SPECIAL ISSUES IN SEMICONDUCTOR MANUFACTURING

III

EECS 290W Class Projects Reports, Spring 1992

Professor:

Costas J. Spanos

Students:

Bart Bombay, Sean Cunningham, Zeina Daoud,
John Helmsen, Joe King, Hao-Cheng Liu, Dave Newmark,
Jorge Noriega-Asturias, John Thomson, and Crid Yu

Also in class:

Eric Boskin and Debra Hebert

Memorandum No. UCB/ERL M92/84

3 August 1992

SPECIAL ISSUES IN SEMICONDUCTOR MANUFACTURING

III

EECS 290W Class Projects Reports, Spring 1992

Professor:

Costas J. Spanos

Students:

Bart Bombay, Sean Cunningham, Zeina Daoud,
John Helmsen, Joe King, Hao-Cheng Liu, Dave Newmark,
Jorge Noriega-Asturias, John Thomson, and Crid Yu

Also in class:

Eric Boskin and Debra Hebert

Memorandum No. UCB/ERL M92/84

3 August 1992

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Preface

This is the third annual edition of the 290W report. This edition includes descriptions of projects completed during the Spring semester of 1992, in the context of the graduate course "Special Issues in Semiconductor Manufacturing". Ten students and two auditors have participated, and according to the course requirements, these students worked with me on their projects during the last six weeks of the semester.

The projects described in this report cover a rather wide range of semiconductor manufacturing applications. These include issues in statistical process control (Chapters 1, 2, 3, 4, 7, 10), experimental design (Chapters 3, 5, 6, 8, 9, 11) automated metrology (Ch. 4, 7, 9, 10), process modeling (Ch.7, 11) circuit design for manufacturability (Ch. 8), process design for manufacturability (Ch. 11). In the area of experimental design, we present work in classical DOE and in Taguchi methods (Ch. 3, 5, 6, 10) as well as in computer based experiments (Ch. 7, 11).

Each of the presented projects covers at least one novel aspect of semiconductor manufacturing. The first project discusses the creation of a novel multivariate exponentially weighted moving average scheme suitable both for closed loop control as well as SPC. The second deals with the application of time series models amenable to real-time control procedures. The next project presents a novel fractional-factorial based scheme for evolutionary operations. Next, the automation of in-situ photoresist monitoring scheme is discussed. The fifth presents a Taguchi experiment for phase shift mask design. The sixth projects deals with the improvement of thin, low temperature oxides. The seventh discusses issues in the distributions of particulate contamination and in-situ monitoring of airborne particles. The eighth deals with the creation of computer-based experiments that can be used to enhance the manufacturability of integrated circuits. The ninth project ties together in-line process measurements to process simulation tools for lithography control and diagnosis. The tenth project is a study of the scheduling and throughput problems that might be introduced by the metrology requirements of a run-to-run control system. Finally, the last project concentrates on a response surface methodology applied towards the design of a robust phase shift lithography procedure.

It is my hope that these reports will add to our understanding of semiconductor manufacturing. My thanks go to the 290W students and auditors whose work made this document possible. I am also grateful to the personnel and management of the Berkeley Microfabrication laboratory for their help with the experimental part of the projects presented here. Finally, I would like to acknowledge High Yield technologies (for letting us use one of their in-situ particle monitoring sensors), Gensym (for allowing us to use their G2 software package), and SC Technologies (for their help in acquiring and installing the SC Inspector monitoring system).

Costas J. Spanos

July, 1992

Table of Contents

A Multivariate Exponentially Weighted Moving Average Control Scheme

Crid Yu

The exponentially weighted moving average (EWMA) control scheme can be designed to detect large or small shifts in the process mean. A multivarite implementation of this scheme and its design guidelines will be presented. A direct vector extension of the univariate EWMA was chosen so that both the magnitude and the direction of the shift in mean can be ascertained. Alarm generation will depend on a single scalar Hotelling statistic. The in-control and out-of-control average run lengths (ARL) have been estimated by monte carlo simulation as a function of the control chart parameters. Multivariate EWMA design guidelines have been established for specified in-control and out-of-control ARL's. In general, it was found that for comparable multivariate CUSUM and EWMA charts, their ARL characteristics are almost the same. It is recommended that a Shewhart chart be used simultaneously with an EWMA chart so that large shifts in the process mean can also be detected quickly.

1.0 Introduction

The exponentially weighted moving average (EWMA) control scheme has been gaining popularity in the manufacturing industry because of certain unique properties. For the purpose of monitoring a process and generating control alarms, its properties are between that of the Shewhart and CUSUM control schemes. That is, an EWMA scheme can be designed so that it would be sensitive to large shifts in the process mean like a Shewhart chart, to small shifts like a CUSUM scheme, or some optimum point between these two extremes. Also, EWMA schemes have a filtering property as they tend to be less sensitive to sporadic shifts in the process control data stream.

In industrial applications, it is often desirable to be able to monitor several process variables at the same time. Furthermore, these variables may be mutually correlated. Any of the popular control schemes mentioned above can be used simultaneouly (in parallel) to monitor several process parameters at one time. However, because the process variables can be correlated, as they are usually on a semiconductor fabrication process, it is necessary to design and implement multivariate versions of these control schemes so that several process variables can be monitored in the form of vectors. In this type of implementation, it is more appropriate to generate control alarms based on a single statistic calculated from the data vectors. Discussions of multivariate CUSUM and Shewhart control schemes are abundant in the SPC literature. Their properties are well known and their design guidelines well established and well characterized.

A multivariate EWMA control scheme will be presented in this report. Its average run length (ARL) values will be presented and design guidlines based on these values will be established. Comparasons of this EWMA implementation will be made with its CUSUM and Shewhart counterparts.

2.0 Methodology

2.1 EWMA Formalism

In the univariate EWMA control scheme the statistic

$$Z_i = \lambda Y_i + (1-\lambda) Z_{i-1}, (0 < \lambda \leq 1) \tag{1}$$

is maintained and an alarm is generated if it is above the upper control limit (UCL) or is below the lower control limit (LCL). The sequentially recorded values, Y_i, can be sample values or averages from a sampling plan while λ is the weighing factor on past observations. If the Y_i's are identically, independently and normally distributed (IIND) with common variance, σ^2_Y, the variance of the control statistic is given by

$$\sigma^2(Z_i) = \left[\frac{\{1-(1-\lambda)^{2i}\}\lambda}{(2-\lambda)}\sigma_Y^2\right] \tag{2}$$

This value converges to the asymptotic value $\sigma_Z^2 = \{\lambda/(2-\lambda)\}\,\sigma_Y^2$ for large i's. Therefore, control limits are expressed by $K\sigma_Z$, where K is typically 3. The vector extension of this implementation is straightforeward. If we simply take the vector variables we have

$$Z_i = \lambda Y_i + (1-\lambda) Z_{i-1}, (0 < \lambda \leq 1) \tag{3}$$

where Y_i's now are the *difference between the vectors of process variables from the vector of targe values.* In most cases, this control scheme will be used to monitor shifts in the process variable means. Then the vector Y_i will contain the sample averages and the scheme will generate alarms if a significant shift in the *vector* is detected. Notice that in this scheme, the *direction* of the shift need not be specified in the chart. In fact, the statistic Z_i can be used to get an idea for the direction of the process shift in n-space. Alarm generation in this scheme will depend on a single statistic,

$$y = \sqrt{(\frac{\lambda}{2-\lambda}) Z_i' \Sigma^{-1} Z_i} \tag{4}$$

where Σ is the covariance matrix determined prior to the control scheme implementation from the process. This can be recognized as the Hotelling statistic. In this implementation, y is the length of the vector Z_i mapped onto a space where the variance in each vector dimension has been normalized to 1. Analogous to the univariate case, this length will be reduced by a factor of $\{\lambda/(2-\lambda)\}^{1/2}$ because of the asymptotic exponential weighing. Alarms will be generated if y exceeds a certain value h=K which needs to be determined depending on the requirements of the control chart. Graphically, this is equivalent to detecting whether Z_i exceeds the n-dimensional ellipsoid determined by h and Σ.

2.2 Average Run Length Calculations

The performance of process control schemes can be determined by its average run length properties. In general we design control charts with a certain in-control run length, or α risk, versus a certain out-of-control run length, or β risk. The optimal control limits are then chosen based on the average run length values.

2.2.1 Markov Chains

ARL values can be calculated analytically by various techniques. For the univariate case, the average run length distributions have been calculated using the Markov chain approach. This procedure involves dividing the interval between the upper and lower control limits into t=2m+1 subintervals of width 2d. With each sample taken, the control statistic makes a transition from state i to state j, not necessarily different, with certain probability ρ_{ij}. The process is considered in control whenever the control statistic is still in

a transient state and out of control if it exceeds either control limit and is in an absorbing state. The average run length and the moments of its distribution can then be calculated by well known methods [1]. The ARL for the continuous transition probilities can be calculated by taking the asymptotic value of ARL as t approaches infinity.

In the multivariate case then, the transition states need to span all n dimensions of the vector statistic. As the number of transition states = t^n, where n is the number of variables in the statistic, the transition probability matrix becomes unwieldly. However, Croiser [2] showed that for the multivariate CUSUM scheme, the Markov chain can be implemented not by all the considering the states of the vector statistic, but by its scalar Hotelling statistic y.

I extend this formalism to the case of the multivariate EWMA scheme. First, $E(\mathbf{Z}_i=\lambda\mathbf{Y}_i + (1-\lambda)\mathbf{Z}_{i-1}) = (1-\lambda)\mathbf{Z}_{i-1}$ for the on target case because we assume $\mathbf{Y}_i$ to have all elements normally distributed around zero and because for the purpose of calculating transitional probabilities, the value of $\mathbf{Z}_{i-1}$ is considered constant. Since $\mathrm{Var}(\mathbf{Z}_i)=\{\lambda/(2-\lambda)\}\Sigma$, the statistic y follows the chi-square distribution with noncentrality parameter

$$(1-\lambda)\,[(2-\lambda)/\lambda\,\mathbf{Z}_{i-1}'\,\Sigma^{-1}\,\mathbf{Z}_{i-1}]^{1/2} = (1-\lambda)[(2-\lambda)/\lambda]^{1/2}\,y_{i-1}. \tag{5}$$

Thus, the transition probability matrix greatly reduces in complexity as we only need to take into account the transitions of a single variable y_i. However, for the cases of the off target ARL, the statistics become complicated. In this case, $E(\mathbf{Y}_i) = \mathbf{u}$ and is no longer **0**. The noncentrality parameter then becomes

$$(1-\lambda)[(2-\lambda)/\lambda]^{1/2}[(\mathbf{Z}_{i-1}+\mathbf{u})'\,\Sigma^{-1}(\mathbf{Z}_{i-1}+\mathbf{u})] \tag{6}$$

and will depend on both **u** and $\mathbf{Z}_{i-1}$, not a single statistic. Thus the transition probability matrix still needs to span states in n space.

2.2.2 Monte Carlo Methods

Usually considered a tedious and time consuming procedure, Monte Carlo methods are becoming increasingly accessible as computing hardware becomes more powerful. For the purposes of ARL calculation, this alternative is appealing because the implementation of the simulation will be relatively simple. Furthermore, Croiser showed [2] that the scalar test statistic for the multivariate CUSUM chart depends only on the value of the noncentrality parameter defined to be:

$$d = [\mathbf{u}'\Sigma^{-1}\mathbf{u}]^{1/2} \tag{7}$$

where **u** is the shift in the vector of means. Note that this could not be applied to the generation of the transition probability matrix because the expectation value of $\mathbf{Z}_{i-1}$ was nonzero for the Markov chain calculations. However, for the purposes of simulation this fact greatly simplifies the simulation procedure.

Again, because the statistic y_i itself already normalizes the vector statistic to the variances in n space, the simulation can be implemented with no loss of generality if we pick an arbitrary covariance matrix Σ whose determinant is 1. For simplicity, the identity matrix is chosen. This reflects the fact that the matrix $\Sigma^{-1/2}$ has transformed the vector Z into a space where the covariances are 0 and the variances of each of the vector elements is one. Similarly the variance of each vector element can be set to one for the purpose of generating random vectors $\mathbf{Y}_i$. From the Croiser reference [2], this transformation does not alter the distribution of the statistic y_i because it depends only on d and is *invariant* under this transformation. Thus, the ARL distribution will only depend on the parameters λ, h (or L), n, and d.

2.3 ARL Simulations

The ARL can be simulated simply by implementing a C program that applies an EWMA control procedure with given λ, K, n, and d. An identity covariance matrix was used and the vectors of observations were generated by a random number generator following N(0,1) distributions. The simulated shift in the

mean vector was implemented by using a N(u,1) distribution for one of the vector elements. This makes d=u and we can determine the ARL versus d by simply changing u. We assume a steady state start of the control chart so that $Z_0 = 0$. The run is stopped and the run length recorded as soon as y_i exceeds h. For consistency I will refer to the value of L instead of h. In all the values I will report, 4 statistical moments (mean, variance, skew, kurtosis) were generated from 1000 iterations. The C program listing is presented in appendix A.

For comparason, some values for the univariate case have been generated and compared against published results calculated by numerical integration [3] and presented below:

d	0	0.5	1.0	1.5	2.0	2.5	3.0
simulated	119.49	23.61	7.5	4.1	2.95	2.31	1.89
calculated	124.18	23.28	7.52	4.18	2.92	2.29	1.91
% dif	3.81	1.4	0.27	1.9	1.02	0.67	1.05

FIGURE 1. Univariate EWMA Comparason of Simulated vs Calculated ARL for λ=0.25 L=2.5 (n=1)

In general they agree within a few percent of one another. The ARL estimates from simulation for d=0 are always going to be worse because in this case the run length distribution approximates that of a geometric distribution[4] (fig.2a) and thus its standard deviation is approximately equal to its mean. However, for cases with some shift in the mean, the estimates are much better. This can be seen in fig. 2b where the distribution peaks at the mean.

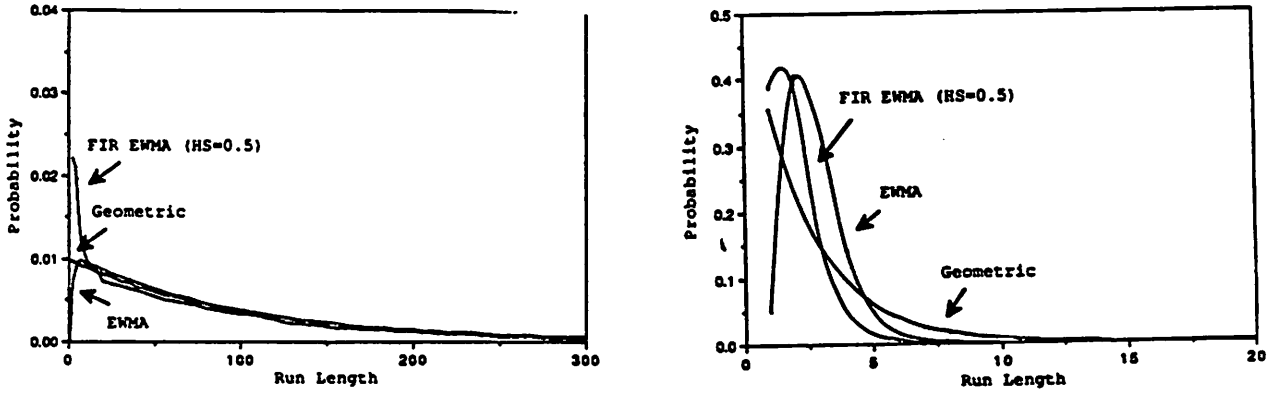


FIGURE 2. Multivariate EWMA: (a) Run length distribution for in control EWMA (b) Run length distribution for out of control EWMA.

3.0 Simulation Results

3.1 ARL Design Curves

Control charts are usually implemented so that they satisfy certain run length requirements. These are often a specified in-control run length and an out of control run length for a certain shift in the mean. Thus, for the multivariate EWMA chart it would be helpful to have design curves that allow us to determine the values of K and λ that would be necessary in order to achieve the desired in-control run lengths while sat-

isfying a run length requirements for a certain shift in the mean which is reflected in the noncentrality parameter d.

