parameter estimates -- improving control performance on future wafers at the cost of adding variance to the current wafer. Finding a closed form solution for the optimal control law, when the optimal control is so elaborate can be extremely difficult or impossible in many problems. From a practical perspective, we may not want to use a controller that deliberately adds unnecessary excitation to the inputs, possibly ruining a wafer to achieve slightly better per-

formance on future wafers. Therefore, in Kalman filter RtR control, we do not bother to find the true optimal controller, and instead use the certainty equivalence approach.

4.2.2. Practical Concerns in Parameter Adaptive Methods

The shortcoming of Kalman filter methods for RTR control in particular, and parameter adaptive control methods in general, is as follows. If there are too many process parameters x_k, estimating them requires a lot of data. By the time we have enough data to estimate the process parameters, they may have drifted considerably. As a result, the estimated process model is poor and RtR control based on this model can in some situations increase process variance. These problems are illustrated in [18].

4.3. Offset Drift Cancellation Applied to Processes with Parameter Drift

The purpose of this section is to demonstrate that under certain conditions, an offset drift cancellation approach has the potential to perform as well as a parameter adaptive approach, even when the process is undergoing parameter drift.

Suppose we have a single input singe output (SISO) process that when run with a fixed input $\bar{u}$ has a measured output y_k that is of form $\delta_k + \bar{y}$ where δ_k is some random deviation from the mean output. One way to model this behavior for control design is to assume that the output deviation δ_k is being caused by offset drift:

$$y_k = Au_k + (\bar{b} + \delta_k) + e_k \tag{4.38}$$

Another model would be to assume that the output deviation is caused by drift in A:

$$y_k = \left(A + \frac{\delta_k}{\bar{u}}\right)u_k + \bar{b} + e_k \tag{4.39}$$

Now, suppose that we build an estimator of the drift process δ_k, and call the estimate, $\hat{\delta}_k$. Then the control law corresponding to the offset drift model (4.38) is to choose u_k satisfying

$$u_k = \bar{u} - \frac{\hat{\delta}_k}{A} \tag{4.40}$$

The control law corresponding to parameter drift model (4.39) is to choose u_k satisfying:

34

$$u_k = \frac{\bar{y}-\bar{b}}{A+\frac{\hat{\delta}_k}{\bar{u}}} = \frac{\bar{y}-\bar{b}}{A\left(1+\frac{\hat{\delta}_k}{A\bar{u}}\right)} = \bar{u}-\frac{\hat{\delta}_k}{A}+\bar{u}o\left(\left|\frac{\hat{\delta}_k}{A\bar{u}}\right|^2\right) \qquad (4.41)$$

We see that the two control laws are nearly identical when the argument of the $o(\ )$, $\left|\hat{\delta}_k/(A\bar{u})\right|^2$ becomes small. Some intuition on this result can be gained by observing Figure 4.2. The figure shows how an offset adaptation and parameter adaptation both account for the same small change in observed output, assuming that $A\bar{u}$ is large. Note that in the operating region, the two models look almost identical. Consequently, the use of either model would lead to almost identical control input selections.

Figure 4.3 shows the case when $A\bar{u}$ is small. The geometry of the problem when $A\bar{u}$ is small causes the Offset adaptation model to have a significantly different slope than the parameter adaptation model. Which model is used will have a drastic effect on the control decision.
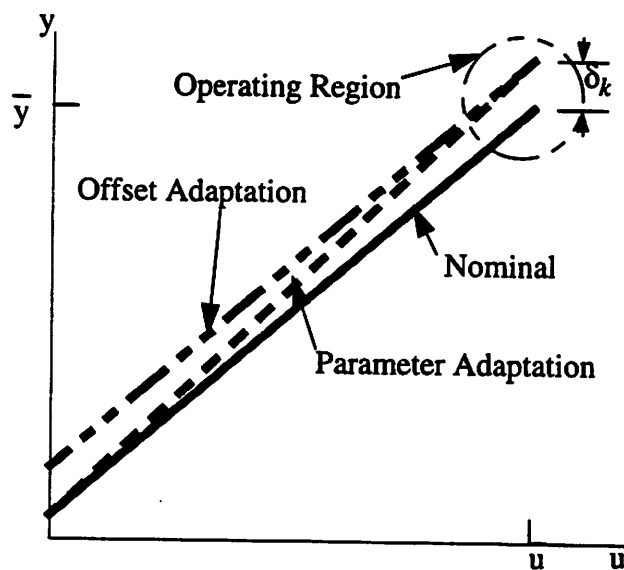

Figure 4.2. Offset vs. Parameter Adaptation: Large $A\bar{u}$ Product

In conclusion, when the $A\bar{u}$ product is small compared to the size of the process output deviations, an offset drift cancellation strategy may work as well as a parameter

35

adaptive approach. This is an extremely useful result because it is often the case that $A\bar{u}$ is large, and it is much easier to build an offset drift model than a parameter drift model.
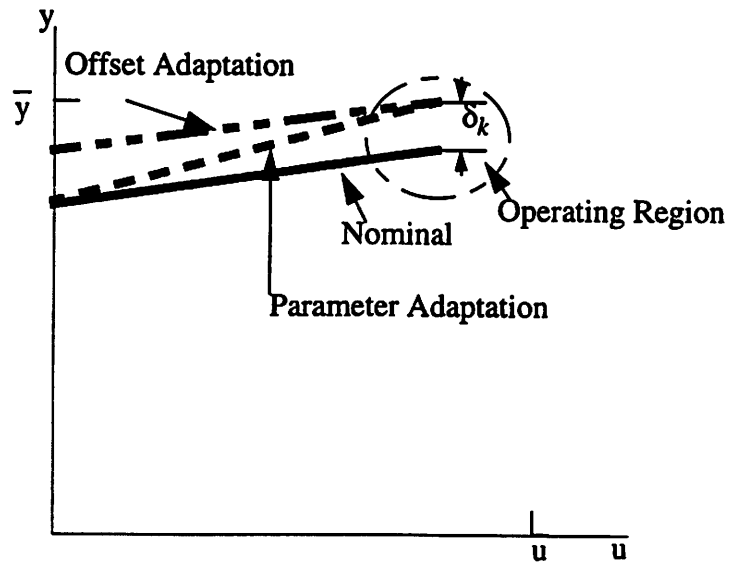

Figure 4.3. Offset vs. Parameter Adaptation. Small $A\bar{u}$ Product.

Chapter 5 RtR Control for DUV Lithography

In previous chapters, we have established a theoretical framework and created tools for RtR control design. In this chapter, we apply these tools to the problem of controlling Deep Ultra-Violet (DUV) lithography. Taking into account our knowledge of the process and available in-line measurements, we develop a control architecture, and simulate the controller's operation in the presence of process drift.

5.1. Motivation for DUV Lithography Control

Deep Ultra Violet (DUV) lithography is anticipated to become the main lithography technology for the next generation of ICs with sub-quarter-micron linewidths. DUV lithography can achieve smaller linewidths than conventional I-line lithography for two important reasons. One reason is that DUV lithography uses shorter wavelength illumination which reduces the diffraction effects for a given grating size. The other reason for the greater capability of DUV technology is the use of extremely sensitive Chemically Amplified (CA) photoresists, which are capable of generating sharp resist profiles, even when the aerial image is blurred by diffraction effects.

Although lower wavelength light and CA resists have increased lithography capability, they have also increased the susceptibility of the lithography process to disturbances. Lower wavelength light reduces the depth of focus, making critical dimensions (CD) more sensitive to wafer topography. CA resist increases the sensitivity to exposure variations from the stepper, and as we shall see later in the chapter, introduce another source of variance during the Post Exposure Bake (PEB) process.

Companies developing DUV lithography processes have indeed observed unacceptable levels of CD variability. For example, some manufacturers have observed a 3σ CD variation of 24 nm within each wafer, between different wafers, and between lots.

To improve the efficiency of DUV lithography, IC producers need to find ways to reduce CD variability to ensure sufficient yields and consistent circuit performance. Having already spent millions of dollars on steppers with ultra-high precision lenses and mechanical components, robotic cluster tools for applying photoresist, and environmental isolation, IC manufacturers trying to reduce CD variability are left with very few alternatives for additional improvements.

RtR control is one of the most promising alternatives for CD variance reduction at the wafer to wafer and lot to lot level. In this chapter, we develop a control architecture for this extremely important problem using the framework developed in previous chapters.

5.2. The DUV Lithography Process

To develop a RtR controller for DUV lithography, it is necessary to understand how DUV lithography works and what are the most likely sources of process drift.

5.2.1. Basic Process Flow

See Figure 5.1 for a flowchart of the DUV lithography sequence. Before the DUV lithography sequence begins, the film to be patterned is deposited onto the wafer. Next the wafer is coated with an antireflective coating, which helps to reduce standing waves of light in the resist during the exposure step. The antireflective coating is baked dry, and then the wafer is coated with a DUV chemically amplified photoresist. This too is baked dry, and then the wafer is brought to the stepper.

The stepper exposes the wafer (die by die) to DUV light, passed through a patterned reticle. Wherever the resist is exposed to light, photo-acid is formed. The two most important process parameters that can be adjusted at the stepper are the exposure dose per unit area, and the focus level.

Following exposure, the wafer is given a Post Exposure Bake (PEB). During the PEB, three simultaneous processes occur [19]. In a chemical reaction called deprotection,

39

photo-acid catalyzes a weakening of organic chemicals in the resist, making it soluble wherever there was acid present from the exposure. At the same time, the photo acid is gradually "quenched" or consumed by a second reaction. Finally, the acid diffuses through the resist somewhat before it is "quenched," which results in a widening of areas where the resist is deprotected. Detailed studies of the reaction kinetics and mechanisms are presented in [19] and [20]. It has been shown that all three effects, deprotection, quenching, and diffusion are temperature dependent.

After PEB, the wafer is immersed in developer, causing the photoresist to dissolve wherever it had been deprotected. The developer is extremely selective between resist that has been deprotected and resist that is not deprotected [21]. Therefore, varying the develop time usually does not effect CD greatly. A study of the relationship between deprotection and develop dissolution rate can be found in [21].

FILM DEPOSITION

SPIN COAT&
SOFT BAKE

EXPOSE

PEB

DEVELOP

Figure 5.1. DUV Lithography Process Flow
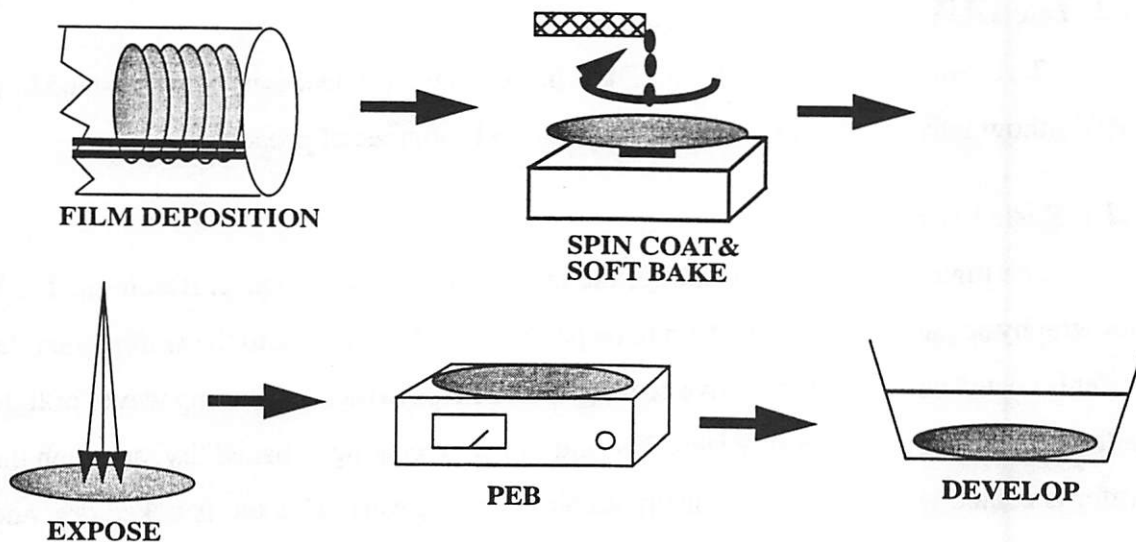
5.2.2. Likely Sources of Variability

As of yet, there have been no definitive studies as to what are the primary sources of CD variability in DUV lithography, but by stepping through our picture of the lithography sequence, we can determine likely sources of variability. Variations in the underlying film thickness (both within wafer and between wafer) change wafer reflectance, which

influences the amount of exposure energy that is coupled into the resist. Variations in the photoresist thickness, can also effect reflectivity, leading to a similar effect.

The light source in the stepper is an excimer laser, and each exposure consists of a discrete number of laser pulses, with each laser pulse having random variations in its energy. Thus, variations between the exposure dose set by the operator, and that actually given by the light source almost certainly play an important role in CD variability.

PEB may also have an important role in CD variance. Variations in bake plate temperature (both spatially across a wafer and between wafers) effect deprotection and diffusion, resulting in variation in CD.

Our picture of the DUV lithography drift is summarized by Figure 5.2.

Abnormal thickness of both resist and underlying films effect reflectivity.

Drifts in wafer reflectivity, photoresist chemistry and light source behavior alter the effective dose coupled to the resist.

PEB Temperature variations effect diffusion and deprotection.

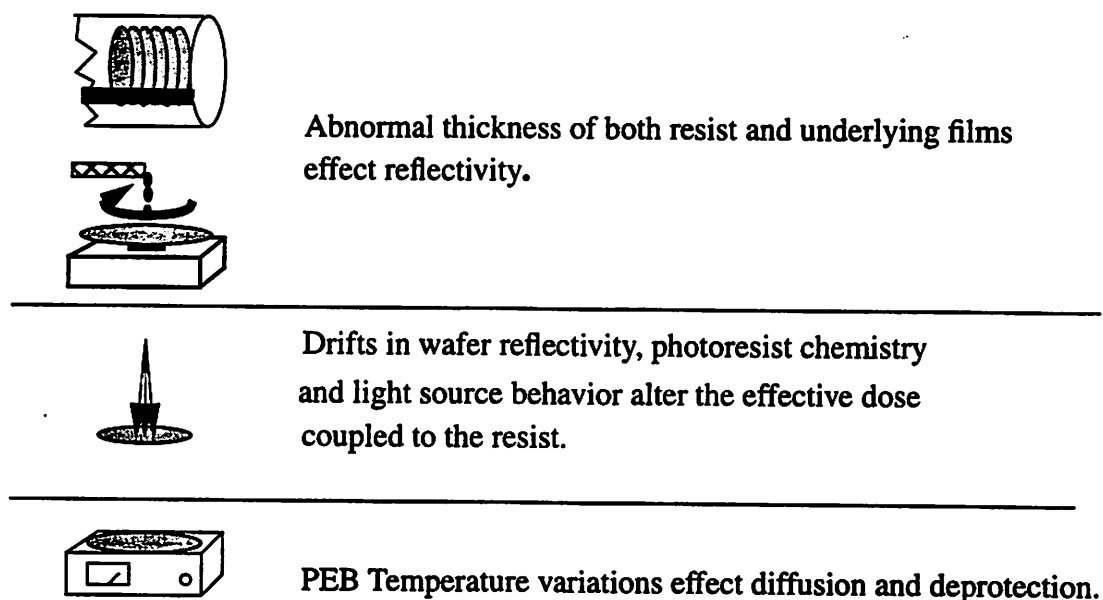Figure 5.2. DUV Lithography Drift

5.2.3. Available Metrology

Having outlined the possible sources of CD variation, we need to establish a means of measuring drifts in the process in order to design a RtR control loop. Ideally, we would like to measure the CD of every wafer before processing the next wafer. However, the only reliable way to measure CD is to use a scanning electron microscope (CD-SEM), and

unfortunately these measurements take too much time to be done between processing each wafer. Instead, The CD-SEM is most effective and has its highest throughput when measuring the CDs of a large batch of wafers. Therefore the CD-SEM cannot be used as the primary feedback of a wafer to wafer RtR control scheme, but can be used to asses the control performance after having processed a lot of wafers. Also, the CD-SEM can be used as the principle feedback of a RtR control scheme that makes adjustments at a lot to lot level -- as was demonstrated by Hershey, et. al. in [4].

One quantity that can realistically be measured in-line is wafer reflectance before exposure, using a small reflectometer such as that manufactured by SC Technologies. This measurement could capture reflectivity variation due to the spin coat process and the underlying film properties, but obviously cannot provide information about light source variation and PEB variation.

Another useful quantity that can be measured in-line is the change in thickness, or thickness loss, of the photoresist after PEB. In [7], Jakatdar shows by experiment that thickness is highly correlated to deprotection. Obviously, a measure of deprotection is extremely useful because it contains information about the cumulative effects of exposure and PEB disturbances. The potential use of this measurement is discussed in some detail later in this chapter.

5.3. Lot to Lot Control for DUV Lithography

RtR control on a lot to lot time scale has already been successfully implemented in industry. In this section, we review the results of the control implementation, and also extrapolate the lot to lot performance of EWMA and RDC control.

5.3.1. Industrial Results

Lot to lot RtR control was implemented at Motorola Semiconductor Corporation, and the implementation and results are discussed in [24], however the details of the control algorithm are unavailable. The controller worked by taking an average DICD (Developed Image Critical Dimension) from a lot, and using this measurement to compute an exposure dose for the subsequent lot -- trying to maintain the DICD as close as possible to a target DICD. In turn, a second, more slowly acting control loop adjusted the DICD target in an

attempt to account for changes in the etch bias, the difference between DICD and CD after etching.

The results presented in [24] indicate that the standard deviation of DICD before implementing the control was 9.4 nm, and after implementing control was reduced to 6.0 nm. The post-etch CD was reduced from 11.1 nm to 7.1 nm.

5.3.2. Results using EWMA and RDC

Because the details of the Motorola controller are unavailable, it is useful to try to estimate how control techniques like EWMA and RDC would have performed if they were used for lot to lot control on Motorola's process. Motorola provided us with a set of lot to lot data for this exercise.