Figures 1-6 in appendix B provide some design guidelines for multivariate EWMA charts. ARL curves of λ=1.0, 0.25, 0.05 versus d are plotted. Figures 1-3 show in-control ARL's of 200 and 4-6 500. The values of K needed to achieve these constraints was determined iteratively converging to an optimum value for K given λ and the targel ARL. Curves were generated for two, five, and ten variable schemes.

For the same in control ARL (ARL(0)), charts with higher values of λ are more sensitive to large shifts in the process means, or larger values of d. This is because for this case the EWMA chart reduces to a Shewhart control chart. As λ decreases, more of the history of the process in incorporated into the vector statistic and the behavior of the EWMA approaches that of the CUSUM scheme; the chart becomes more sensitive to small shifts in the mean. From this a design methodology is apparent. Given ARL(0), a value of l is chosen so that the chart will be most sensitive to an anticipated value of d.

There has been much discussion concerning the performance of the EWMA compared to the CUSUM procedures. In figures 1-3 in the appendix, I've also plotted the ARL curves for 3 CUSUM schemes (from ref. 2) with the same in control ARL values. This reveals that the ARL characteristics of the CUSUM schemes are amost exactly the same as the the EWMA schemes with λ=0.05. It is conceivable that with certain values of λ the EWMA could be made to be more sensitive to this particular CUSUM chart. However, the CUSUM chart also has two design parameters at its disposal, h and k. It could also be adjusted so that certain optimal values of ARL's are obtained. Thus, by comparing ARL behavior alone it is not clear which of the two schemes has "superior" performance. This has also been the general conclusion for the univariate case in the literature.

One often cited disadvantage of EWMA charts is the inertia. Although we often refer to the CUSUM scheme as an extreme case of the EWMA scheme, this statement is not really true. This is because the alarm generation procedure for the CUSUM is inherently different from the EWMA scheme, in that the V-mask procedure used to generate alarms in the CUSUM procedure takes into account the slope of Q_i to i, where Q is the sum. Thus, as soon as the mean shifts, the CUSUM chart will start to respond even if the the sum was pulled far in the opposite direction. On the other hand, the EWMA scheme only considers the cumulative weighed sum. Thus, a more accurate statement would be that the EWMA scheme for λ approaching 0 is the CUSUM with k=0. Thus, the EWMA chart would not compare favorably to the CUSUM chart under worst case conditions where Z had been taken close to the control limit and a shift in the mean occurs in the opposite direction. It would take the statistic some time to recover to the point of Z=0. Thus, it is recommended that for EWMA charts with low values of λ a Shewhart control chart also be implemented to safeguard against this phenomenon. However, in general we observe that in order to have sensitivity to both large and small shifts in the mean, it might be necessary to implement EWMA charts with different values of λ at one time. The combined ARL value will then be approximately $(1/ARL_1 + 1/ARL_2)^{-1}$.

4.0 Conclusions

A multivariate implementation of the EWMA has been developed and its performance analyzed by its run-length characteristics. This approach is a direct vector extension of the univariate version. Notably, a vector statistic is maintained so that information about the direction of the shift in mean can be estimated while a scalar statistic was used to generate alarms. The run lengths were determined by simulation and, for the univariate case, was found to agree with calculated results in the literature to within 3 percent. ARL distribution moments were generated from 1000 trials for each data point. Design curves for multivariate EWMA was generated for on target ARL's of 200 and 500, and show the ARL dependencies on n, L, and λ.

In general, it was found that the ARL characteristics for multivariate CUSUM and comparable multivariate EWMA schemes are almost the same. However, because of the difference in the alarm generation schemes used in each technique, the EWMA may suffer from an inertia problem under worst case conditions. Thus, it is recommended that a Shewhart type chart be implemented to safeguard against this. It is also recommended that multiple EWMA charts be implemented with different values of λ so that both

large and small shifts in the process can be detected. An example of this is to implement a Shewhart chart with an EWMA chart

Some additional properties of the EWMA have not yet been explored for the multivariate case. The fast initial response feature could improve the sensitivity of the chart to certain shifts in the mean. The forecasting feature of the EWMA can also be implemented in multivariate form for use in adaptive control schemes.

5.0 References

[1] D. Brooks and D.A. Evans, "An Approach to the Probability distriobution of CUSUM Run Length," *Biometrika*, vol. 59, p. 539-549, 1972.

[2] R. B. Croiser, "Miltivariate Generalizations of Cumulative Sum Quality-Control Schemes", *Technometrics*, Vol. 30, no. 3, Aug. 1988.

[3] S. V. Crowder, "A Simple Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts," *Technometrics*, vol. 29, no. 4, 1987.

[4] J. M. Lucas and M. S. Saccucci, "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements", *Technometrics* , vol. 32, no. 1, Feb. 1990.

Appendix A

arl.c Sat May 9 13:41:21 1992

```
#include <stdio.h>
#include <string.h>
#include <math.h>

double rg()
{
    double x1 =(double)random()/ (1024*1024*2048-1);
    double x2 =(double)random()/ (1024*1024*2048-1);
    return (sqrt(-2.*log(1.0-x1)) * cos(2*M_PI*x2));
}

main()
{
float x,y[20],
  d,lam,l,arl,sum,h,
  qt[20],qto[20],z,rl1[100001],
  mean,sd, sum2,sum3,sum4,vl;
int i,j,k, n,it,rl;
FILE *fp1, *fp2,*fp3,*fp4;
printf("it, lam, l, h,n:");
scanf("%d %f %f %d",&it,&lam,&l,&h,&n);
printf("\n");

sum=0;
for (i=1;i<=it;i++)
{

  for (j=0;j<n;j++)
    {
      qto[j]=0;qt[j]=0;
    }

h=l*sqrt(lam/(2-lam));
rl=0;
z=0;
  while(z<h)
    {

    for(j=0;j<n;j++)
      {

        qto[j]=qt[j];
        qt[j]=(1-lam)*qto[j]+lam*rg();

      }

    qt[0]+=lam*d;
    rl++;
    z=0;
    for (j=0;j<n;j++)
      {
        z+=qt[j]*qt[j];
      }
      z=sqrt(z);

    }
  sum+=rl;
  rl1[i]=rl;

}
mean=sum/it;
for (sum3=0,sum4=0,sum2=0,i=1;i<=it;i++)
{
 vl=(rl1[i]-mean)*(rl1[i]-mean);
 sum2+=vl;
 sum3+=vl*(rl1[i]-mean);
 sum4+=vl*vl;
}
sd=sqrt(sum2/(it-1));
printf("arlm(l-4) %.2f %.2f %.2f %.2f %.2f\n",mean,sd,sum3/(pow(sd,3)*it),sum4/(pow(sd,4)*it)
}
```

Appendix B

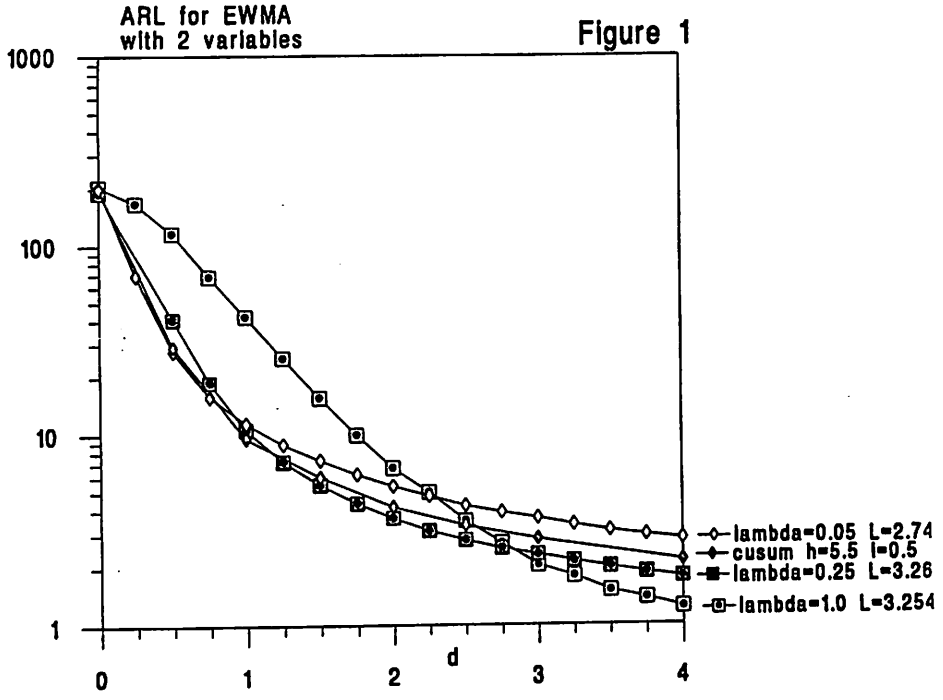


Figure 1

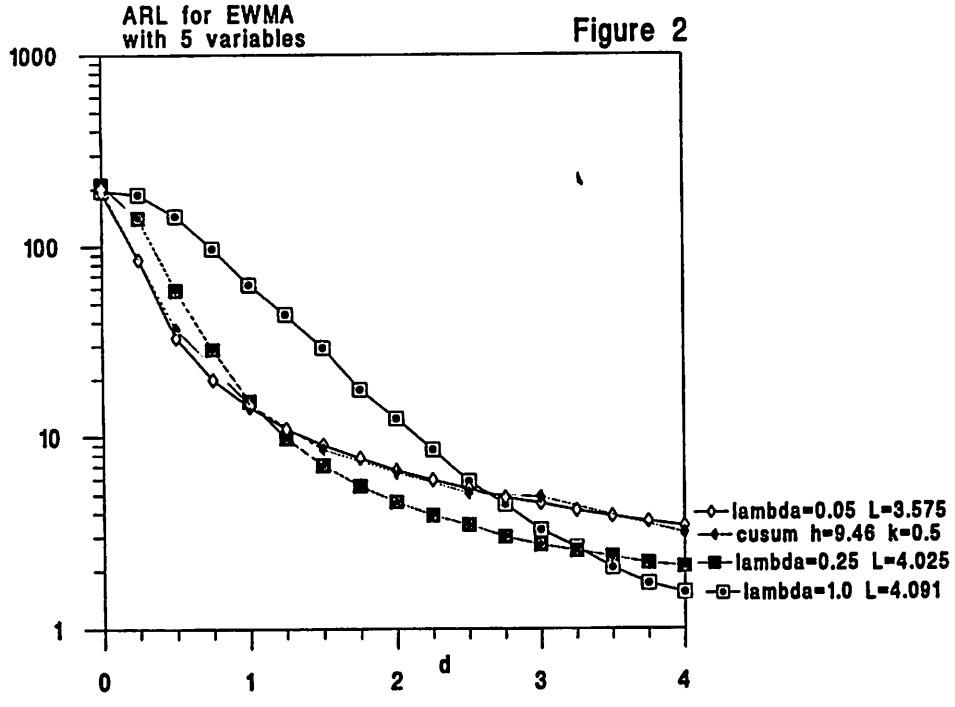


Figure 2

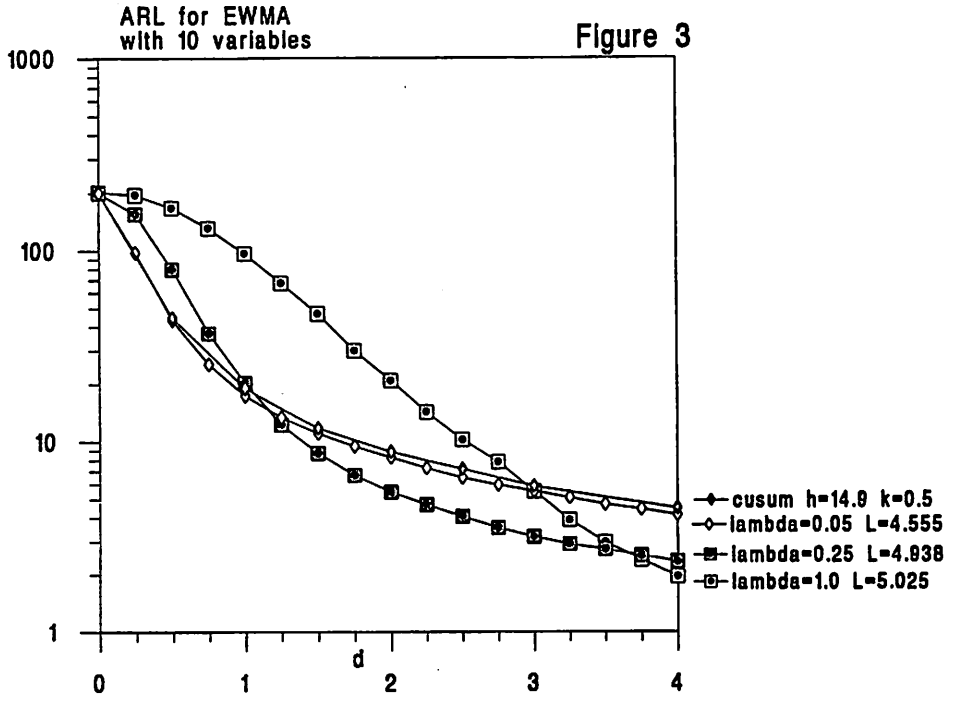


Figure 3

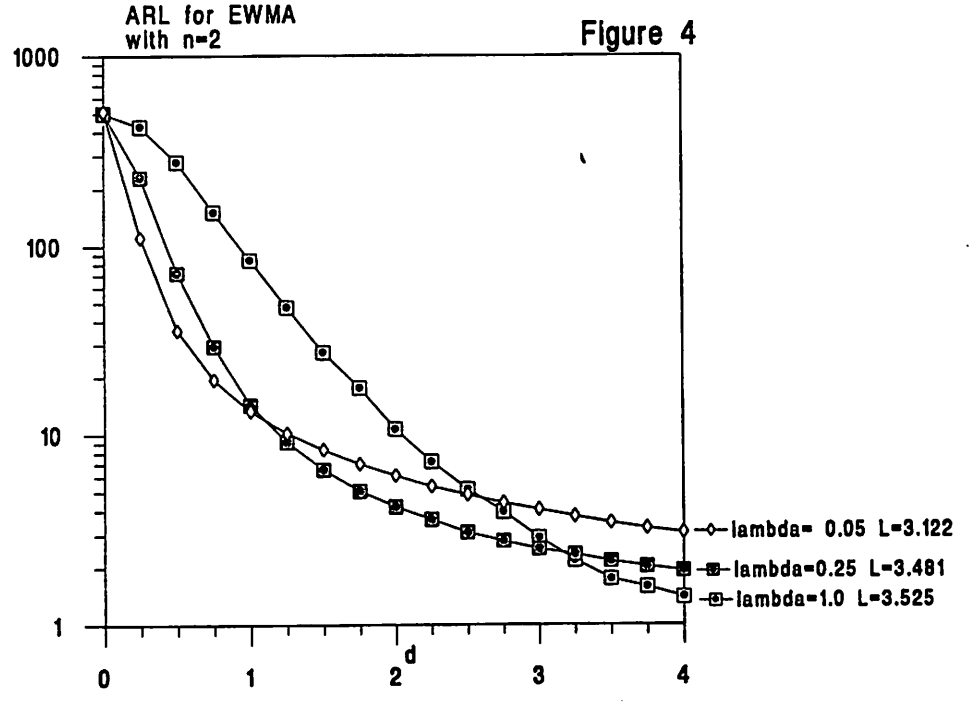


Figure 4

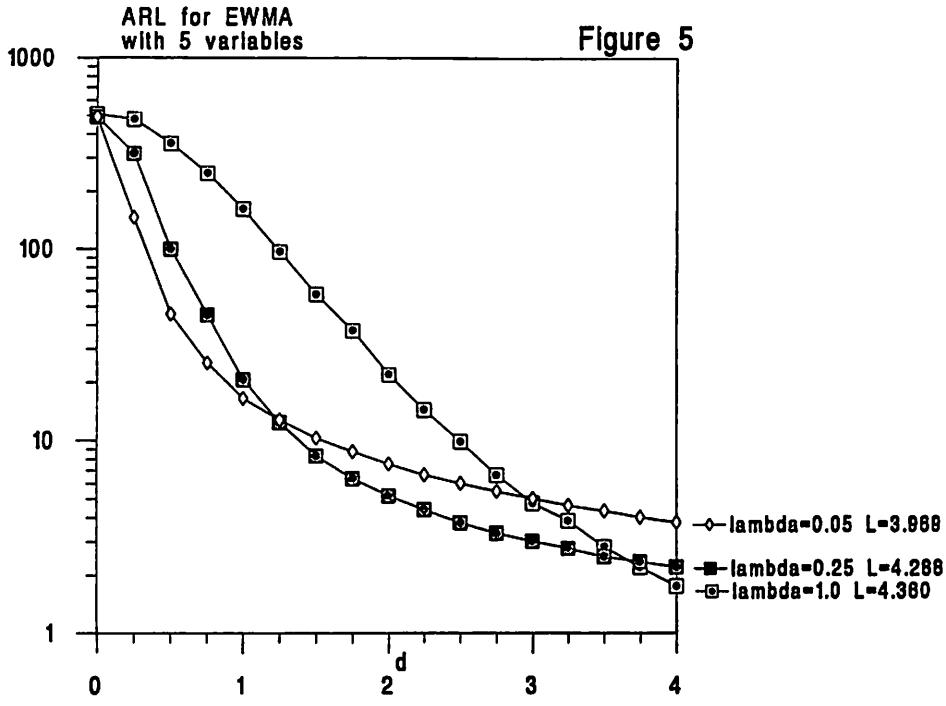


Figure 5

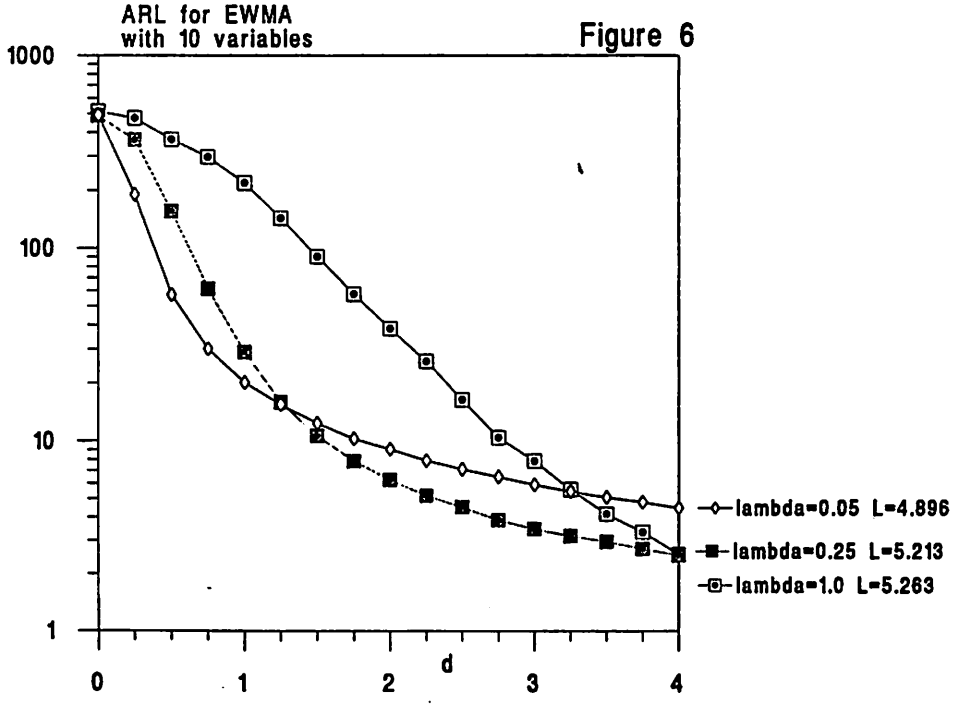


Figure 6

Advanced Empirical Equipment Modeling Using ARIMAX Time-Series Transfer Function Models

Hao-Cheng Liu

The goal of this project is to develop advanced empirical equipment models for semiconductor manufacturing using *Auto-Regressive Integrated Moving Average eXogenous (ARIMAX) Time Series Transfer Functions.* By using ARIMAX transfer functions, we are able to model not only the relationship between process outputs and inputs, but also time dependencies, if they exist. We have applied this modeling scheme to the GCA wafer stepper and compared our new model with that derived using simple regression. The results show that the ARIMAX transfer function is able to model time dependences that simple regression was unable to capture.

1.0 Introduction

In the semiconductor manufacturing industry the development of highly accurate equipment models is critical. These equipment models are used for predictions and simulations, and also in the implementation of feed-forward and feedback control, malfunction diagnosis, etc.

Traditional equipment models are derived empirically using simple regression analysis [1]. These equipment models are unsatisfactory because they fail to model any time dependencies in the equipment behavior. These time dependencies can exist as a result of changes in equipment inputs, process aging, equipment aging, maintenance events, etc. In order to properly capture these dependencies, we propose using *Auto-Regressive Integrated Moving Average eXogenous variable (ARIMAX) Time Series Transfer Functions.* As we will show below, the ARIMAX transfer functions combine the power of regression analysis with that of ARIMA time series modeling.

In section 2, we will first give a brief introduction to ARIMA time series functions and explain its application to equipment modeling. The details for the implementation of our study will be discussed in section 3, with the results presented in section 4. We will present our conclusions in section 5 and give some directions for future work in section 6.

2.0 Methodology

2.1 Background on ARIMA Time Series Transfer Functions

In many forecasting situations, information contained in timed observations of one variable will systematically influence the time dependent behavior of a dependent variables. The objective is to build a forecasting model that properly relates the time dependent behavior of several variables, thus capturing the dynamic characteristics of the system. ARIMAX time series transfer functions are useful in modeling such dynamic systems in which the output time series depends not only on its own past behavior, but also on the

input time series as well. An example of such a dynamic system simplified with one input and one output is shown in Figure 1.

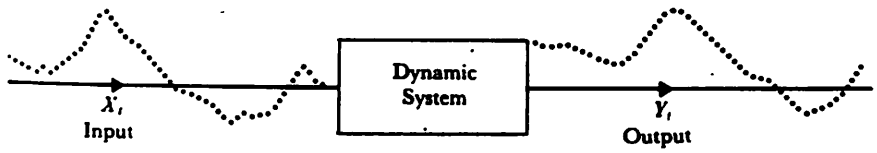


FIGURE 1. Input to and output from a dynamic system [2].

A simple transfer function model with one input X and one output Y can be written as follows:

$$Y_t - \delta_1 Y_{t-1} - \ldots - \delta_r Y_{t-r} = \omega_0 X_{t-b} - \omega_1 X_{t-b-1} - \ldots - \omega_l X_{t-b-l} + \varepsilon_t \tag{1}$$

The variable b is the *dead time*, the number of periods it takes before X_t starts influencing the dependent variable. If a change in X instantaneously affects Y, the dead time b is equal to zero. If we define

$$\omega(B) = \omega_0 - \omega_1 B - \ldots - \omega_l B^l \tag{2}$$

$$\delta(B) = 1 - \delta_1 B - \ldots - \delta_r B^r \tag{3}$$

where B is the backward shift operator defined as

$$BX_t = X_{t-1} \tag{4}$$

we can rewrite Equation 1 as follows:

$$\delta(B) Y_t = \omega(B) X_{t-b} + \varepsilon_t \tag{5}$$

Rearranging, we obtain

$$Y_t = \frac{\omega(B)}{\delta(B)} X_{t-b} + e_t \tag{6}$$

where

$$e_t = \frac{1}{\delta(B)} \varepsilon_t \tag{7}$$

We can rewrite Equation 6 as follows:

$$Y_t = \nu(B) X_{t-b} + e_t \tag{8}$$

where

$$\nu(B) = \frac{\omega(B)}{\delta(B)} = \nu_0 + \nu_1 B + \nu_2 B + \ldots \tag{9}$$

The weights $v_0, v_1, v_2, \ldots$ in Equation 9 are called the *impulse response weights* and a graph of these weights is called an *impulse response function.* An example of an impulse response function is given in Figure 2.

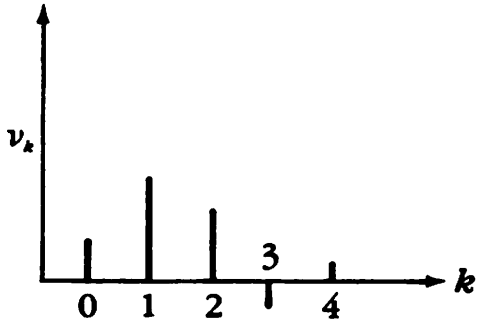


FIGURE 2. The impulse response function [7].

The error term e_t in Equations 6, 7, and 8 is not necessarily white noise. However, it might be possible to represent e_t with the following univariate ARIMA process [5], assuming that e_t is statistically independent of the explanatory variable X_t,

$$\nabla^d e_t = \frac{\theta(B)}{\phi(B)} a_t \tag{10}$$

or

$$\phi(B)\nabla^d e_t = \theta(B)\, a_t \tag{11}$$

with

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p \tag{12}$$

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q \tag{13}$$

where ∇^d is the consecutive difference operator used to induce stationarity in the series e_t,

$$\nabla^d X_t = (1-B)^d X_t \tag{14}$$

and a_t is assumed to be white noise. Finally, substituting Equation 10 into Equation 6, we obtain

$$\nabla^d Y_t = \frac{\omega(B)}{\delta(B)} \nabla^d X_{t-b} + \frac{\theta(B)}{\phi(B)} a_t \tag{15}$$

which is a general form of the ARIMA time series transfer function for a simple dynamic system with one input and one output. This transfer function model is represented in Figure 3. At the top of the figure, we have the transfer function structure determining the nature of the influence of the explanatory variable on

the dependent variable. In the lower part, we have the noise model representing a standard univariate ARIMA process. Finally, these two parts are put together to form the complete transfer function model [7].

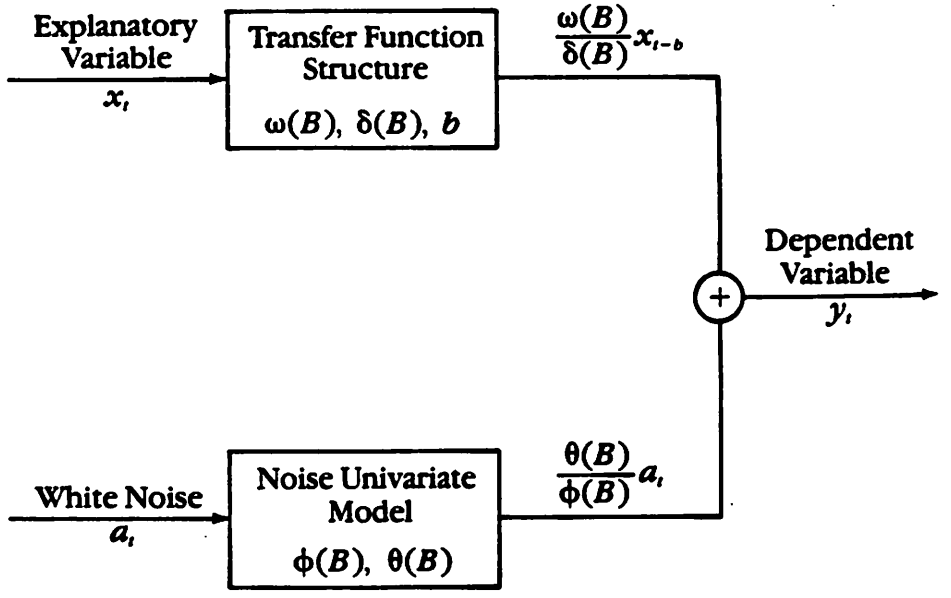


FIGURE 3. The transfer function model [7].

If seasonal patterns exist in our transfer function model, we may use the more general multiplicative seasonal transfer function model, which for a single explanatory variable is:

$$\nabla_S^D \nabla^d Y_t = \frac{\omega(B)\Omega(B^S)}{\delta(B)\nabla(B^S)} \nabla_S^D \nabla^d X_{t-b} + \frac{\theta(B)\Theta(B^S)}{\phi(B)\Phi(B^S)} a_t \tag{16}$$

with

$$\Omega(B^S) = \Omega_0 - \Omega_1 B^S - \Omega_2 B^{2S} - \ldots - \Omega_L B^{LS} \tag{17}$$

$$\nabla(B^S) = 1 - \nabla_1 B^S - \nabla_2 B^{2S} - \ldots - \nabla_R B^{RS} \tag{18}$$

$$\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \ldots - \Theta_Q B^{QS} \tag{19}$$

$$\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \ldots - \Phi_P B^{PS} \tag{20}$$

Notice that if there are also regular numerator parameters, we normalize the transfer function with $\omega_0 \neq 1$ and $\Omega_0 = 1$. If, however, there are no regular numerator parameters, we assume $\Omega_0 \neq 1$ [7].

Identification tools such as the *autocorrelation function*, the *partial autocorrelation function*, and the *cross-correlation function* are useful in determining the structure of the ARIMA transfer function model [5][7]. Once the structure of the model has been determined, we may use the Yule-Walker Equations [3, 4] and estimation methods such as least squares estimation and maximum likelihood estimation [5, 6] in order to determine the coefficients of the transfer function model.

It is clear that cross correlations can be a helpful tool for checking the dependencies between two time series. However, when a series, Y or X, is highly autocorrelated, the cross correlation function between the two time series can be difficult to interpret and it can even be misleading. It is quite possible that two time series which are not related at all show high spurious correlation if each one of the series is highly autocorrelated.

In order to obtain valuable identification information from cross correlations, it is recommended to first filter, or *prewhiten*, the data before calculating the cross correlations. This prewhitening of the data amounts to first obtaining the appropriate univariate models for each series involved, and then, at the second stage, cross correlating the (residual) white noise series [2,7].

2.2 Use of ARIMAX Transfer Functions in Equipment Modeling

For equipment modeling, we will use ARIMAX transfer functions with multiple inputs. These inputs may be measurement inputs or controlled inputs as shown in Figure 4. The outputs will be dependent on not only past values of the outputs themselves, if equipment aging exists, but also on current and past values of the inputs due to inherent process characteristics and process aging. This method of equipment modeling is particularly useful in feed-forward and feedback control, where the controlled inputs are highly correlated due to the fact that the control mechanism is constantly trying to bring a particular output measurement to target.

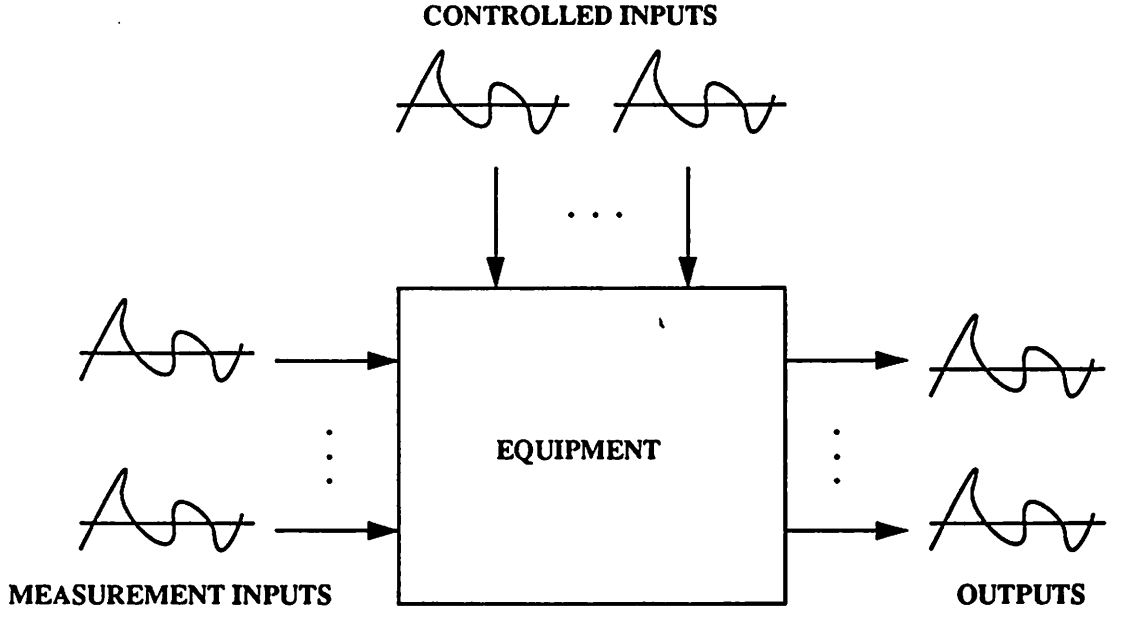


FIGURE 4. Inputs and outputs used in ARIMAX transfer function models.