The data set contained the exposure dose used for the lot and the DICD for the lot. Our analysis was done as follows. From the data, we computed the sensitivity of DICD to dose, A. By subtracting Au_k (Where u_k is the exposure dose of the kth lot) from each DICD data point we arrived at the sequence of offset terms b_k. Having a sequence of b_k terms and the sensitivity A, it is straightforward to reconstruct how any offset drift cancellation control would have performed on this lithography line. In our simulations, we add measurement noise, normally distributed with $3\sigma = 5$ nm, to the DICD to generate the hypothetical CD measurements, or y_k.
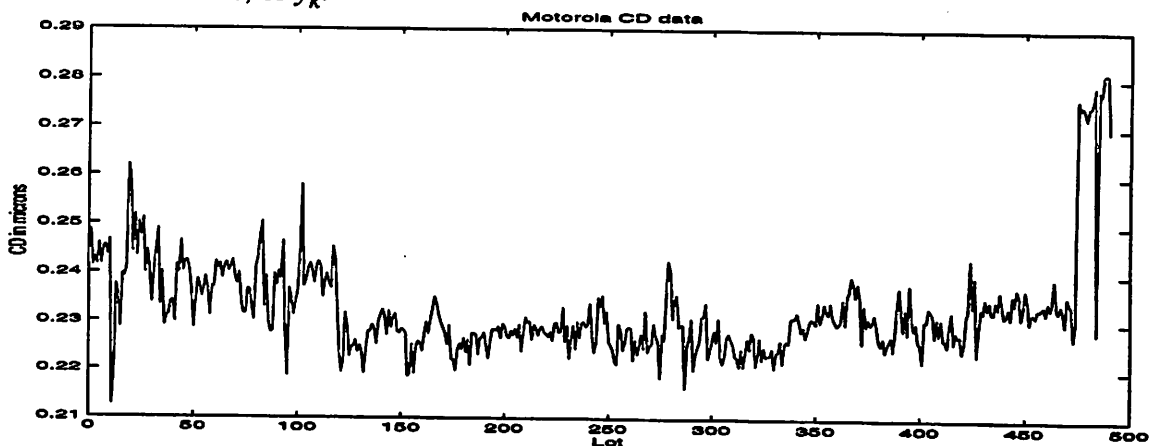
Figure 5.3. Motorola DICD Data

Figure 5.3 shows the DICD data supplied by Motorola. The target CD is 0.24, and the exposure dose is being adjusted from lot to lot by a controller. The standard deviation of the CD

here is 10.4 nm. It should be stressed that this data set is not the same data which was used for the results presented in [24]. Figure 5.4 shows CD vs. lot number, calculated to reflect what would have happened if the exposure dose were held fixed. The standard deviation of this data is 11.3 nm.

Figure 5.4. Lot to lot DICD had no controller been used

Figure 5.5 plots DICD vs. lot index when EWMA control is used ($w = 0.3$). The standard deviation has been reduced dramatically to 5.8 nm. Finally, the results of the RDC technique (using a window size of three wafers) are shown in Figure 5.6. Here, the standard deviation is 5.7 nm, nearly the same as for the EWMA control.
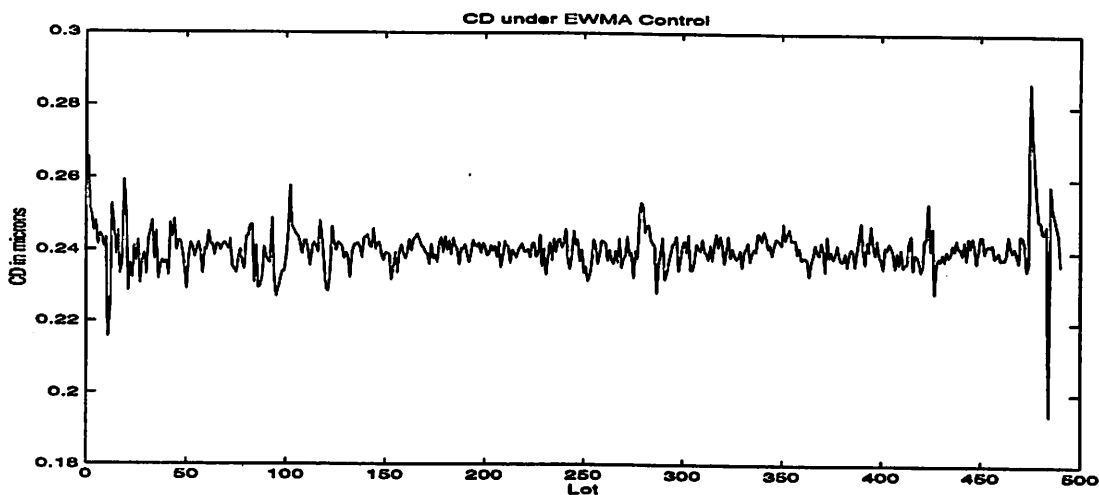
Figure 5.5. CD under EWMA control

44

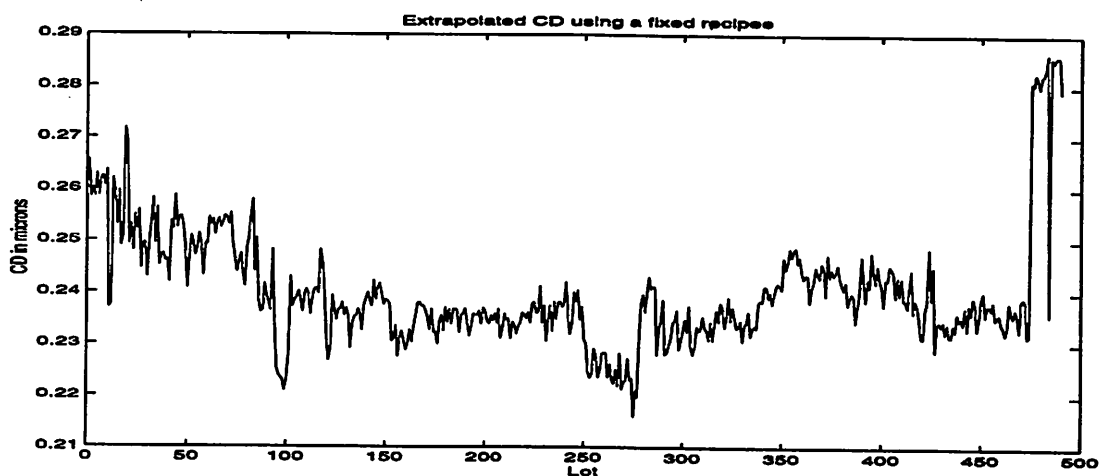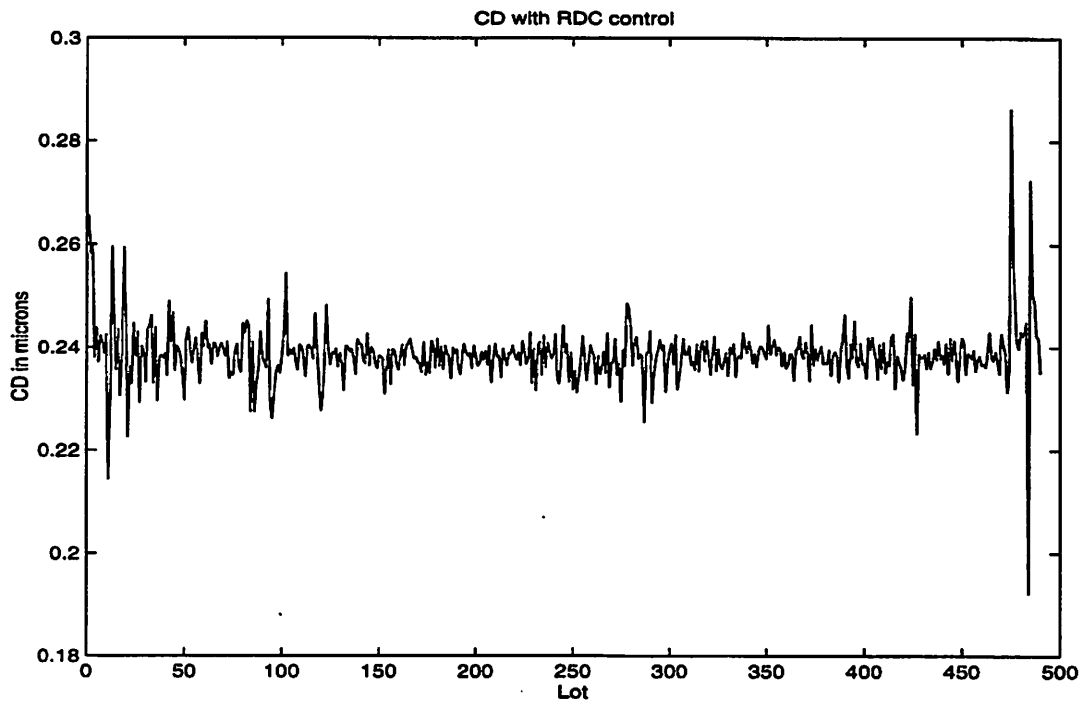Figure 5.6. CD under RDC control

The results of both our analyses, and the Motorola results from [24] indicate that significant lot to lot CD variability reductions can be achieved with RtR control. This is important because according to [24] lot to lot variability accounts for 30% of overall CD variance. Another 10% of the overall CD variance is due to variations between wafers. To reduce this portion of the variance, we must build a controller that acts on a wafer to wafer time scale. The development of such a controller is the topic of the remainder of this chapter.

5.4. Wafer to Wafer RtR Control Design Process

5.4.1. Lithography Modeling Experiment

To quantify the relationships between exposure dose, bake time, thickness loss, and CD, a modeling experiment was performed at National Semiconductor Corporation in Santa Clara, CA.

The experiment was carried out as follows. Four wafers were spun with 660 Angstroms of AR2 antireflective coating, and approximately 6550 Angstroms of UV5 deep

ultra-violet photoresist. Then one set of die on each wafer was exposed to CD test patterns, with the exposure dose varying across the wafer. To facilitate thickness loss measurements, another set of die on each wafer were given blanket exposures, again with a varying exposure dose. Figure 5.7 details how each die was exposed. To ensure that the blanket and adjacent patterned areas received comparable doses taking into account their different diffraction efficiencies, the blanket area doses were scaled down relative to the patterned area doses. Following exposure, the wafers were given a PEB, with each wafer receiving a different bake time as summarized in Table 5.1. After PEB, the photoresist thickness was measured in all blanket areas. These thicknesses were compared to thickness measurements taken before exposure to compute the thickness loss for each blanket exposed die. Finally, all of the wafers were developed and the CDs in each of the CD test pattern die was measured with a CD-SEM.