Furthermore, seasonal transfer function models are useful in modeling seasonal effects that might arise due to processing of wafers in lots or batches. However, care must be taken in gathering the data used for empirically derived seasonal transfer function models.

3.0 Experimental Results from Lithographic Data

Experimental data from a previous experiment on feedforward control done on the GCA wafer stepper in the Berkeley Microfabrication Laboratory was used for our study. The experimental data is shown in Appendix 8.1.

This experimental data was fitted with a simple regression model, an ARIMA univariate time series model, and an ARIMAX transfer function model using the *SAS* Statistical Software Package [6]. The inputs of the models were the input thickness of the wafer, the input reflectance, and the normalized dose. The output of the models was the change in reflectance in the wafer. Appendix 8.2 contains the SAS code used for generating the above models.

For simplicity, no interaction terms were considered in our regression model. Furthermore, *non-seasonal* ARIMA and transfer function models were used. Although our data showed seasonal correlations, we felt that this seasonal pattern might be misleading due to the fact that our seven batches of wafers were not processed continuously in time. However, we are not discounting the possibility that seasonal patterns might exist in actual continuous processing.

4.0 Results

The resulting models are shown in the SAS output in Appendix 8.3. The forecasts and residuals for the regression model, the univariate ARIMA model, and the transfer function model are shown in Figures 5, 6, and 7, respectively.

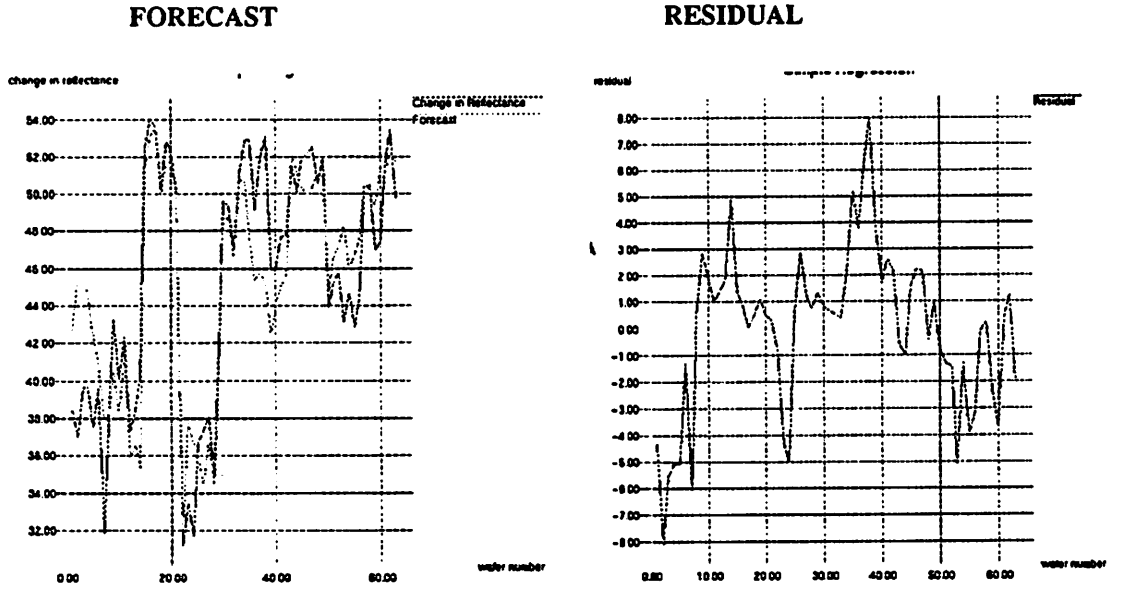


FIGURE 5. Forecasts and residuals for the simple regression model.

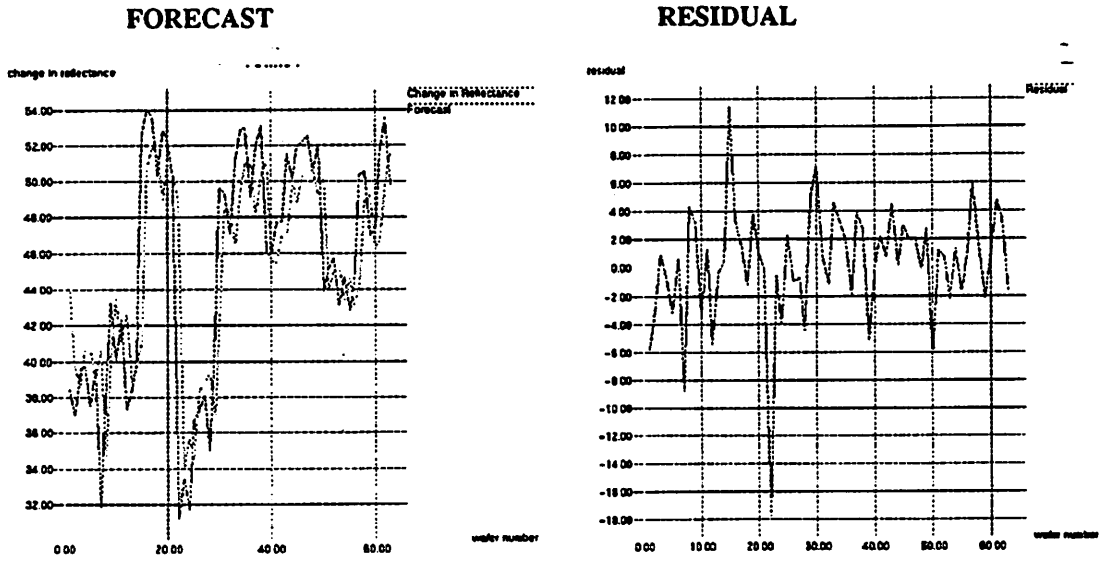


FIGURE 6. Forecasts and residuals for the univariate ARIMA time series model.

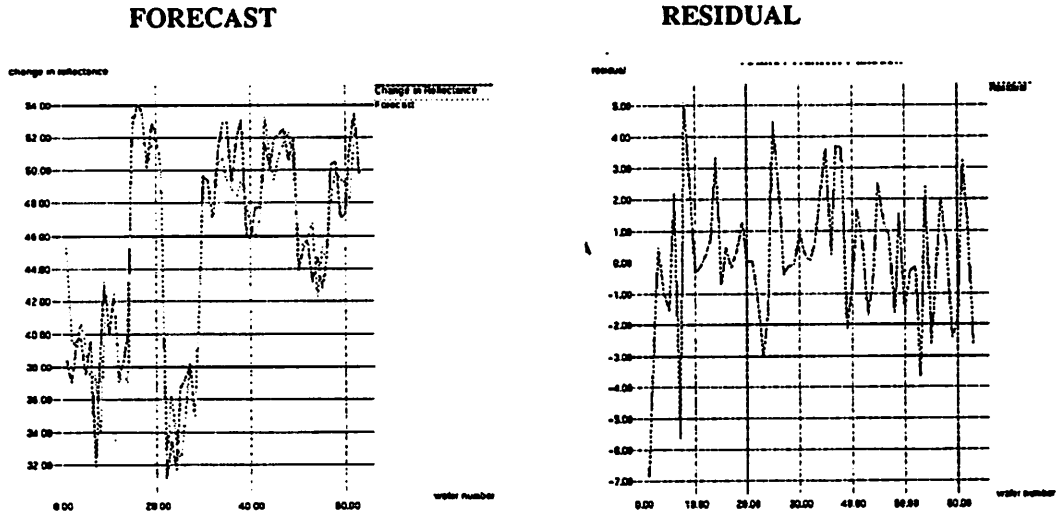


FIGURE 7. Forecasts and residuals for the ARIMAX transfer function model.

Looking at the results, we see that the regression model predicts the process changes well. Furthermore, the residuals are fairly tight with a variance of approximately 10% of the measured reflectance. However, we see that the residuals are highly autocorrelated. This can be inferred by looking at the auto-

correlation plot of the residuals for the regression model (See Appendix 8.3). Thus, we conclude that there are some time dependencies not modeled using our simple regression model.

A simple univariate ARIMA time series model was fitted to the output reflectance changes in the wafers. This model, which depends only on past output changes in reflectance, turned out to be an AR(1) model (See Appendix 8.3). As a result, we see that this model did not model the process changes well. This is because we did not consider the relationships between the inputs and the output. However, the residuals for this model do appear to be identically, independently, normally distributed (IIND), despite having a large variance of approximately 18% in measured reflectance.

Finally, we fitted our GCA data with an ARIMAX transfer function model. By looking at the forecasts, we see that this new model models captures the process changes as well as, if not better than the regression model. Furthermore, the residuals are IIND. And lastly, the variance of the residuals is a low 5% of the measured reflectance (See Appendix 8.3).

All in all, we see that by combining the power of a regression model with that of a time series model, we are able to built equipment models that explain the equipment process better than either a regression or time series model could do separately.

5.0 Conclusions

From our study, we see that although simple regression models model relationships between inputs and outputs well, they lack the capability of modeling time dependencies that may very well exist in semiconductor manufacturing processes. We therefore conclude that an ARIMAX transfer function model, which is a combination of a regression and a time series model, provides us with not only better insight into the process characteristics, but also allows us to model any time dependencies. Such time dependencies can arise due to changes in the inputs, process aging, equipment aging, maintenance events, etc.

6.0 Future Work

The use of ARIMAX time series transfer functions for equipment modeling should be pursued further in order to prove their validity and their superiority to traditional regression models. We will explore the effects of changing input parameters, such as temperature changes in wafer furnaces. Furthermore, we will look into the process decay during batch runs. These can include decreasing etch rates in etchers and decreasing deposition rates in deposition equipment.

We will also attempt to apply the use of these transfer function models to feed-forward and feedback control in order to see if we can obtain superior process characteristics. Furthermore, we will continue to look into the development of algorithms for the automatic generation of ARIMAX transfer function models. This will prove to be useful later in applications to feedback control.

7.0 Acknowledgment

I would like to thank Sovarong Leang for supplying me with the data he gathered for his feedforward control experiment.

8.0 Appendix

8.1 Experimental Data

TABLE 1. Experimental Data From the GCA Wafer Stepper

OBS (n)	REFL OUT (percent)	THICK IN (Å)	REFL IN (%)	EXP TIME (seconds)	STD EXP TIME (seconds)
1	79.96	12112	41.54	0.77	0.77
2	76.83	12078	39.83	0.77	0.77
3	79.27	12092	39.68	0.77	0.77
4	79.78	12081	40.02	0.77	0.77
5	78.92	12145	41.40	0.77	0.77
6	82.30	12162	42.68	0.77	0.77
7	77.54	12102	45.68	0.77	0.77
8	83.61	11929	44.74	0.70	0.77
9	86.51	11998	43.24	0.73	0.77
10	84.57	12014	44.48	0.71	0.77
11	85.10	11966	42.77	0.73	0.77
12	83.37	11952	46.04	0.68	0.77
13	84.19	11950	45.77	0.69	0.77
14	86.83	11944	46.64	0.68	0.77
15	89.49	11798	36.98	0.77	0.77
16	89.37	11790	35.35	0.77	0.77
17	88.56	11794	34.94	0.77	0.77
18	88.31	11814	37.95	0.77	0.77
19	89.09	11826	36.21	0.77	0.77
20	88.57	11814	36.38	0.77	0.77
21	87.88	11844	37.62	0.77	0.77
22	69.97	12281	38.77	0.41	0.77
23	68.57	12309	35.15	0.46	0.77
24	67.58	12296	35.85	0.45	0.77
25	72.91	12321	36.18	0.45	0.77
26	74.56	12366	37.25	0.45	0.77
27	73.72	12343	35.62	0.46	0.77
28	72.45	12283	37.48	0.43	0.77
29	86.63	12017	44.79	0.82	0.80
30	87.32	12024	37.71	0.82	0.80
31	87.34	11959	38.01	0.81	0.80
32	86.65	11994	39.61	0.81	0.80
33	87.36	11962	36.26	0.81	0.80
34	88.91	11990	36.03	0.81	0.80
35	91.80	11962	38.86	0.81	0.80
36	85.55	12466	36.41	0.80	0.80
37	88.37	12424	36.38	0.80	0.80
38	89.70	12465	36.63	0.80	0.80
39	84.38	12542	38.36	0.80	0.80
40	82.84	12526	37.01	0.80	0.80
41	84.18	12488	36.51	0.80	0.80
42	83.79	12492	36.03	0.80	0.80
43	86.68	12258	35.16	0.90	0.80
44	85.97	12242	35.87	0.89	0.80
45	87.98	12298	36.09	0.89	0.80
46	88.39	12357	36.20	0.90	0.80

TABLE 1. Experimental Data From the GCA Wafer Stepper

OBS (n)	REFL OUT (percent)	THICK IN (Å)	REFL IN (%)	EXP TIME (seconds)	STD EXP TIME (seconds)
47	88.43	12351	35.91	0.90	0.80
48	86.46	12252	35.88	0.89	0.80
49	87.93	12243	36.00	0.89	0.80
50	81.49	12386	37.61	0.80	0.80
51	81.29	12386	36.09	0.80	0.80
52	81.27	12385	35.55	0.80	0.80
53	78.15	12340	34.98	0.80	0.80
54	81.16	12392	36.43	0.80	0.80
55	78.74	12383	35.94	0.80	0.80
56	79.54	12390	35.22	0.80	0.80
57	88.10	12389	37.76	0.98	0.80
58	88.29	12410	37.74	0.98	0.80
59	85.37	12364	38.39	0.96	0.80
60	84.68	12406	37.26	0.99	0.80
61	88.75	12419	37.23	0.99	0.80
62	90.01	12390	36.52	0.99	0.80
63	86.66	12368	36.82	0.98	0.80

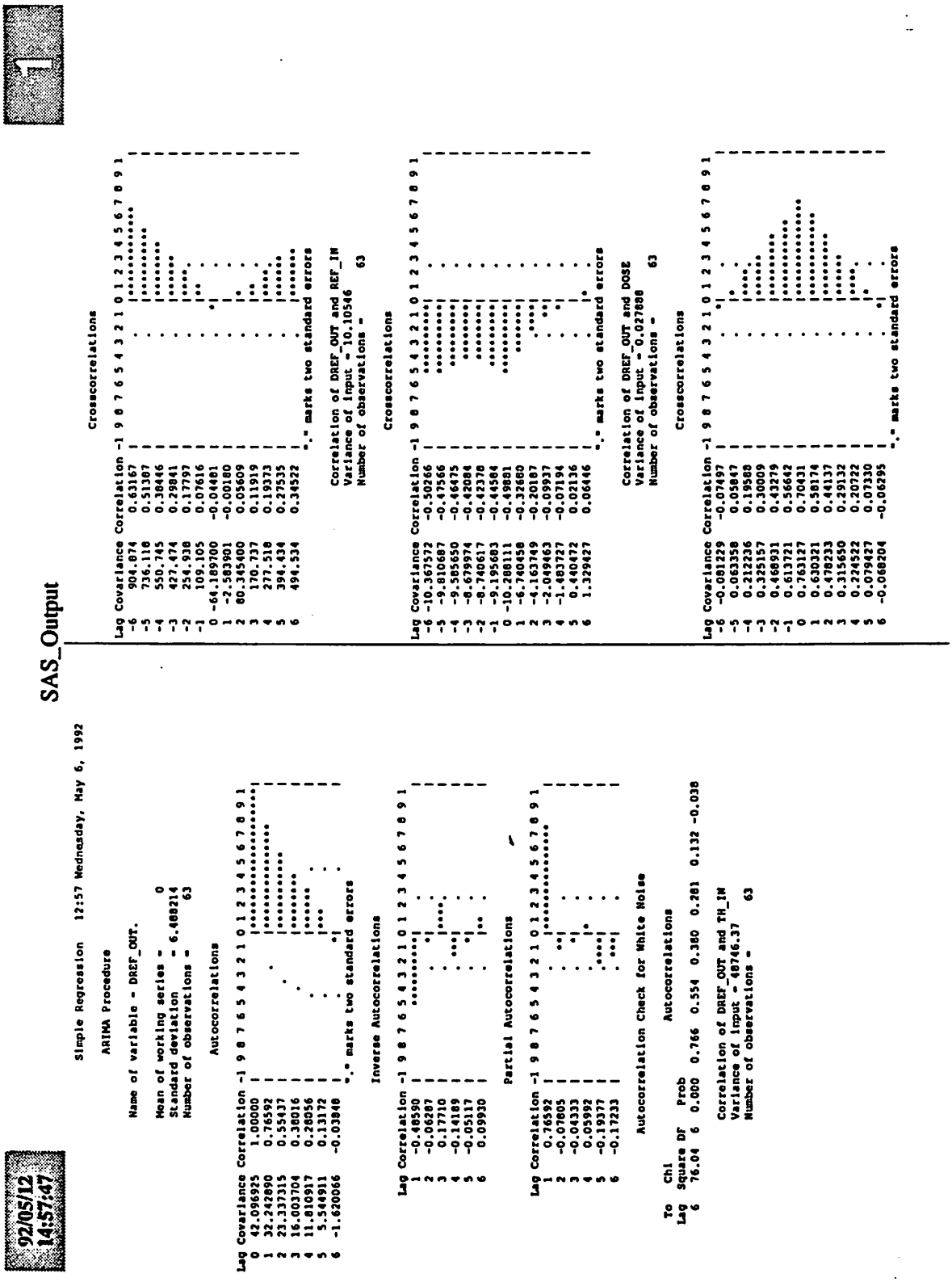
NOTE: The wafers were processed in batches of seven, and the batches were not processed consecutively in one day. The change in reflectance was used as the output of our models instead of the absolute reflectance. The normalized dose, which is obtained by dividing the exposure time by the standard exposure time, was used as an input instead of the absolute exposure time.

8.2 SAS Code

```
/* SPECIFY LIBRARY */
libname mydata 'c:\hao\arimax';
/* FIT REGRESSION EQUATION */
proc arima data=mydata.gcad2;
title 'Simple Regression';
identify var=dref_out crosscor=(th_in ref_in dose) nlag=6 center;
estimate input=(th_in ref_in dose) plot maxit=30;
forecast out=b1 back=0 lead=0 id=n printall;
run;
/* PLOT REGRESSION FORECASTS */
proc plot data=b1;
title 'Simple Regression';
plot dref_out*n='*' forecast*n='F' l95*n='L' u95*n='U' /overlay;
run;
proc gplot data=b1;
title 'Simple Regression';
plot dref_out*n forecast*n l95*n u95*n /overlay;
plot residual*n;
symbol1 i=join;
run;
/* FIT ARIMA MODEL */
proc arima data=mydata.gcad2;
title 'ARIMA';
identify var=dref_out nlag=6 center;
estimate p=1 plot maxit=30;
```

```
forecast out=b2 back=0 lead=0 id=n printall;
run;
/* PLOT ARIMA FORECASTS */
proc plot data=b2;
title 'ARIMA';
plot dref_out*n='*' forecast*n='F' l95*n='L' u95*n='U' /overlay;
run;
proc gplot data=b2;
title 'ARIMA';
plot dref_out*n forecast*n l95*n u95*n /overlay;
plot residual*n;
symbol1 i=join;
run;
/* FIT TRANSFER FUNCTION */
proc arima data=mydata.gcad2;
/* IDENTIFY OUTPUT REFLECTANCE WITHOUT PREWHITENTING INPUT */
title 'TRANSFER FUNCTION IDENTIFICATION WITHOUT PREWHITENED INPUTS';
identify var=dref_out crosscor=(th_in ref_in dose) nlag=6 center;
/* PREWHITEN INPUT THICKNESS */
title 'PREWHITENING OF INPUT THICKNESS';
identify var=th_in nlag=6 center;
estimate p=1 plot maxit=30;
/* PREWHITEN INPUT REFLECTANCE */
title 'PREWHITENING OF INPUT REFLECTANCE';
identify var=ref_in nlag=6 center;
estimate p=1 plot maxit=30;
/* PREWHITEN DOSE */
title 'PREWHITENING OF DOSE';
identify var=dose nlag=6 center;
estimate p=1 plot maxit=30;
/* IDENTIFY AND FIT TRANSFER FUNCTION MODEL */
title 'TRANSFER FUNCTION IDENTIFICATION WITH PREWHITENED INPUTS';
identify var=dref_out crosscor=(th_in ref_in dose) nlag=6 center;
title 'TRANSFER FUNCTION';
estimate p=1 input=(th_in ref_in dose) plot maxit=30;
forecast out=b3 back=0 lead=0 id=n printall;
run;
/* PLOT TRANSFER FUNCTION FORECASTS */
proc plot data=b3;
title 'TRANSFER FUNCTION';
plot dref_out*n='*' forecast*n='F' l95*n='L' u95*n='U' /overlay;
run;
proc gplot data=b3;
title 'TRANSFER FUNCTION';
plot dref_out*n forecast*n l95*n u95*n /overlay;
plot residual*n;
symbol1 i=join;
run;
```

8.3 SAS Output



92/05/12
14:57:47

SAS_Output

Conditional Least Squares Estimation

Parameter	Estimate	Approx. Std Error	T Ratio	Lag	Variable	Shift
MU	-2.558E-14	0.39654	-0.00	0	DREF_OUT	0
NUM1	-0.00963	0.0019709	-4.89	0	TH_IN	0
NUM2	-1.17565	0.13705	-8.58	0	REF_IN	0
NUM3	26.34688	2.38472	11.05	0	DOSE	0

Constant Estimate = -2.558E-14

Variance Estimate = 9.90661577
Std Error Estimate = 3.14747768
AIC = 327.125402*
SBC = 335.697941*
Number of Residuals= 63
* Does not include log determinant.

Correlations of the Estimates

Variable	Parameter	DREF_OUT MU	TH_IN NUM1	REF_IN NUM2	DOSE NUM3
DREF_OUT	MU	1.000	0.000	0.000	-0.000
TH_IN	NUM1	0.000	1.000	0.407	-0.037
REF_IN	NUM2	0.000	0.407	1.000	0.062
DOSE	NUM3	-0.000	-0.037	0.062	1.000

Autocorrelation Check of Residuals

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	65.21	6	0.000	0.687	0.470	0.399	0.290	0.182	0.049
12	67.79	12	0.000	-0.024	-0.022	-0.056	-0.066	-0.098	-0.121
18	83.83	18	0.000	-0.199	-0.268	-0.215	-0.089	-0.083	-0.132
24	85.04	24	0.000	-0.009	0.083	0.063	0.007	-0.013	-0.036

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	9.906616	1.00000	|********************|
1	6.803309	0.68674	|************** . |
2	4.658237	0.47021	|********* . |
3	3.955264	0.39925	|******** . |
4	2.869905	0.28970	|****** . |
5	1.805159	0.18222	|**** . |
6	0.480472	0.04850	|* . |

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.52259	|********** . |
2	0.18789	| . |**** . |
3	-0.19432	| .****| . |
4	0.08076	| . |** . |
5	-0.11101	| . **| . |
6	0.09677	| . |** . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.68674	| . |************** |
2	-0.00265	| . | . |
3	0.14630	| . |*** . |
4	-0.07200	| . *| . |
5	-0.03663	| . *| . |
6	-0.15200	| . ***| . |

Model for variable DREF_OUT

Data have been centered by subtracting the value 45.40444454.
Estimated Intercept = -2.558E-14

Input Number 1 is TH_IN.
Overall Regression Factor = -0.00963

Input Number 2 is REF_IN.
Overall Regression Factor = -1.17565

Input Number 3 is DOSE.
Overall Regression Factor = 26.34688

92/05/12
14:57:47

SAS_Output

ARIMA 12:57 Wednesday, May 6, 1992 32

Conditional Least Squares Estimation

Parameter	Estimate	Approx. Std Error	T Ratio	Lag
MU	-1.20706	2.15097	-0.56	0
AR1,1	0.78342	0.08167	9.59	1

Constant Estimate = -0.2614255

Variance Estimate = 17.6939293
Std Error Estimate = 4.20641526
AIC = 361.766782*
SBC = 366.053051*
Number of Residuals= 63
* Does not include log determinant.

Correlations of the Estimates

Parameter	MU	AR1,1
MU	1.000	-0.148
AR1,1	-0.148	1.000

Autocorrelation Check of Residuals

To Lag	Chi Square	DF	Prob	Autocorrelations						
6	2.14	5	0.830	0.044	0.011	-0.080	0.133	0.057	0.039
12	13.40	11	0.268	-0.336	-0.148	-0.095	0.093	0.011	0.014
18	18.59	17	0.353	-0.122	0.056	0.155	0.027	0.132	0.021
24	19.42	23	0.676	-0.007	0.014	0.029	0.047	-0.068	-0.018

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	17.693929	1.00000	|********************|
1	0.782335	0.04421	| . |* . |
2	0.196993	0.01113	| . | . |
3	-1.409584	-0.07966	| . **| . |
4	2.355499	0.13312	| . |*** . |
5	1.001465	0.05660	| . |* . |
6	0.691902	0.03910	| . |* . |

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.05163	| . *| . |
2	-0.01406	| . | . |
3	0.08772	| . |** . |
4	-0.13176	| . ***| . |
5	-0.04232	| . *| . |
6	-0.02376	| . | . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.04421	| . |* . |
2	0.00920	| . | . |
3	-0.08072	| . **| . |
4	0.14131	| . |*** . |
5	0.04605	| . |* . |
6	0.02451	| . |* . |

Model for variable DREF_OUT

Data have been centered by subtracting the value 45.40444454.
Estimated Mean = -1.2070624

Autoregressive Factors
Factor 1: 1 - 0.78342 B**(1)

SAS_Output

TRANSFER FUNCTION 12:57 Wednesday, May 6, 1992

Correlation of DREF_OUT and TH_IN
Both series have been prewhitened.
Variance of transformed series = 17.80631 and 10827.68
Number of observations = 63

Crosscorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-6	65.844437	0.14996	|***. |
-5	15.923295	0.03626	|* |
-4	-47.355429	-0.10785	.** |
-3	59.586646	0.13570	|*** |
-2	-14.487622	-0.03299	.* |
-1	18.547031	0.04224	|* |
0	-219.981	-0.50099	|********** |
1	-19.486705	-0.04438	.* |
2	-6.483640	-0.01477	| |
3	-10.712892	-0.02440	| |
4	-2.298185	-0.00523	| |
5	17.762491	0.04045	|* |
6	8.283247	0.01886	| |

"." marks two standard errors

Crosscorrelation Check Between Series

To Lag	Chi Square	DF	Prob	Crosscorrelations
5	16.09	6	0.013	-0.501 -0.044 -0.015 -0.024 -0.005 0.040

Both variables have been prewhitened by the following filter:

Prewhitening Filter

Autoregressive Factors
Factor 1: 1 - 0.89081 B**(1)

100

Correlation of DREF_OUT and REF_IN
Both series have been prewhitened.
Variance of transformed series = 17.24 and 4.601781
Number of observations = 63

Crosscorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-6	-1.226801	-0.13773	|. ***| . |
-5	-0.339159	-0.03808	|. *| . |
-4	-1.135443	-0.12748	|. ***| . |
-3	0.088392	0.00992	|. | . |
-2	-0.323098	-0.03627	|. *| . |
-1	-0.082655	-0.00928	|. | . |
0	-4.067944	-0.45671	|********* | . |
1	0.288206	0.03236	|. |* . |
2	0.030141	0.00338	|. | . |
3	1.012614	0.11369	|. |** . |
4	-1.165251	-0.13082	|. ***| . |
5	0.832024	0.09341	|. |** . |
6	-0.190749	-0.02142	|. | . |

"." marks two standard errors

Crosscorrelation Check Between Series

To Lag	Chi Square	DF	Prob	Crosscorrelations
5	15.65	6	0.016	-0.457 0.032 0.003 0.114 -0.131 0.093

Both variables have been prewhitened by the following filter:

Prewhitening Filter

Autoregressive Factors
Factor 1: 1 - 0.74675 B**(1)

Correlation of DREF_OUT and DOSE
Both series have been prewhitened.
Variance of transformed series = 17.49912 and 0.008207
Number of observations = 63

Crosscorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-6	0.029115	0.07683	|. |** . |
-5	-0.011615	-0.03065	|. *| . |
-4	0.037512	0.09899	|. |** . |
-3	-0.0050039	-0.01320	|. |* . |
-2	0.018287	0.04826	|. | . |
-1	0.00066022	0.00174	|. |********** |
0	0.244684	0.64567	|. |** . |
1	0.030778	0.08122	|. |* . |
2	0.020702	0.05463	|. |* . |
3	-0.052658	-0.13895	|. ***| . |
4	0.054520	0.14387	|. |*** . |
5	0.0038074	0.01005	|. | . |
6	0.016580	0.04375	|. |* . |

"." marks two standard errors

Crosscorrelation Check Between Series

To Lag	Chi Square	DF	Prob	Crosscorrelations
5	29.39	6	0.000	0.646 0.081 0.055 -0.139 0.144 0.010

Both variables have been prewhitened by the following filter:

Prewhitening Filter

Autoregressive Factors
Factor 1: 1 - 0.85436 B**(1)

Conditional Least Squares Estimation

Parameter	Estimate	Approx. Std Error	T Ratio	Lag	Variable	Shift
MU	-0.67938	0.97610	-0.70	0	DREF_OUT	0
AR1,1	0.72527	0.09515	7.62	1	DREF_OUT	0
NUM1	-0.0090516	0.0028709	-3.15	0	TH_IN	0
NUM2	-1.01473	0.13632	-7.44	0	REF_IN	0
NUM3	26.58300	3.34343	7.95	0	DOSE	0

Constant Estimate = -0.1866476

Variance Estimate = 5.14139702
Std Error Estimate = 2.26746489
AIC = 286.728142*
SBC = 297.443815*
Number of Residuals= 63
* Does not include log determinant.