Wafer #	Post Exposure Bake Time (s)
1	105
2	75
3	90
4	90

Table 5.1. Post Exposure Bake Times used in the Modeling Experiment

5.4.2. Empirical Lithography Models

Several regression models were built using the measured thickness losses, CDs, and process input values. Square root terms were introduced into some of the models to improve the fit to the experimental data.

The results are summarized in Table 5.2 [22]. The first and second models relate thickness loss and CD (respectively) to the process: inputs bake time and exposure dose. The last model relates CD, the output we wish to regulate, to thickness loss, the quantity we can measure *in-situ*. The additional terms involving PEB time were added to the model to improve the fit to the experimental data. A physical justification for these terms is that

thickness loss does not capture the effects of acid diffusion during PEB, but these effects are strongly related to PEB time.

	Model	Adjusted R^2	Std. Eror
I.	$\sqrt{Thloss} = -23.4615 + 2.5176 Dose + 0.0334 PEBtime$	0.9802	0.22 $A^{0.5}$
II.	$CD = -0.0160 Dose + 0.0906\sqrt{PEBtime} - 0.0052 PEBtime$	0.9996	3.837 nm
III.	$CD = -0.0029 PEBtime + 0.0508\sqrt{PEBtime} - 0.000255 Thloss$	0.9995	4.216 nm

Table 5.2. Modeling Experiment Results
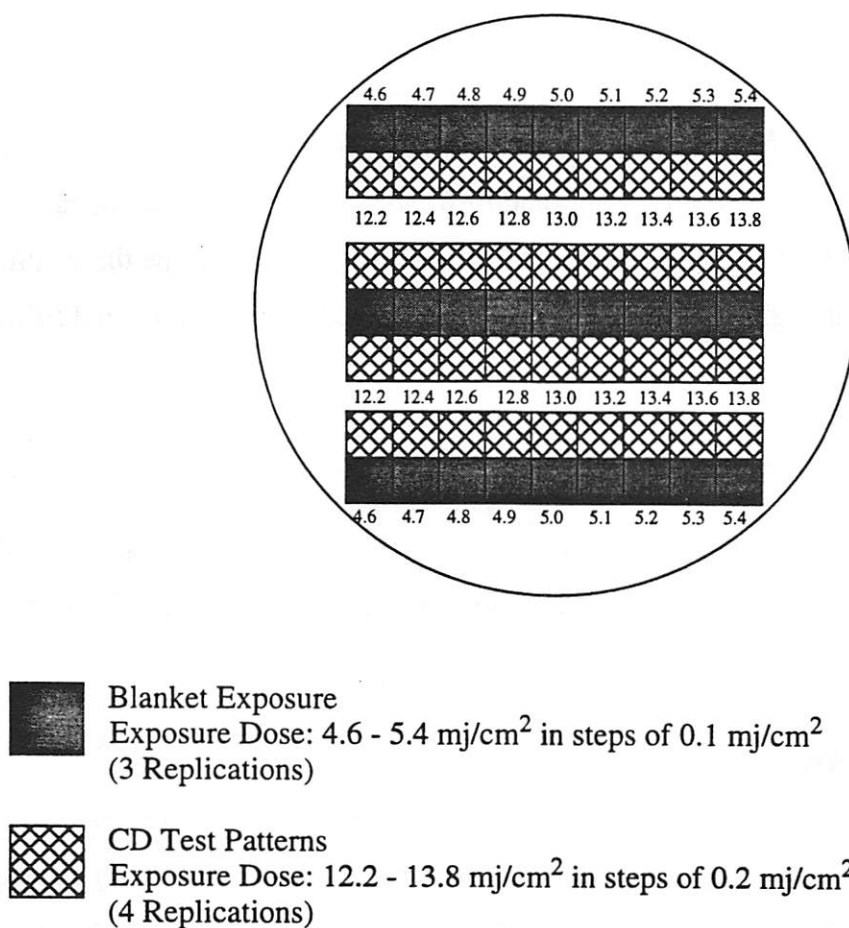

Figure 5.7. Modeling Experiment 8" Wafer Layout

Blanket Exposure
Exposure Dose: 4.6 - 5.4 mj/cm^2 in steps of 0.1 mj/cm^2
(3 Replications)

CD Test Patterns
Exposure Dose: 12.2 - 13.8 mj/cm^2 in steps of 0.2 mj/cm^2
(4 Replications)

5.4.3. DUV Lithography Drift Model

Having found static models relating the inputs and outputs of the DUV lithography process, it is now necessary to develop a model of how the process drifts. As stated in Section 5.1 the typical wafer-to-wafer 3σ variation is about 24 nm. The issue addressed in this section is where in our process model should we account for this variation.

In Section 5.2.2 we surveyed the likely sources of variability in DUV lithography. Some likely sources were effects like incoming wafer reflectivity variation causing effective dose fluctuations, and PEB temperature drifts which affect diffusion and deprotection.

These sources of variability as well as the others we discussed in Section 5.2.2 can be modeled (at least to first order) as additive noise on the process inputs -- exposure dose and PEB time.

$$Dose_{eff} = Dose + Noise_{dose} \qquad (5.1)$$

$$PEBtime_{eff} = PEBtime + Noise_{PEB} \qquad (5.2)$$

Here, we have introduced the terms $Dose_{eff}$ and $PEBtime_{eff}$ to denote the sum of each input setting and noise term. Our final models for thickness loss and CD will be those summarized in Table 5.2, but with $Dose_{eff}$ and $PEBtime_{eff}$ substituted wherever $dose$ and $PEBtime$ appeared.

$$\sqrt{Thloss} = -23.4615 + 2.517 Dose_{eff} + 0.0334 PEBtime_{eff} \qquad (5.3)$$

$$CD = -0.0160 Dose_{eff} + 0.0906\sqrt{PEBtime_{eff}} - 0.0052 PEBtime_{eff} \qquad (5.4)$$

Finally, we assume that some measurement noise is incurred in measuring thickness loss.

$$Thloss_{meas} = Thloss + noise$$

5.4.4. RtR Control Architecture

The control strategy is straightforward: measure the photoresist thickness loss after PEB, estimate the post-develop CD, and use this estimate in conjunction with a standard RtR control algorithm to prescribe a recipe for the subsequent wafer. A schematic of the control architecture is shown in Figure 5.8, and a summary of notation we will use in our controller development is presented in Table 5.3.
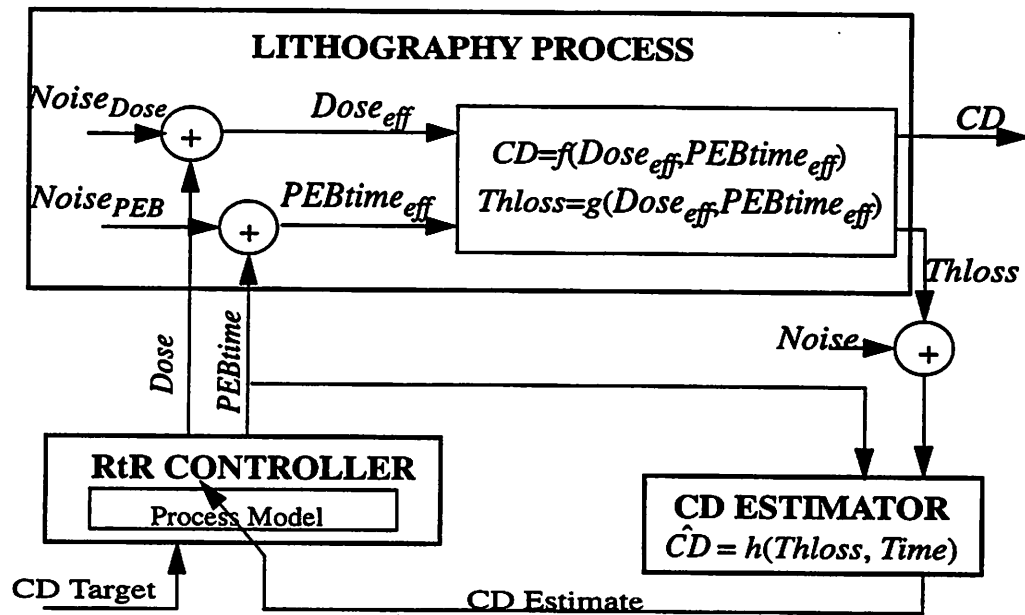
Figure 5.8. RtR Control Architecture for DUV lithography

Definition	Explanation
T = CD Target	Desired CD
$y^o_k = CD_k$	True CD of wafer k
$y_k = \hat{CD}_k$	CD estimate of wafer k
$e_k = y_k - y^o_k$	CD Estimation error of wafer k
$u_k = \begin{bmatrix} Dose_k \\ PEBtime_k \end{bmatrix}$	Input vector for wafer k

Table 5.3. Control Architecture Notation

The CD Estimation block of Figure 5.8 takes a presumably noisy thickness loss measurement and the wafer's PEB time and applies Model III. of Table 5.2 to arrive at an estimate for CD.

$$y_k = \hat{CD}_k = -0.0029PEBtime_k + 0.0508\sqrt{PEBtime_k} - 0.000255Thloss_k \tag{5.5}$$