Correlations of the Estimates

Variable	Parameter	DREF_OUT MU	DREF_OUT AR1,1	TH_IN NUM1	REF_IN NUM2	DOSE NUM3
DREF_OUT	MU	1.000	-0.214	0.024	-0.083	-0.011
DREF_OUT	AR1,1	-0.214	1.000	-0.151	0.160	-0.096
TH_IN	NUM1	0.024	-0.151	1.000	0.095	0.390
REF_IN	NUM2	-0.083	0.160	0.095	1.000	-0.022
DOSE	NUM3	-0.011	-0.096	0.390	-0.022	1.000

Autocorrelation Check of Residuals

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	3.36	5	0.645	-0.011	-0.114	0.103	0.041	0.143	-0.050
12	4.84	11	0.939	-0.087	0.040	-0.012	0.044	0.062	0.065
18	16.14	17	0.380	-0.060	-0.187	-0.122	0.176	0.086	-0.245
24	21.47	23	0.552	0.053	0.155	0.079	-0.009	-0.034	0.018

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	5.141397	1.00000	|**********|
1	-0.054587	-0.01062	| . | . |
2	-0.585351	-0.11385	| . **| . |
3	0.531149	0.10331	| . |** . |
4	0.212461	0.04132	| . |* . |
5	0.737523	0.14345	| . |*** . |
6	-0.259312	-0.05044	| . *| . |

"." marks two standard errors

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.01771	| . | . |
2	0.13401	| . |*** . |
3	-0.14609	| . ***| . |
4	-0.01609	| . | . |
5	-0.16024	| . ***| . |
6	0.05071	| . |* . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.01062	| . | . |
2	-0.11398	| . **| . |
3	0.10209	| . |** . |
4	0.03030	| . |* . |
5	0.17098	| . |*** . |
6	-0.05398	| . *| . |

Crosscorrelation Check of Residuals with Input TH_IN

To Lag	Chi Square	DF	Prob	Crosscorrelations					
5	0.32	6	0.999	0.023	-0.053	-0.010	0.028	0.021	-0.021
11	1.87	12	1.000	-0.138	-0.026	-0.015	0.007	0.047	0.050
17	11.10	18	0.890	0.062	0.050	-0.142	0.141	0.119	-0.293
23	12.36	24	0.976	-0.110	-0.057	0.014	-0.009	-0.004	-0.065

Crosscorrelation Check of Residuals with Input REF_IN

To Lag	Chi Square	DF	Prob	Crosscorrelations					
5	0.23	6	1.000	-0.000	-0.044	0.014	0.025	-0.017	-0.025
11	7.74	12	0.805	0.050	0.163	0.127	0.088	-0.200	-0.162
17	10.63	18	0.909	0.078	0.139	-0.081	-0.099	0.025	0.059
23	12.68	24	0.971	-0.037	-0.045	0.084	-0.074	-0.080	0.100

Crosscorrelation Check of Residuals with Input DOSE

To Lag	Chi Square	DF	Prob	Crosscorrelations					
5	2.85	6	0.828	0.012	0.102	0.004	-0.155	-0.038	0.096
11	4.78	12	0.965	0.020	-0.080	0.024	0.050	-0.128	-0.066
17	8.43	18	0.972	-0.086	-0.164	-0.073	-0.105	-0.061	0.059
23	9.80	24	0.995	0.041	-0.095	0.016	-0.029	0.020	-0.098

Model for variable DREF_OUT

Data have been centered by subtracting the value 45.40444454.
Estimated Intercept = -0.6793797

Autoregressive Factors
Factor 1: 1 - 0.72527 B**(1)

Input Number 1 is TH_IN.
Overall Regression Factor = -0.00905

Input Number 2 is REF_IN.
Overall Regression Factor = -1.01473

Input Number 3 is DOSE.
Overall Regression Factor = 26.583

9.0 References

[1] Box, George E. P., Hunter, William G., and Hunter, J. Stuart, *Statistics for Experiments*, John Wiley & Sons, pp. 453-628, 19

[2] Box, George E.P. and Jenkins, Gwilym M., *Time Series Analysis: Forecasting and Control*, Holden-Day, pp. 337-491, 1976.

[3] Chow, Joseph C., "On Estimating the Orders of an Autoregressive Moving-Average Process with Uncertain Observations", *IEEE Transactions on Automatic Control*, pp. 707-709, October 1972.

[4] Kay, Steven M. and Marple, Stanley Lawrence, Jr., "Spectrum Analysis - A Modern Perspective", *Proceedings of the IEEE*, Vol. 69, No. 11, pp. 1380-1419, November 1981

[5] Pankratz, Alan, *Forecasting with Univariate Box-Jenkins Models*, John Wiley & Sons, 1983.

[6] SAS Institute Inc., *SAS/ETS User's Guide, Version 5 Edition*, Cary, North Carolina: SAS Institute Inc., pp. 127-181, 1984.

[7] Vandaele, Walter, *Applied Time Series and Box-Jenkins Models*, Academic Press, Inc., pp. 257-347, 1983.

Evolutionary Operation with Fractional Factorials

John Thomson

An evolutionary operation software package for use in any manufacturing environment has been written. Simulations have been run using equipment models of the Eaton photoresist spin-coat & bake station. The effects of changing various parameters of the simulations have been observed.

1.0 Introduction

Evolutionary operation (EVOP) is a popular process control technique that is useful in optimizing equipment performance during production. Factorial experiments centered around the current operating point are constructed and the operating point may be adjusted if a favorable effect on the output is likely. When running the experiment, only small deviations may be introduced to the inputs in order for the process capability to remain acceptable. However, if the deviations are made too small, then the effects of the input variables will be invisible due to the noise of the process.

EVOP attempts to position the operating point at its optimal value even for noisy, dynamic equipment. Unlike traditional off-line experimental designs, EVOP is applied on a sequential run-by-run basis during actual production. The goal of this project is to design and implement generic EVOP algorithms for use in any manufacturing environment.

2.0 Methodology

2.1 Design of Experiments during EVOP

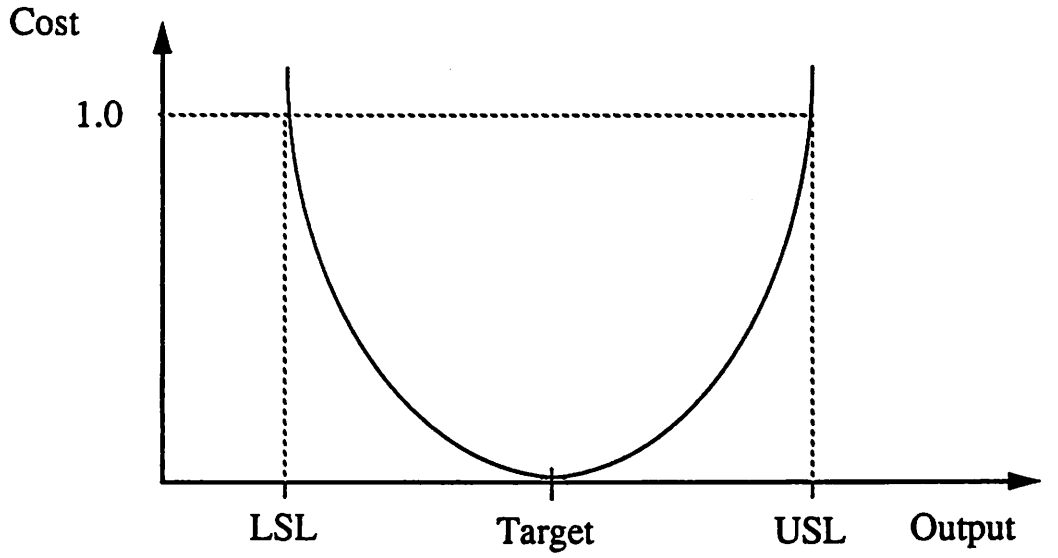
Any EVOP approach uses the common idea of a structured factorial experiment. Still, substantial flexibility exists in the design of the experiment and the actions taken as a result. When performing EVOP, a decision must be made regarding the magnitude of the deviations introduced to the inputs. If the ideal operating point is far from the current operating point, then large deviations are useful to shift the inputs as quickly as possible. This is especially true if the current operating point is at a relatively insensitive location of the response surface or if appreciable noise exists, since small input deviations will have negligible effect. At the other extreme, if the current operating point is at its optimal location and the process is sensitive to the inputs, then the deviations introduced must be small in order for an acceptable capability to be maintained.

Fractional factorial designs should be used, especially if the number of inputs is larger than 3 or 4. By using fractional factorials, the important information is often deduced using fewer runs compared to EVOP using full factorial designs. The danger is that high order fractional designs run the risk of excessive confounding of effects which may lead to incorrect conclusions. As a minimum, the resolution of the design must be at least III so that first order effects do not confound with each other. It is important to realize that resolution III designs are still not immune to first order effects confounding with second order effects. If

second order effects are considered to be significant, based on operator experience or theoretical foundations, then a higher resolution design should be utilized.

EVOP assumes that a single performance measure is being minimized or maximized. Many applications strive to bring an output close to a certain target, a task that is usually accomplished with the help of a cost function. A popular quadratic cost function for a single output is shown in figure 1. For systems with multiple outputs, the total cost would be the sum of the cost associated with each output.

Figure 1: Quadratic cost function for a single output



2.2 Estimation of Effects

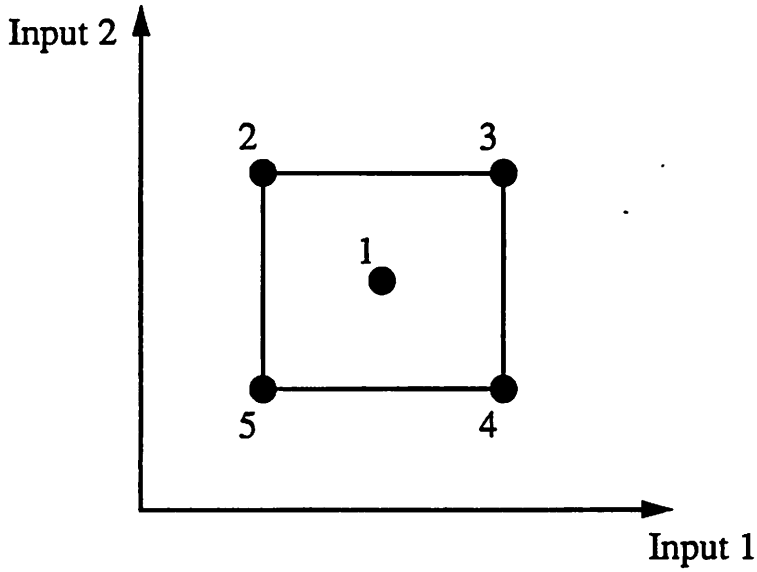
Figure 2 shows a 2^2 full factorial design which would be used for a system with only two inputs. After the equipment is run at each of the five locations, a *cycle* is said to have been completed, and a decision is made whether or not to change the current operating point. To make this decision, the *effects* of the inputs are calculated. A first order effect for an input variable is defined to be the average change in the output when going from the input's low value to its high value. For example, referring to figure 2, the effect of input 1 would be the average of the output at points 3 and 4, minus the average of the output at points 2 and 5. A similar calculation is required for the effect of input 2. (The center point is not used for the effect estimation. It is only used to determine if a minimum or maximum has been reached.)

With R runs per cycle, excluding the center point, the error of the first order effects after n cycles at a particular factorial location is:

$$\sigma_{effect} = \frac{2\sigma_{exp}}{\sqrt{Rn}} \tag{1}$$

The 95% confidence interval for the estimated effects is $\pm 2\sigma_{effect}$. If the estimate of the effect exceeds this interval, then the effect is considered to be significant.

Figure 2: 2^2 Factorial design with center point



The change in mean effect is defined to be the difference in the average response at the factorial locations of the experiment minus the average response at the center point. For the 2-input case shown in figure 2, it is:

$$CIMeffect = \frac{1}{5}(\bar{y}_2+\bar{y}_3+\bar{y}_4+\bar{y}_5-4\bar{y}_1) \tag{2}$$

The standard deviation of the estimated CIM effect is:

$$\sigma_{CIMeffect} = \sqrt{\frac{R}{(R+1)n}}\sigma_{exp} \tag{3}$$

The CIM effect is used to determine if a minimum or maximum has been reached in the response surface. The significance of the CIM effect is established with the help of the 95% confidence interval which is $\pm 2\sigma_{CIMeffect}$. A significantly positive CIM effect indicates that a minimum may have been reached. A significantly negative CIM effect indicates that a maximum may have been reached. These conclusions are only valid if no first order effects are significant.

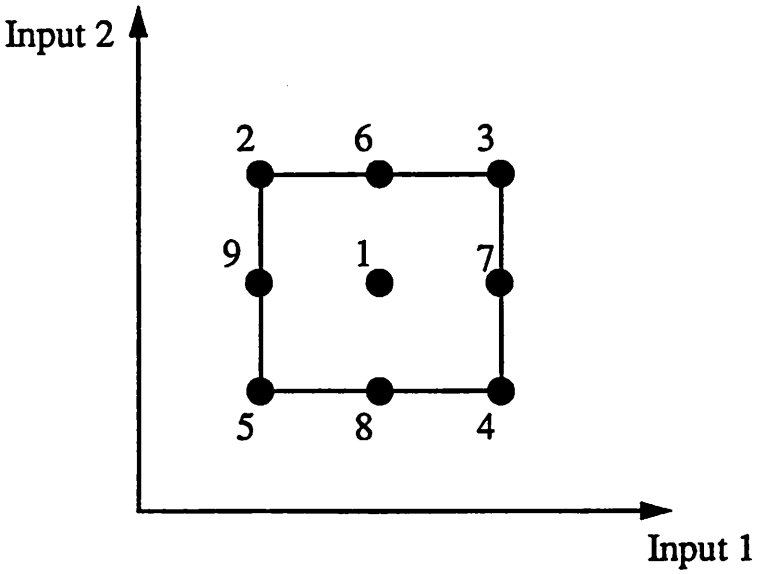
The numbers assigned to the points in figure 2 are not related to the actual order in which the experiment is run. The actual order is randomized in order to keep time effects unconfounded with at least first order effects for any given cycle.

2.3 Shifting the Experiment

If a first order effect is significant, then the position of the factorial is moved for the next cycle. The location of the new factorial is dependent upon which effects are significant, whether the effects are positive or negative, and whether the output is being minimized or maximized. For example, if an effect is significantly positive and the output is being minimized, then the input will be decreased for the next cycle. Each first order effect is examined independently. Note that interaction effects need not be calculated even if they exist, since the gradient of the response surface can be determined from first order effects alone.

After any given cycle, the center point of the next cycle will be shifted positively, negatively or not at all, for each input. Thus, for k inputs, there are 3^k possible locations for the next cycle as shown in figure 3. (Locations 1 through 5 are the settings used for the previous cycle.) It is quite possible that no effects will be significant, resulting in a cycle repeated at the same location.

Figure 3: Possible locations for a center point for the next cycle



Some implementations of EVOP unnecessarily restrict the possible new positions of the center point to be one of the locations of the previous experiment: positions 1 through 5. Doing this results in a reduced number of possible new locations, but actually makes the move decisions more complicated. Consider, for example, that the output is to be minimized, the effect of input 1 is negative and the effect of input 2 is negligible. We are motivated to increase input 1, implying that positions 3 and 4 are candidates for the next center point. We must make the choice between 3 and 4 arbitrarily, or wait for the effect of input 2 or the interaction effect to become significant in order to make an intelligent decision. However, waiting for other effects to become significant will slow the response time of EVOP and therefore is not recommended.

2.4 Estimation of Experimental Error

Testing the significance of the effects requires an estimate of the experimental error. This can be acquired by using points which are replicated during a repeated cycle. Any time a cycle is repeated, a new estimate of the experimental error is obtained by taking the difference of the last average at a particular location and the new value. It can be shown that this difference follows the distribution $N(\mu, \sigma^2_{exp}n/(n-1))$ where n is the number of cycles run and σ_{exp} is the experimental error. μ will be 0 if the process has not shifted. In general, σ_{exp} is unknown and must be estimated from the differences. There is a total of $2^{inputs}+1$ differences which are samples from the above distribution. The sample variance can be used as an unbiased estimator of $\sigma^2_{exp}* n/(n-1)$, from which an estimator of σ^2_{exp} is easily obtained. Simply taking the square root of this estimator gives a biased estimator of σ_{exp}. An unbiased estimator may be obtained by dividing by c_4 where:

$$c_4 = \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \tag{4}$$

$G(k) = (k-1)!$, for integer k. $G(1/2) = \sqrt{\pi}$

When repeated cycles are needed using a fractional arrangement, each repetition should use the same fraction. Different fractions could be used for each cycle, but then new estimates of σ could not be made.

It is possible to use the range of the differences to estimate the experimental error. This is computationally simpler, but the computing power evident in modern-day computer integrated manufacturing frameworks makes this difference unnoticeable. Further, as the number of runs per cycle increases, the relative efficiency of the range estimator diminishes, so it should only be used in problems with small dimensionality.

Once an estimate of the experimental error is obtained from a repeated cycle, it is combined with the previous estimate using an exponentially weighted moving average to form the new estimate to be used. This weighted average is used to signify that recent estimates are more important than older ones.

2.5 Other Issues

In general, if any effect becomes significant after only one cycle at a particular location, then the experiment is moved for the next cycle and no updated estimate of the noise is obtained. This situation is called a *quick move*. There are two instances when a quick move is not allowed. First, since the experimental error is estimated only through replication, at least two cycles will always be run when EVOP starts in order for the initial noise estimate to be derived.