The target CD and CD estimate of the previous wafer are fed into the RtR Controller block of Figure 5.8. In Section 5.4.3, we arrived at a drift model in which drift terms add to the exposure dose and PEB time. Under this drift model, the sensitivies to the process inputs are fixed to first order (If we were to write (5.5) as a taylor series about a nominal *PEBtime* and nominal *dose*, the drift terms would multiply only with second order and higher terms of *PEBtime* and *dose*.) Therefore the drift is primarily an *offset drift*, and thus we use an offset cancellation RtR control technique. In our first design, we choose the simplest in this class of techniques, EWMA. Recall from Chapter 4, an EWMA RtR control design assumes a process model of the form.

$$y_k^o = Au_k + b_k \tag{5.6}$$

Our process model will be a modification of the process model obtained from our modeling experiment (Model II. from Table 5.2). We define the function A ( ) using our empirical model:

$$A\left(\begin{bmatrix} Dose_k \\ PEBtime_k \end{bmatrix}\right) = -0.0160Dose + 0.0906\sqrt{PEBtime_k} - 0.0052PEBtime_k \tag{5.7}$$

We then add a drifting offset term to (5.7) to create the process model for our EWMA control design.

$$y_k^o = A(u_k) + b_k \tag{5.8}$$

The offset term b_k used by the EWMA controller accounts for the cumulative effects of PEB and exposure drifts on CD. We could modify (5.8) to be affine by defining the input vector to have a $\sqrt{PEBtime_k}$ term, but it is more convenient to leave the equation in its present form. The b_k term is recursively estimated according to the equation:

$$\hat{b}_{k+1} = (1-\omega)\hat{b}_k + \omega(y_k - A(u_k)) \tag{5.9}$$

where ω is the EWMA weight and $\hat{b}_k$ is the estimate of b_k. The input setting is selected by finding a u_k that will meet the target output T by cancelling the estimated drift:

$$T = A(u_k) + \hat{b}_k \qquad (5.10)$$

The choice of u_k that satisfies (5.10) is not unique, so two practical design consid-
erations are used to select a u_k. First, we would like to stay within the region of the input
space which was used in the modeling experiment. Second, the exposure dose and PEB
time have a minimum step size of 0.1 mj/cm^2 and 1 second respectively. By trial and error
using the simulator to be discussed in Section 5.5, we arrive at an input selector algorithm
that performs well with the above input limitations. The algorithm is outlined in Table 5.4.
Though this input selector is quite problem specific, it illustrates how one can incorporate
input limitations into a control design using ad-hoc rules.

---

1. Assume a nominal PEB time of 91 seconds and calculate an exposure dose, α_k
   that satisfies (5.10).

$$T = A\left(\begin{bmatrix} \alpha_k \\ 91\,\text{seconds} \end{bmatrix}\right) + \hat{b}_k$$

2. Scale α_k according to its difference from the nominal dose of 13.2 mj/cm^2 to
   form $\widetilde{dose}_k$.

$$\widetilde{dose}_k = 0.85\left(\alpha_k - 13.2\frac{\text{mj}}{\text{cm}^2}\right) + 13.2\frac{\text{mj}}{\text{cm}^2}$$

3. Round $\widetilde{dose}_k$ to the nearest multiple of 0.1 mj/cm^2 and apply hard limits of 12
   mj/cm^2 and 14mj/cm^2 to yield $dose_k$.

4. Find an appropriate PEB time for the selected $dose_k$.

$$T = A\left(\begin{bmatrix} dose_k \\ \widetilde{PEBtime}_k \end{bmatrix}\right) + \hat{b}_k$$

5. Round $\widetilde{PEBtime}_k$ to the nearest second, and apply hard limits of 75 seconds
   and 105 seconds to give $PEBtime_k$.

Table 5.4. Input Selection Rules

5.5. RtR Control Simulation

Having designed a RtR controller for DUV lithography, we would now like to simulate the controller to assess its potential performance. In this section, we develop an architecture for simulating the closed loop system, outline the parameter selections and outcome of a baseline case, and then evaluate the change in performance when parameters are perturbed.

5.5.1. Simulator Architecture
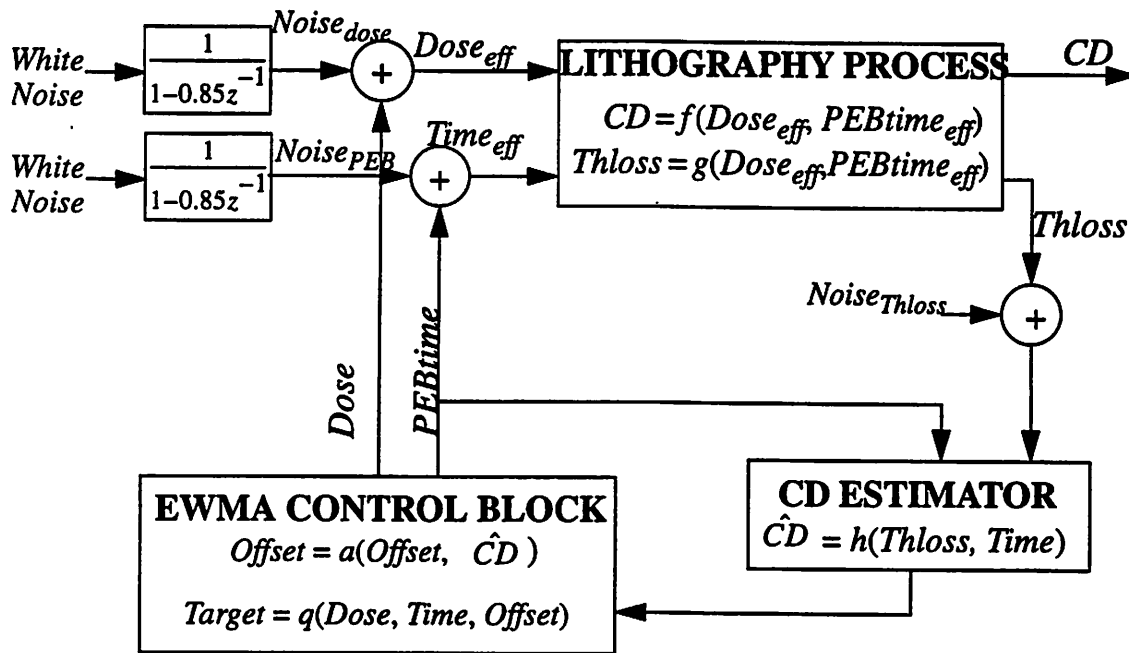
Figure 5.9. Simulation Architecture

Figure 5.9 outlines the simulation architecture. The Simulator simulates three blocks: the drifting lithography process, the CD estimator, and the EWMA controller. The CD estimator and EWMA Control blocks are implemented with the equations given in Section 5.4.4. Simulating the lithography process block also requires specifying a stochastic drift model, which is the purpose of this section.

The lithography process block takes exposure dose and PEB time as inputs, and outputs CD and Thickness Loss. The model for the lithography process is that described by Section 5.4.3. Thus, $f(\ )$ and $g(\ )$ of Figure 5.9 are the models for CD and thickness loss

52

found from the modeling experiment, with the substitution of a $Dose_{eff}$ and $PEBtime_{eff}$ for $Dose$ and $PEBtime$.

We must assume that $Noise_{dose}$ and $Noise_{PEB}$ has some autocorrelation structure between runs. Because we do not know the exact autocorrelation structure we will encounter in the real lithography process, by default we assume the simplest model possible, a first order AR. The modeling technique is that illustrated in Section 2.3.1. We define

$$x_k = \begin{bmatrix} \{Noise_{dose}\}_k \\ \{Noise_{PEB}\}_k \end{bmatrix} \tag{5.11}$$

Now suppose we expect the variance of $Noise_{dose}$ to be σ^2_{dose} and the variance of $Noise_{PEB}$ to be σ^2_{PEB}. Also suppose we expect the autocorrelation coefficient between successive runs to be $P_{d,1}$ and $P_{d,2}$ for $Noise_{dose}$ and $Noise_{PEB}$ respectively. With the following definitions:

$$K = \begin{bmatrix} \sigma^2_{Dose} & 0 \\ 0 & \sigma^2_{PEB} \end{bmatrix}, F = \begin{bmatrix} P_{d,1} & 0 \\ 0 & P_{d,2} \end{bmatrix}, G = \sqrt{K - FKF^*} \tag{5.12}$$

the stochastic process for x_k is

$$x_{k+1} = Fx_k + Gw_k \tag{5.13}$$

Where w_k is a 2 dimensional gaussian random vector with identity covariance. As indicated by Figure 5.9, the random process generated by (5.13) can be generated by passing white noise through two filters, each with a pole at $P_{d,i}$ and a zero at 0.

The thickness loss measurement error, $Noise_{Thloss}$, is assumed to be white and normally distributed with variance σ^2_{Thloss}.

5.5.2. Baseline Simulation

Section 5.5.1 described the basic architecture for our simulations, but left many parameter values unspecified. In this section, we will pick a set of parameter values that, to the best our knowledge, characterizes how the DUV sequence typically behaves. After selecting baseline parameters, we simulate the baseline case and evaluate control perfor-

mance. In Section 5.5.3 we explore what happens when the baseline parameters are perturbed.

5.5.2.1. Parameter Selection

The baseline parameter choices are summarized in Table 5.5. σ^2_{dose} and σ^2_{PEB} were selected to give a CD variation of 8 nm. Clearly, many combinations of these two variances would yield a CD variance of 8 nm. The values chosen as the baseline values are the best guesses from a experienced lithography engineer [23]. His choices were primarily based upon the typical variation in the optimal exposure dose, as periodically measured by test wafers with focus/exposure matrices. σ^2_{Thloss} was selected to be two times the known measurement error variance of the reflectometer used for the thickness loss measurements (Two times the measurement error because two uncorrelated measurements are subtracted to compute thickness loss.)

Parameter	Parameter Description	Baseline Value
σ_{dose}	Effective Dose Variance	0.4 mj/cm^2
σ_{PEB}	Effective PEB Time Variance	7.1 seconds
σ_{Thloss}	Thickness Loss Measurement Error Variance	13 Angstroms
$P_{d,1}$	Drift Model Parameter from (5.12) & (5.13) (The correlation coefficient between successive $\{Noise_{Dose}\}_k$ terms.)	0.85
$P_{d,2}$	Drift Model Parameter from (5.12) & (5.13) (The correlation coefficient between successive $\{Noise_{PEB}\}_k$ terms.)	0.85
ω	EWMA Weight	0.6

Table 5.5. Baseline Parameters

At the time of this writing, no wafer to wafer drift data is available from the National DUV Pilot line to compute the wafer drift statistics. Thus, the values of the drift model parameters $P_{d,1}$ and $P_{d,2}$ are selected by analysis of CD data from a different source, a DUV lithography line at Motorola in XXX, TX. A first order AR model was fit to the data set, and the resulting pole was located at around 0.85. Thus, we concluded that 0.85 is a

reasonable choice for $P_{d,1}$ and $P_{d,2}$. However, the CD data from Motorola was taken on a lot to lot basis rather than wafer to wafer. So we are essentially assuming that the drift statistics on a wafer to wafer time scale at National Semiconductor are similar to the lot to lot statistics at a different factory, of a different manufacturer, making a different product. Clearly, this is a very bold assumption and must be revisited as soon as a large set of wafer to wafer DUV data becomes available from National Semiconductor.

The EWMA is selected to be 0.6, because this value seemed to give relatively good performance for both the baseline case, and cases where the baseline parameters are perturbed.
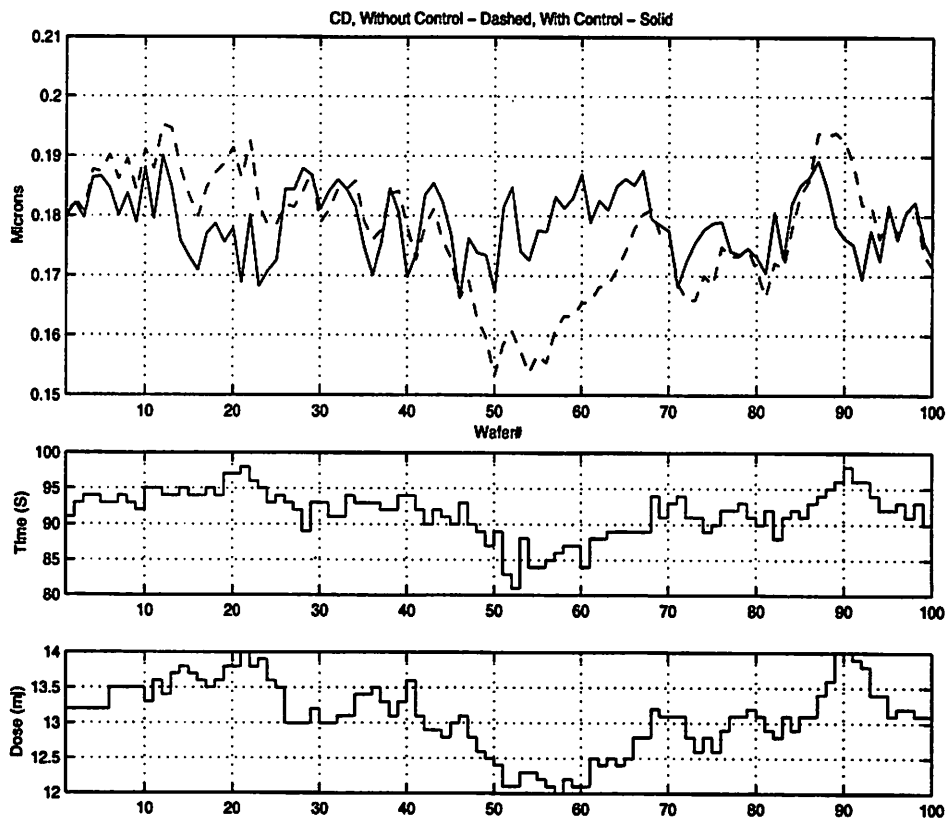
5.5.2.2. Baseline Simulation Results

Figure 5.10. Baseline Simulation Performance and Inputs for 100 wafers.

The baseline case was simulated for a sequence of 1000 wafers. Figure 5.10 illustrates the results of the simulation. In the first plot, the solid trace represents the CD of the first 100 wafers using the EWMA controller, while the dashed trace shows what the CD would have been had the input been held fixed. The variation in with-control trace varies noticeable less than the open-loop trace. The other plots of Figure 5.10 show the inputs that were used by the controller. Note that the inputs stay within the range of 80 - 100 seconds and 12 - 14 mj / cm^2.

Table 5.6 quantifies the relative performance of the EWMA control vs. no control. The statistics -- RMS error and mean -- are evaluated using all 1000 wafers.

	Mean CD	RMS Error
Without Control	0.1785 μm	0.0071 nm
With Control	0.1804 μm	0.0052 nm

Table 5.6. Control Performance Statistics

To obtain a more concrete picture of what the drift models in this simulation are doing, the $Noise_{Dose}$ and $Noise_{PEB}$ sample paths used in the baseline simulation for the first 100 wafers are shown in Figure 5.11. The sample paths appear to be quite noisy but do exhibit autocorrelation, which is indeed the type of drift we hoped to synthesize with our drift model.
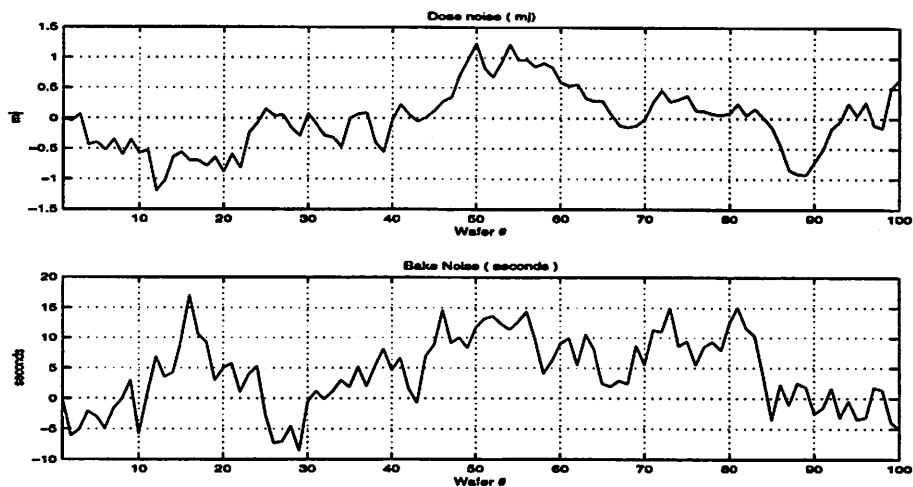
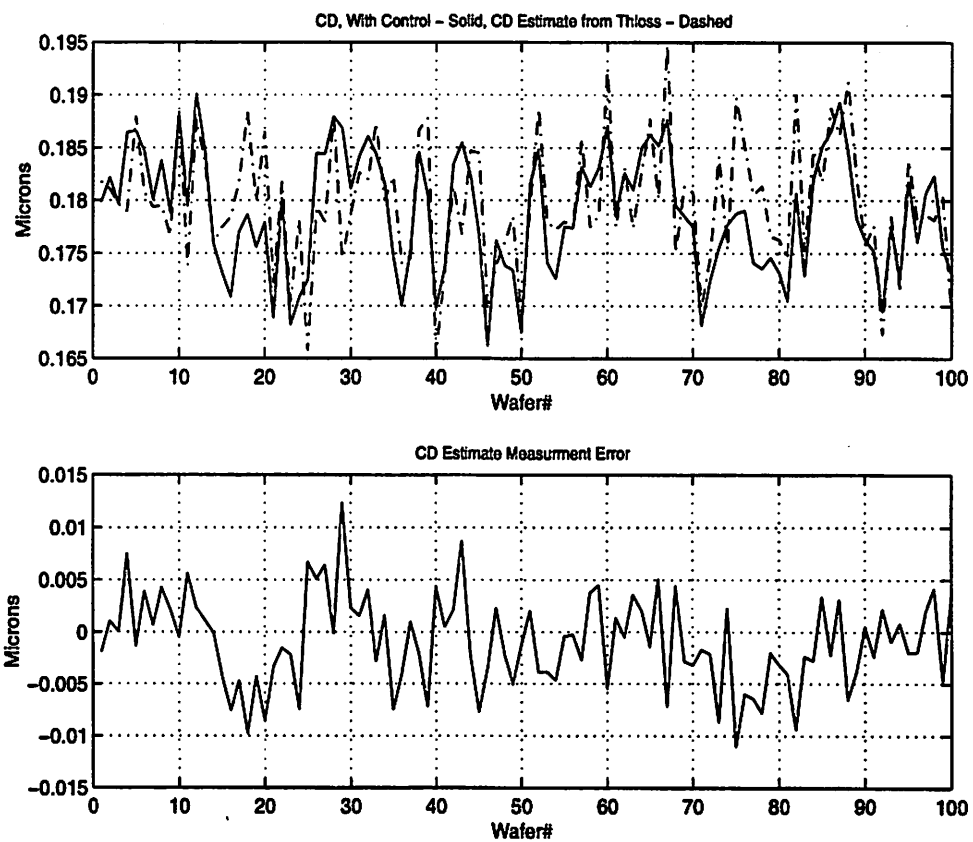Figure 5.11. Dose and Bake Noise Sample Paths for Baseline Simulation.