The second exception is if too many consecutive quick moves have been performed. If this were to occur, then the estimate of the error would not be updated and instead would be based only on relatively old data. Thus, a cycle is inserted to update the estimate. In a dynamic system where the noise level is constantly changing, it is crucial to maintain an accurate estimate of the noise at all times. But even in static systems where the noise level of the outputs is constant, the sensitivity of the cost function to noise is variable, since it will be dependent on the distance of the outputs to their targets. The addition of noise to outputs which are close to their targets will have a small impact on the cost since the cost function is in its flat region. The addition of noise to outputs which are distant from their targets will have a large impact on the cost since the cost function is in a steep region. Thus, even for static systems, updated estimates of the error are necessary.

To ensure that a current estimate of the error is used, a limit has been set on the maximum number of consecutive quick moves allowed. This heuristic was developed to guard against the situation where the current estimate of σ is smaller than the true σ. Without the heuristic, numerous consecutive quick moves may occur without having a chance to update σ, causing moves may be made as a result of noise only. Having an overestimate of σ is not a problem since it will be more difficult to shift the experiment. Consequently, cycles will be repeated and updated estimates of σ will be made.

3.0 Implementation

The algorithms described above have been implemented in C++ and have been combined with equipment models developed by the Berkeley Computer Aided Manufacturing group. Various parameters of the experiment are specified by the user. These include:

- énumber of inputs and outputs
- degree of fractionation
- range of the inputs
- the starting center point for the inputs
- specifications for the outputs
- step size

The step size specifies the distance between the center point and the factorial points along each input direction, as a fraction of the range of each input. The other parameters that may be modified are the maximum number of consecutive quick moves and the forgetting factor used in the exponentially weighted average calculation for the estimate of the noise.

Simulated optimization runs were completed on the Eaton photoresist spin and bake station, but any equipment can be simulated with trivial modification to the code. In addition, small changes are required in order to run the experiments on the actual equipment instead of using a simulator. The simulations determine the number of runs required to find the optimum as a function of the step size and whether or not a full factorial is used. Once the optimum has been found, the effects of continuously changing the recipe on the cost function have been analyzed.

Generators for the fractional factorials have been taken from page 410 in "Statistics for Experimenters" by Box, Hunter and Hunter.

4.0 Simulated Optimization of the Spin-Coat & Bake Procedure

A few parameters were set somewhat arbitrarily before the simulation began. Specification limits were set to be 12300 - 12500 Å for the output thickness and 37.5 - 42.5% for the output reflectance. Standard errors of 70 Å and 1.5% were added to the thickness and reflectance respectively. The starting center point for the factorial experiment was 5200 rpm for spin speed, 30 seconds for spin time, 115°C for bake temperature and 90 seconds for bake time. Further, the maximum number of consecutive quick moves was 5 and the forgetting factor was 0.5.

The result of using full and half factorials with various step sizes is shown in Table 1.

TABLE 1.

Type of Factorial	Step Size	Runs to reach optimum	Average cost after optimum is reached
Full	0.01	1207	0.95
	0.02	408	1.32
	0.05	136	3.30
	0.10	51	11.77

TABLE 1.

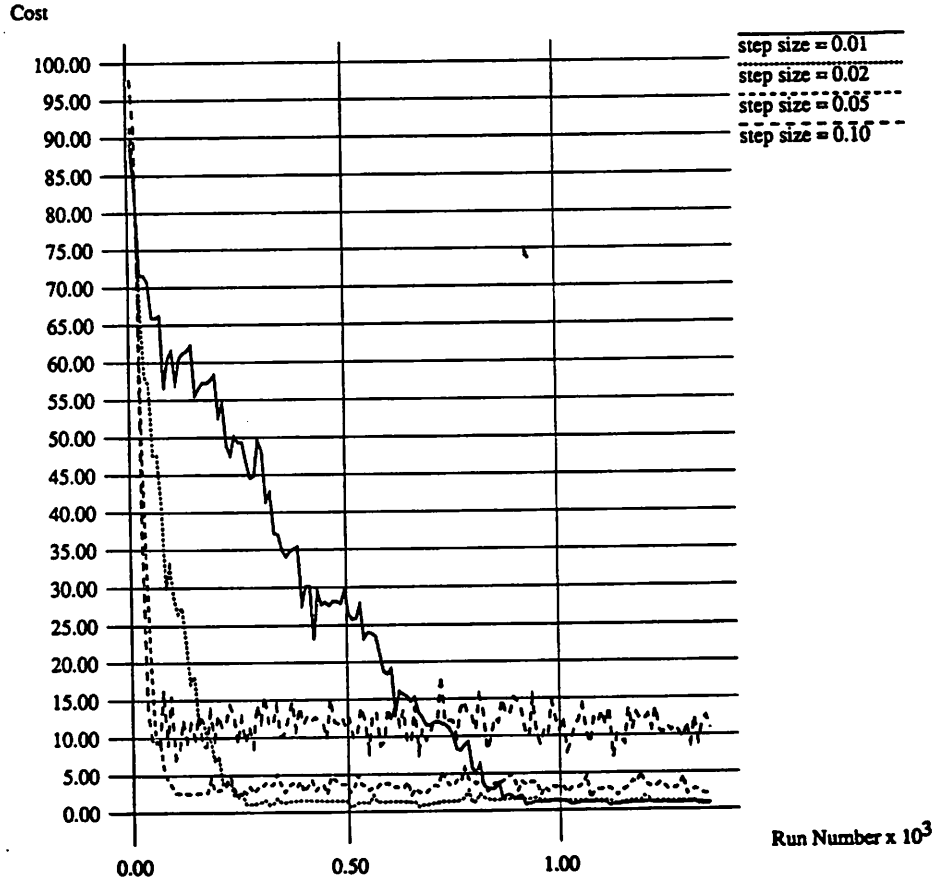
Type of Factorial	Step Size	Runs to reach optimum	Average cost after optimum is reached
Half	0.01	909	0.97
	0.02	252	1.33
	0.05	72	3.42
	0.10	27	11.65

As the step size increases, fewer runs are required to find the optimum but the average cost after the optimum is reached is larger. For the given values of the specifications and noise levels, running the system continuously at its optimum without introducing the deviations required for the factorial experiment results in an average cost of 0.87. Note that when the step size is small, the average cost is only slightly above the cost obtained when no deviations are introduced. Thus, such a small step size would have a minimal impact on the process capability. Using a large step size around the optimum degrades the process capability substantially.

However, a small step size requires a ridiculously large number of runs to reach the optimum. If the current operating point is far from the optimum, such as when the original optimizations are being done at start-up or if the equipment response has shifted, large step sizes are desirable. In a system without any noise, doubling the step size will cut the number of runs required to find the optimum in half. In a system with noise, doubling the step size will cut the runs required by more than half. Figure 4 shows the cost as a function of run number for a variety of step sizes. The simulations were done using half fractional designs.

Figure 4

Cost vs. Run Number



The average costs around the optimum for a full factorial and half fraction are virtually the same, but the number of runs to reach the optimum are significantly different, as expected. If the half fraction was equally effective as the full factorial in determining the direction to move, then both approaches would take the same number of cycles. This is true when the step size is large since the effects are much larger than the noise level. But as the step size decreases, the half fraction takes more cycles than the full factorial, but still fewer total runs. In the limiting case where the step size is made arbitrarily small, we would expect the half fraction to take the same number of runs as the full factorial to find the optimum.

5.0 Conclusion

An EVOP software package has been written and applied to the Eaton photoresist spin and bake station. Simulations have been run to verify the operation of the software and also to examine the impact of fractional factorials and step size.

Clearly, the ideal scenario would be to have a variable step size depending on the current conditions: a large step size when movement is required, a small step size when the optimum has been found. It is important to note that even after the optimum is found, the recipe does not become fixed. Instead, EVOP continues so that adaptations to shifts in the equipment can be made.

There are a couple of enhancements that could be made to decrease the response time of EVOP to changes in the equipment. For example, if large effects are calculated, then the position of the center point for the next factorial could be shifted by an amount larger than the step size used within a single factorial. In addition, if several moves are currently being made in a certain direction, then the step size in that direction could be increased as well.

When an extra cycle is inserted to update σ, the response time will be lengthened. If extra cycles need to be inserted frequently and the factorial is large, one alternative would be to simply repeat the runs at the center point to update the estimate instead of repeating the entire factorial. This would reduce the total number of runs required.

6.0 Acknowledgments

I would like to acknowledge the work of Gary May and Bart Bombay for creating the structure of the C++ equipment models and also of Sovarong Leang for developing the model for the Eaton resist coating and bake station.

7.0 Appendix

Source code for the program and a description of files are available upon request.

Use of SPC on a Wafer Track

Jorge M. Noriega-Asturias

Statistical process control is applied to a photoresist dispensing Wafer Track System. Control Charts for the individual photoresist thickness values as well as their moving range are calculated. The history of the process is also studied for comparison purposes. The process range was found to be in control, but the process average was not. Thus, the control limits calculated are useful for a short term control of the process only. Recommendations are given for further analysis and control of the process.

1.0 Introduction

The purpose of this project is to apply Statistical Process Control (SPC) to an Eaton Wafer Processing system. The Eaton wafer processing system is a spindle/hot plate/cold plate combination used to dispense photoresist to wafers in an automated manner. Track #1 is for dispensing the KTI 820 photoresist and for doing the post exposure bake. The hot plate temperature at track #1 is set at 120 °C. Track #2 is for dispensing Olin Hunt I-line photoresist and Shipley 1400-31 photoresist, and the hot plate temperature is set at 90 °C.

Two types of control charts are used in this project. One of them is the control chart for individuals, in which every single measurement is plotted. The control chart for individuals is often used in conjunction with the moving-range chart. The moving range is the absolute value of the difference between consecutive measurements. This type of chart is used in cases where it is inconvenient or impossible to obtain more than one measurement per sample, when automated testing and inspection allow measurements of every unit produced, or when data become available very slowly, and waiting for a larger sample will be impractical or make the control procedure too slow to react to problems.

The other type of chart used is the $\bar{x}$ and R chart. This chart is often useful when a new product is being manufactured by an existing process, for diagnostic purposes when the process is in trouble, or where a set-up must be evaluated. For this type of chart, a sample of size n is chosen. For each sample both the average and range are plotted in the control chart. The control limits are a function of the sample size n.

In this project it is desired to control the photoresist thickness dispensed by the Eaton wafer track. The following sections describe in detail the methodology, implementation, results and conclusion of this project.

1.1 Methodology

This section describes the approach followed to implement SPC to the Eaton Wafer Track. Data is obtained, control limits are calculated for the individual thicknesses measurements and moving range charts, and previous available data for 8 weeks is analyzed.

The new data is obtained using the "Inspector INS-800-1". The Inspector is an automated thin film thickness measurement equipment. Through the use of fiber optics and multi-wavelength reflection interferometry the Inspector makes these measurements in-situ and in quasi real time. The system is based on a 486 PC. It consists of the data station (PC), and the detector/spectrometer module coupled to a fiber optic cable. The use of fiber optics makes it possible to do in-situ measurements.The computer acquires the data from the detector/spectrometer module and computes thickness values in approximately 250 milliseconds. Multi-layer measurements as well as multi-point measurements are also possible.

Once the single measurements are taken, a simple algorithm is developed to calculate the control limits for the individual units/moving range control charts. Once the data is entered, the algorithm computes the control limits for the moving range chart. If any moving range exceeds the limits, it is eliminated, and the algorithm modifies the control limits. This iteration is repeated until all moving ranges conform to the control limits. The individual unit control limits are calculated once the moving range is stable. The same test/modify iteration as in the moving range control limits is used.

In addition to taking new data, the history of the thickness values for the previous 8 weeks are analyzed. The analysis of the history is necessary to evaluate the dispensing of photoresist between relatively long periods of time. Both the KTI-820 and the I-line photoresists are analyzed. This data was supplied by the Staff of the Berkeley Microfabrication Laboratory. It was measured by the Nanospec using program #10. $\overline{X}$ and R charts are used for the analysis of the thickness values of the wafer center.

1.2 Implementation

The measurement equipment was set up during this project. Eventually, it will be installed permanently over the wafer track, where it will take measurements over the cold plate. For this project, the measurements were taken in the test stand included with the equipment. 20 wafers were used and one measurement was taken for each wafer. Also, all measurements were taken at the center of the wafer only.

The algorithm was implemented in BASIC, and the code is included in Appendix A. It was applied to the 20 measurements obtained from the Inspector to get control limits for both the individual units, x, and moving range, MR, control charts. The control limits are given by:

Moving Range:

$$UCL = 3.267MR_{ave} \tag{1}$$

$$LCL = 0.00MR_{ave} \tag{2}$$

Individual units:

$$UCL = x_{ave} + 2.659MR_{ave} \tag{3}$$

$$LCL = x_{ave} - 2.659MR_{ave} \tag{4}$$

$$CL = x_{ave} \tag{5}$$

The history of the thickness values was analyzed using the $\overline{X}$ and R charts. A sample size n=5 was used in this analysis. For each sample, the average, $\overline{X}$, and the range, R, was calculated. For the 8 samples the total average, $\overline{X}_{ave}$, as well as the range average, R_{ave}, was also calculated. The control limits are giving by:

Range chart:

$$UCL = 2.115R_{ave} \quad (6)$$

$$LCL = 0.00 \quad (7)$$

$$CL = R_{ave} \quad (8)$$

X chart:

$$UCL = \overline{X}_{ave} + 0.577R_{ave} \quad (9)$$

$$LCL = \overline{X}_{ave} - 0.577R_{ave} \quad (10)$$

$$CL = \overline{X}_{ave} \quad (11)$$

1.3 Results

Figure 1 and figure 2 show the control charts for the individual units and moving range, respectively, for the resist thickness values measured by the Inspector. We can see that at the beginning of the resist coating, the thickness values plot outside the control limits. It is not known when the Wafer Track was last used, so the cause for that out of control situation was assigned to be the problems associated with the beginning of a new resist coating process. The algorithm eliminated those points and recalculated both the moving range and individual units control limits. The final limits are the ones shown in figure 1 and figure 2.

Figure 3 and figure 4 show the control charts for the KTI-820 and I-line photoresists, respectively. For both resists, the range between samples is seen to vary much less than the $\overline{X}$ between samples. The process is not in control, therefore control limits do not make sense for this data. Anyway, the variation of the I-line resist is less than that of the KTI-820.

1.4 Conclusions

SPC was applied to an Eaton Wafer Track. The purpose was to control and monitor the photoresist coating done by that Wafer processing system. New measurements were done by the "Inspector INS-800-1", a personal computer controlled measurement equipment. This equipment has the advantage of allowing quick, in-situ measurements.

Control limits for the individual units and moving range control charts were calculated by a simple algorithm implemented in BASIC. The control charts are shown in figure 1 and figure 2. Analysis of the history of the last 8 weeks showed that even though the range of the process between weeks does not change noticeably, $\overline{X}$ does. In the long term the process is not in control, and control limits are meaningless.

From the analysis of the results obtained, I recommend that several issues be addressed to fully implement SPC to the Wafer Track. As the thickness values seem to vary at least from week to week, new control limits should be calculated every week. The cause for the week to week variations should be investigated. Those causes are producing a shift in the average of the sample, even though the range seems to remain in control. If no particular cause is found, autocorrelation functions could be used to model the time dependance, and control charts could be applied to the residuals.