Figure 5.12. CD Estimator Performance in Baseline Simulation. RMS error = 4.7 nm

57

Figure 5.12 shows the performance of the CD Estimator block in the baseline simulation. The first plot shows the actual CD and the estimated CD for the first 100 wafers, while the second plot shows the difference between the two, the estimation error. If one compares this second plot with the traces in Figure 5.11, one notices some correlation between the estimation error and the PEB noise (For example $Noise_{PEB}$ has a peak at 17 seconds while the estimation error has a trough, also at 17 seconds.) An explanation for this can be found by examining (5.5), the equation describing the CD estimator. In (5.5), the CD estimate is coupled to the PEB time through all three terms, $PEBtime_k$, $\sqrt{PEBtime_k}$, and also by the $Thloss_k$ term, which is a function of the effective bake time in the simulation. Changes in the effective bake time from $\{Noise_{PEB}\}_k$ can be "seen" by the estimator through the $Thloss_k$ term, but not through the other terms. So a large, positive, $Noise_{PEB}$ will decrease the estimated CD by a smaller amount than it decreases the true CD, resulting in an increased CD estimation error.

5.5.3. Simulation of Perturbations in the Drift Model

Having established a baseline simulation, in this section we investigate the effects of parameter perturbations on performance.

We first explore variations in $P_{d,1}$ and $P_{d,2}$, the correlation coefficient between successive $\{Noise_{Dose}\}_k$ and $\{Noise_{PEB}\}_k$ terms respectively. In the baseline case, we set both $P_{d,1}$ and $P_{d,2}$ to be 0.85, but recall we had very limited confidence in our choices in $P_{d,1}$ and $P_{d,2}$ because the choices were made by analyzing lot-to-lot data taken from a completely different lithography setup. Here we will explore an array of possible values for $P_{d,1}$ and $P_{d,2}$. In this first parameter exploration we vary both $P_{d,1}$ and $P_{d,2}$ between 0.5 and 0.95 in steps of 0.05. For each $(P_{d,1}, P_{d,2})$ pair, we run our simulator with a sequence of 1000 wafers, and set all other parameters to the baseline values.
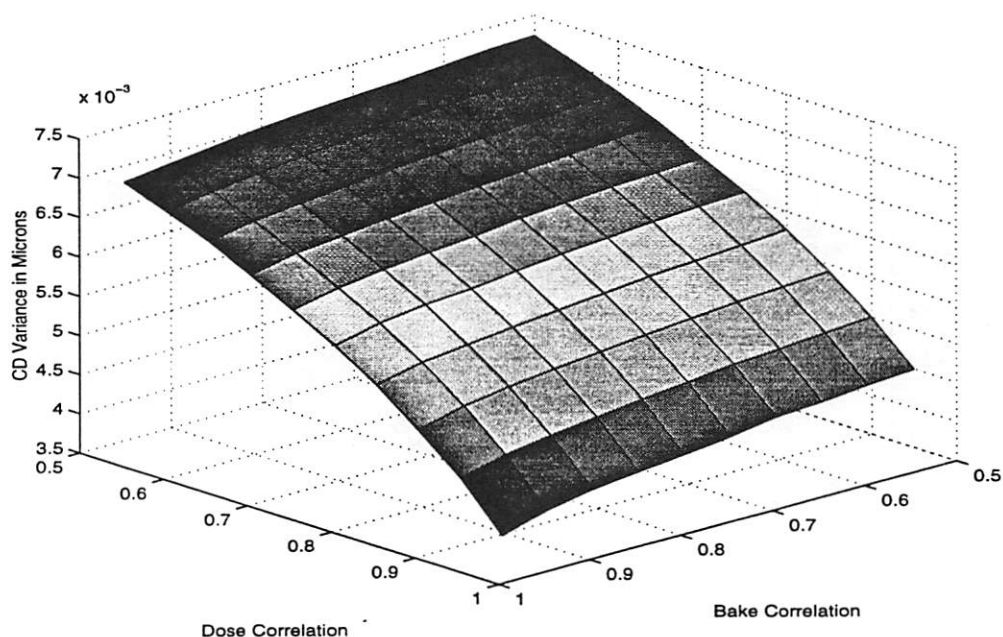
Figure 5.13. CD Variance with respect to $P_{d,1}$ and $P_{d,2}$ other parameters set to baseline.

Figure 5.13 shows the results of the first simulation. The x and y axis show the values of $P_{d,1}$ and $P_{d,2}$ while the CD variance is shown on the z axis. As expected, the higher values of $P_{d,1}$ and $P_{d,2}$, or more autocorrelation in the process drift, result in reduced CD variability. In this simulation, the CD variance is most strongly dependent on the Dose correlation, $P_{d,1}$. This is probably because the dose variance constitutes the larger portion of the overall process variability. Thus under the assumptions of this simulation, good control performance requires strong autocorrelation in the Dose Noise.

The second parameter exploration is similar to the first. Again, we test an array of different $(P_{d,1}, P_{d,2})$ pairs, but now we change σ^2_{dose} from 0.4mj/cm^2 to 0.3 mj/cm^2 and σ^2_{PEB} from 7.16 s to 9.55 s. Like the original values for σ^2_{dose} and σ^2_{PEB}, the new values result in a open loop CD variation of 8.0 nm σ. However in the situation described by the new parameter choices, the PEB Noise constitutes an increased portion of the overall variability. All other parameters were set to the baseline values.
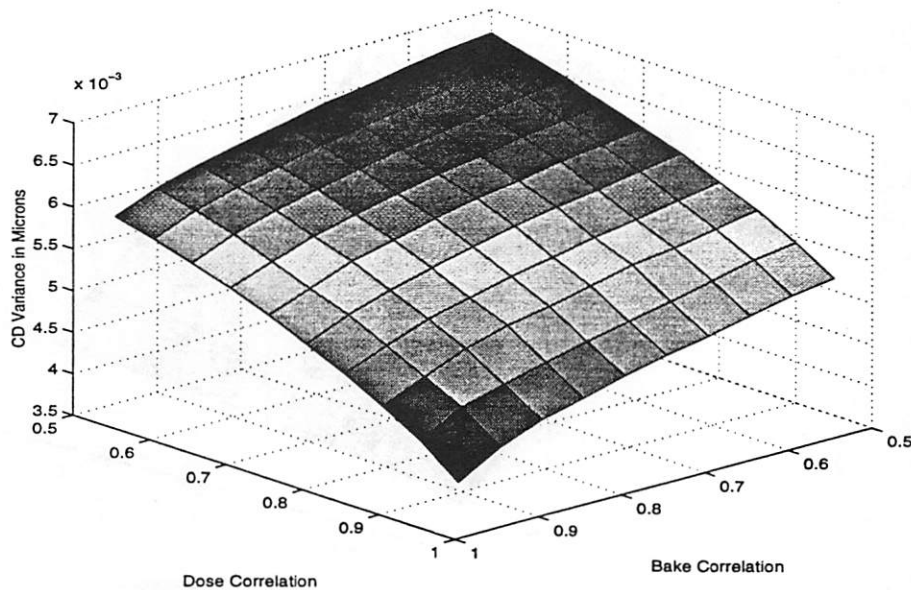
Figure 5.14. CD Variance with respect to $(P_{d,1}, P_{d,2})$

$$\sigma^2{}_{dose} = 0.3 \text{ mj/cm}^2, \quad \sigma^2{}_{PEB} = 9.55 \text{ s.}$$

Figure 5.14 shows the results of this second simulation. As before, higher values of $P_{d,1}$ and $P_{d,2}$, or more autocorrelation in the process drift, result in reduced CD variability. However, in the first simulation CD variance was much more dependent on $P_{d,1}$, but in this simulation CD variance is almost equally dependent on the values of $P_{d,1}$ and $P_{d,2}$. This should be expected because we have increased the contribution of Bake Noise to overall variability.

Our parameter exploration leads us to two conclusions. One is that CD variance reduction can be achieved for a large range of parameters and not just the ones used in the baseline case. The second point is that the potential performance of RtR control is extremely dependent upon the autocorrelation of the process drift.

5.5.4. Other Relevant Properties of the Process Drift

Though the degree of autocorrelation in the process drift is extremely important in determining the potential impact of RtR control, other properties of the process drift have a role as well. One property in particular is how long it takes the process drift to exhibit its autocorrelation. For example, suppose we have a process that undergoes step changes (perhaps from maintenance cycles) in a scalar offset term, b_k, with some additional white noise. A simple wide sense stationary model for this behavior is:

$$b_k = f(k+V) + w_k \qquad (5.14)$$

$$f(\ ) = \left\{ \begin{array}{ll} -1 & k \ mod \ 1000 \in [0, 499] \\ 1 & k \ mod \ 1000 \in [500, 999] \end{array} \right. \qquad w_k \sim N(0, 0.1) \qquad (5.15)$$

where V is a discrete random variable, uniformly distributed on [0,999]. Simple calculations show that the autocorrelation coefficient between successive runs is:

$$\frac{E[b_k b_{k+1}]}{E[b_k^2]} = \frac{1 - \frac{2}{500}}{1 + \sigma_w^2} \approx 0.90 \qquad (5.16)$$

From the value of this autocorrelation, we might expect that this process would be an excellent candidate for RtR control. However, if we think more carefully about (5.15), we see that the process is basically white noise on the scale of 10's or 100's of runs, and periodically shifts means on a very long time scale. If we were to run a RtR controller over something like 50 runs, it is extremely likely that the controller would actually worsen process variance because it would basically be operating on a process with white noise drift. If the RtR controller ran for 1000's of wafers, then we would see a substantial variance reduction, but only because the RtR controller would "re-center" the process every time a step change occurred. Such basic changes could much more easily done by an equipment operator -- recalibrating the tool after each maintenance cycle for example.

What this example illustrates is that for a process to benefit from RtR control, it not only must have high autocorrelation, and but also it must exhibit its autocorrelation on a reasonable time scale.

Chapter 6 Conclusions

6.1. The Need for RtR Control

The economic pressure to produce ICs with denser layouts and smaller feature sizes will continue to drive manufacturers to reduce variability in critical processes such as lithography and plasma etch for the forseeable future [25]. In this work, we have seen that RtR control can be extremely effective at reducing variability, and because it requires much fewer equipment changes than redesigning a process, it can achieve variability reduction at a comparitively low cost.

6.2. On the Design of RtR Control

The proper design of a RtR control algorithm is a subtle, an interesting engineering problem in itself. In this work we have emphasised one approach to solving these problems, which is to formulate the problem in the context of random processes, and to build controllers based upon standard filtering techniques. This approach ultimately leads to the techniques emphasisized in this work: EWMA, RDC, and Kalman Filtering. In the development of all these techniques, we have seen that the design of an effective RtR Control balances two objectives: the need to aggresively compensate for drifts in a process, and the need to not introduce more variability into a process by misjudging how the process has drifted. The advantage of designing a RtR controller in the context of random processes is that it leads to mathematical formulas that tell us where the best balance between these two objectives lies, in terms of first order statistics of the process drift.

6.3. Practical Considerations

In Chapter 5, our study of RtR control for DUV Lithography, we saw that not a clever mathematical technique for designing a RtR control was not enough. What we saw is that the

key for developing an effective RtR control is to have a metrology that samples the process as often as possible and which measures a variable that is or is closely linked to the parameter we are trying to control. However such a metrology does not yet exist in DUV lithography. If we want to regulate CD -- we can measure it directly with a CD-SEM between every lot, or we can estimate it indirectly for each wafer using thickness loss. In Chapter 5 we explored RtR controllers based on both approaches, and saw that both control on a lot to lot and wafer to wafer time scale led to modest variability reductions in CD.

6.4. Issues for Further Research

A RtR controller wich uses both metrologies might be the ideal control for DUV lithography. Such a controller would have to blend multi-rate information: lot-to-lot CD-SEM measuremnts and wafer-to-wafer thickness loss measurments. How to best cope with this multiple time-scale problem would be an interesting topic for future research. Multiple time-scale RtR control is almost certainly applicable to other semiconductor manufacturing process where the wafer-to-wafer information is noisier than the lot-to-lot information.

Another important area of future research is in the development of new metrologies. For example in DUV lithography, measuremetnt of drifts in develop rate, de-focus, and bake plate temperature could compliment thickness-loss information to yield a better CD estimate [26]. Additional metrology, such as a de-focus measurment, may also allow a RtR controller to focus as well as bake time and dose. This would be usefull becuase a RtR control that neglects focus may improve CD variability at the expensive of increased variability in other properties, such as side wall angle.

One extension of basic RtR control that is applicaple to DUV is to develop an end-point controller for PEB. Because the thickness loss metrology we discussed in Chapter 5 is capable of estimating CD just after PEB, one possible control scheme would be to stop the PEB process as soon as the thickness loss metrology indicates the CD is on target. In principle, this should ensure that CD variability is reduced to simply the metrology variability.

Another extension to simple RtR control worth future work is to develop controllers which deal with spatial variability. CD variability within a wafer is just as significant as CD

variability between wafers [24]. Dealing with this problem will require a spatially resoilved actuation, perhaps by using a spatially resolved bake plate [27] , or by using different exposure doses for different fields (though much variability is actually *within* field.) Controling systems with so many inputs, and so much coupling between the inputs, would be a very challenging control problem.

6.5. Closing Remarks

Clearly RtR control will be of crucial, and growing importance to the industry in the coming years, and the development of better control techniques will continue to challenge researchers for the forseeable future.

References

[1] C. Schaper, M. Moslehi, K. Saraswat, T. Kailath, "Temperature Control of MMST RTP: repeatability, uniformity, and integration forflexible manufacturing," *IEEE Transactions on Semiconductor Manufact.*, vol.7, no.2, pp 202-19, May 1994.

[2] M. Hankinson, T. Vincent, K. Irani, and P. Khargonekar, "Integrated Real-Time and Run-to-Run Control of Etch Depth in Reactive Ion Etching," *IEEE Trans. Semiconductor Manufact.*, vol. 10, no. 1, pp 121 - 30, Feb. 1997.

[3] S. Leang, S-Y Ma, J. Thompson, B. Bombay, and C. Spanos, "A Control System for Photolithographic Sequences," *IEEE Trans. Semiconductor Manufact.*, vol. 9, no. 2, pp 191 - 207, May 1995.

[4] D. Gerold, R. Hershey, K. McBrayer, and J. Sturtevant, "Run-to-Run Control Benefits to Photolithography," Sematech AEC/APC Workshop IX, Incline Village, NV, Sept. 1997.

[5] D. Bonning, W. Moyne, and T. Smith, "Run by Run Control of Chemical-Mechanical Polishing," 1995, IEEE CPMT Int'l Electronics Manufacturing Technology Symposium, 1995.

[6] R. Rosenthal, P. Solomon, S. Charpenay, A. Bonanno, W. Zhang, and W. Eiklebarry, "Run-to-Run Control of a Single Wafer Epitaxial Silicon Fabrication Process," 1997 Sematech AEC/APC Symposium, Incline Village, NV, Sept. 1997.

[7] N. Jakatdar, "In-Situ Metrology for Deep Ultra Violet Lithography Control," M.S. Thesis, Dec. 1997.

[8] S. Haykin, *Adaptive Filter Theory*, 3rd Ed., Englewood Cliffs, NJ: Prentice Hall, 1996.

[9] L. Ljung, *System Identification: Theory for the User*, Englewood Cliffs, NJ: Prentice-Hall, 1987.

[10] B. Anderson, J. Moore, *Optmal Filtering*, New Jersey: Prentice Hall 1979.

[11] T. Kailath, *Lectures on Wiener and Kalman Filtering*, New York: Springer Verlag, 1981.

[12] R. Gallager, "Estimaton," Class Notes for Electrical Engineering 226A at University of California, Berkeley, 1996.

[13] S. Orfanidis, *Optimum Signal Processing: an Introduction*, New York: Macmillan, 1985.

[14] E. Sachs, A. Hu, and A. Ingolfsson, "Run by Run Process Control: Combining SPC and Feedback Control," *IEEE Trans. Semiconductor Manufact.*, vol. 8, no. 1, pp 26 - 43, Feb. 1995.

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

[16] S. Boyd, et. al., *Linear Matrix Inequalities in System and Control Theory*, Philadelphia : Society for Industrial and Applied Mathematics, Series title: SIAM studies in applied mathematics, 1994.

[17] J. Musacchio, S. Rangan, C. J. Spanos, and K. Poolla, "On the Utility of Run to Run Control in Semiconductor Manufacturing," Int'l Symp. on Semicon. Manufact., San Francisco, CA, Oct. 1997.

[18] E. Palmer, W. Ren, C. Spanos, and K. Poolla, "Control of Photoresist PropertiesL A Kalman Filtering Based Approach," *IEEE Transactions on Semiconductor Manufact.*, vol. 9, no. 2, pp 208-14, May 1996.

[19] J. Peterson, C. Mack, J. Thackeray, R. Sinta, T. Fedynyshyn, J. Michael, J. Byers, and D. Miller, "Characterization and Modeling of a Positive Acting Chemically Amplified Resist," SPIE Vol. 2438, 1995.

[20] M. Zuniga, G. Walraff, and A. R. Neureuther, "Reaction Diffusion Kinetics in Deep-UV Positive Tone Resist Systems," SPIE Vol 2438, 1995.

[21] M. Zuniga, and A. R. Neureuther, "Reaction-Diffusion Modeling and Simulations in Positive DUV Resists," *Journal of Vacuum Science & Technology B* (Microelectronics and Nanometer Structures), vol.13, no.6, pp2957-62, Nov.-Dec. 1995.

[22] N. Jackatdar, J. Musacchio, and C. Spanos, "Control Models for DUV Workcell," SRC Deliverable Report, April 1998.

[23] N. Jackatdar, Personal Communication, Jan. 1998.

[24] J. Sturtevant, M.R. Weilemann, K. G. Green, J. Dwyer, E. Robertson, R.R. Hershey, "Implementation of a Closed-Loop CD and Overlay Controller for Sub - 0.25 - μm patterning," SPIE Vol. 3332, 1998.

[25] Semiconductor Industry Association National Technology Roadmap for Semiconductors 1997 Edition.

[26] M. Classen, "Study on Real Time Develop Monitors for DUV Lithography," M.S. Thesis, Dec. 1998.

[27] K. El-Awady, C. Schaper, and T. Kailath, "Thermal Cycle Module for Photoresist Processing," Sematech AEC/APC Symposium X, Vail, CO, Oct. 1998.