FIGURE 1. Control Chart for Individual thickness values

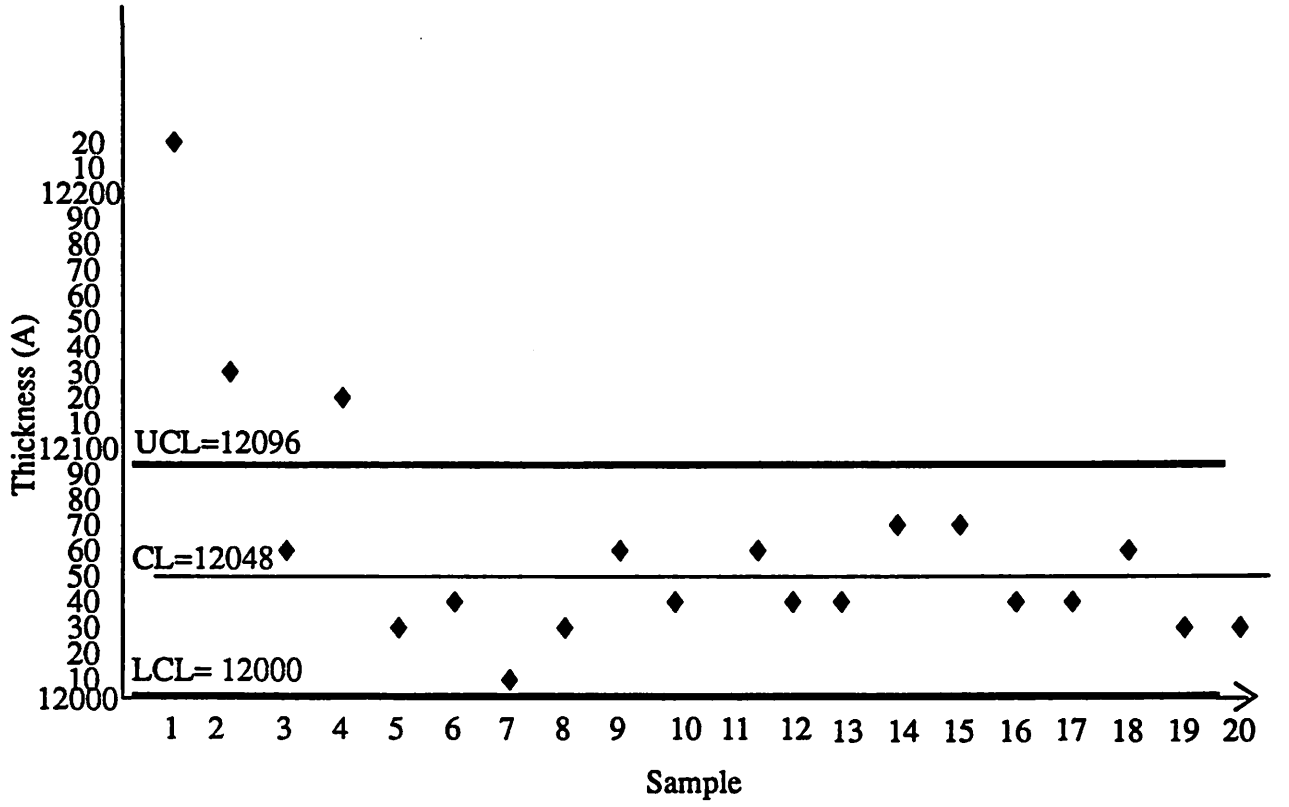


FIGURE 2. Control chart for moving average of thickness values

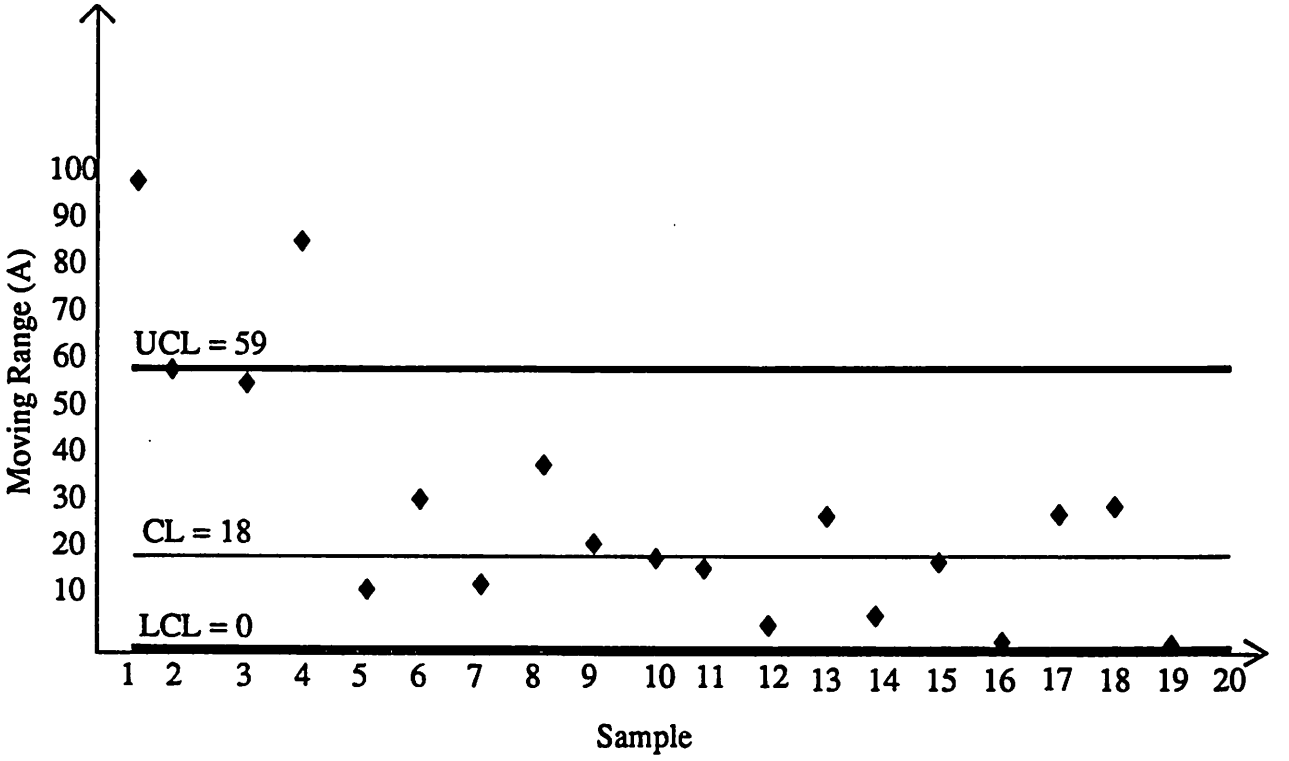


FIGURE 3. Control Chart for KTI-820 photoresist

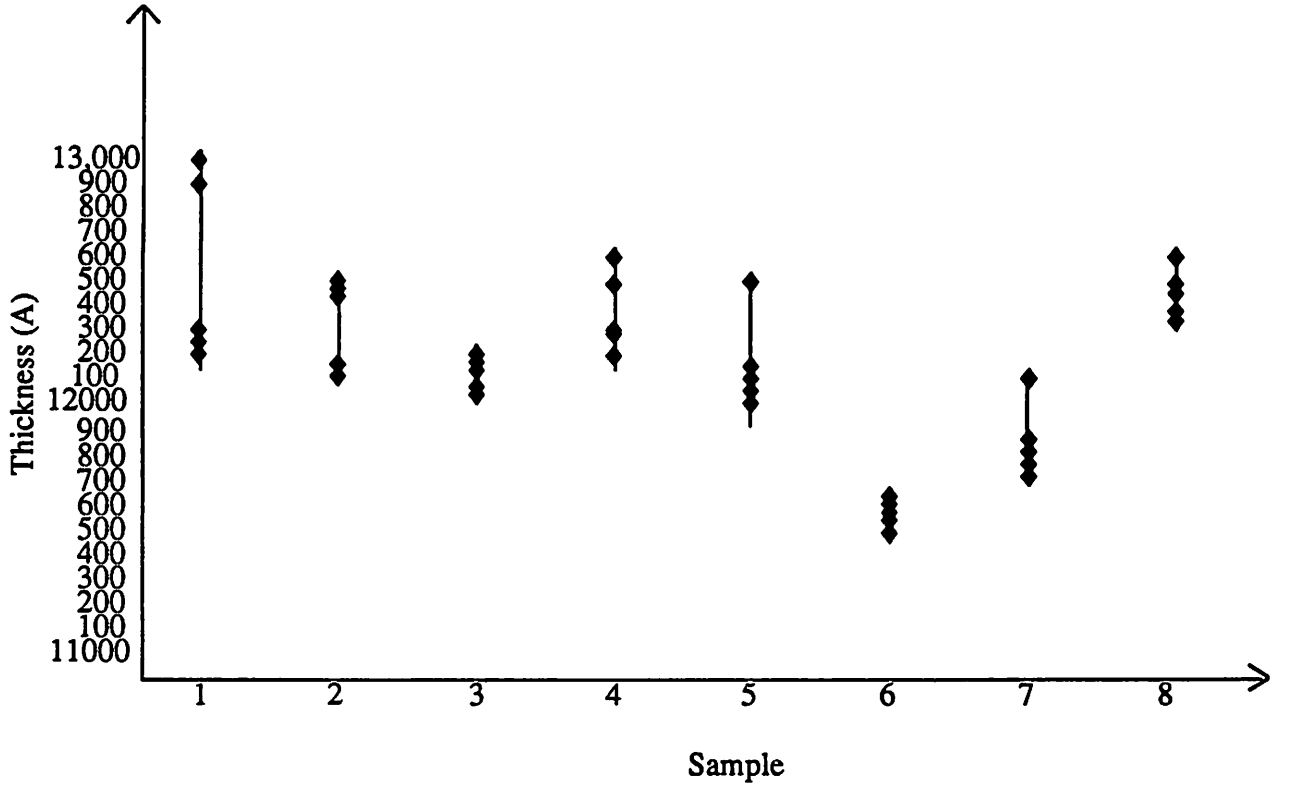
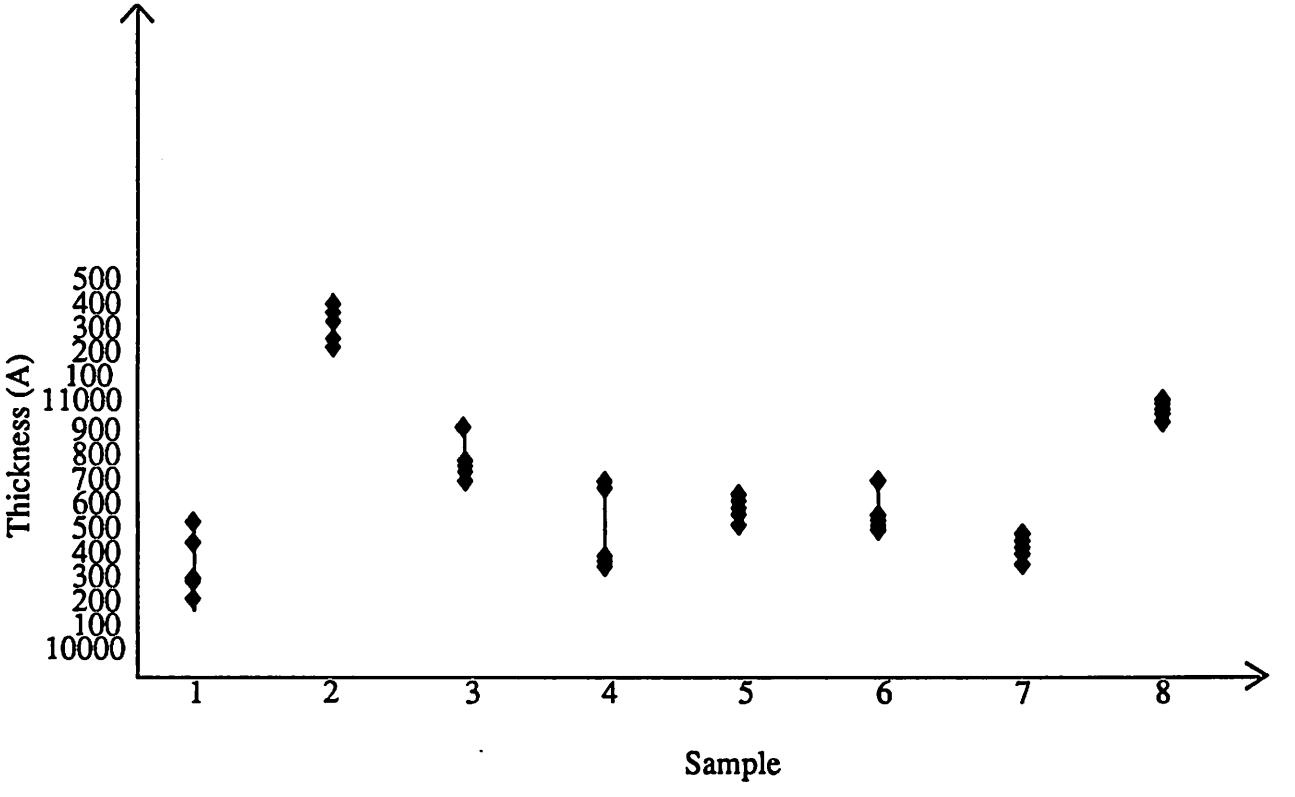


FIGURE 4. Control chart for I-line photoresist



Appendix A

BASIC code for the algorithm to calculate/modify control limits for individual units and moving average:

```
10 REM CALC/MODIFY CONTROL LIMITS FOR INDIV. UNITS/MOVING AVG.
20 CLEAR
30 INPUT "NO. OF WAFERS=";N
40 DIM X(N), MR(N-1)
100 REM READ DATA
110 FOR I=1 TO N
120 PRINT "THICKNESS("; I ; ")";
130 INPUT X(I)
140 NEXT I
200 REM CALC. MR, MR-LIMITS
210 PRINT "USING"; N ; "SAMPLES",
220 MRSUM=0
230 FOR I=1 TO N-1
240 MR(I)=ABS(X(I)-X(I+1))
250 PRINT "MR(";I")=";MR(I),
260 MRSUM=MRSUM + MR(I)
270 NEXT I
280 MRAVE=MRSUM / (N-1)
290 MRCL = MRAVE : MRLCL=0
300 MRUCL = 3.267 * MRAVE
400 REM MODIFY MR C-LIMITS IF NECESSARY
410 M=N-1
420 A=0
430 FOR I=1 TO N-1
440 IF MR(I)>MRUCL OR MR(I) < MRLCL THEN
PRINT "MR(";I;") IS OUT OF RANGE", :
MRAVE=(MRAVE - MR(I)/M)*M/(M-1) :
M=M-1:
MRCL=MRAVE:
MRUCL = 3.267 * MRAVE :
MR(I)=0 :
A=A+1
450 NEXT I
460 PRINT N-1-M; "MR OUT OF RANGE",
470 IF A>0 THEN GOTO 420
500 REM CALC. XAVE, XUCL AND XLCL
510 XSUM=0
520 FOR I=1 TO N
530 XSUM = XSUM + X(I)
540 NEXT I
550 XAVE = XSUM / N
560 XCL=XAVE
570 XUCL = XAVE + 2.6596*MRAVE
580 XLCL = XAVE - 2.6596*MRAVE
600 REM MODIFY X CONTROL LIMITS IF NECESSARY
610 X=N
620 A=0
630 FOR I=1 TO N
640 IF X(I) < XLCL OR X(I) > XUCL THEN
PRINT "X("; I ;") IS OUT OF RANGE",:
A=A+1
XAVE = ( XAVE - X(I) / X ) * X / (X-1):
X=X-1:
X(I)=0:
XCL=XAVE:
XUCL = XAVE + 2.6596*MRAVE:
XLCL = XAVE - 2.6596*MRAVE
650 NEXT I
660 PRINT N-X; "X'S WERE OUT OF RANGE",
700 REM RECALCULATE X(I) IF NECESSARY
710 A=0
720 FOR I=1 TO X
730 IF X(I+A)=0 THEN
A=A+1:
GOTO 730
740 X(I)=X(I+A)
750 NEXT I
760 N=X
770 IF A>0 THEN GOTO 200
800 REM PRINT CONTROL LIMITS
810 PRINT "MOVING RANGE CHART:",
820 PRINT "CL=";MRCL,
830 PRINT "UCL=";MRUCL
840 PRINT "LCL=";MRLCL,
850 PRINT "INDIVIDUAL THICKNESS CHART:",
860 PRINT "CL=";XCL,
870 PRINT "UCL=";XUCL,
880 PRINT "LCL=";XLCL,
900 END
```

Using Orthogonal Arrays to Optimize a Phase-shift On Substrate Process

Debra L. Hebert

This report describes a Taguchi-based experiment, as it was applied to optimize a Phase-Shift On Substrate (POST) lithographic process. In this experiment we have employed the L_8 orthogonal array in order to analyze and optimize the main effects and to estimate their first order interactions. The experiment was completed in the Berkeley Microfabrication Laboratory.

1.0 Introduction

Continued miniaturization of semiconductor devices is largely dependent on whether photolithographic technologies can be developed that will produce features in the deep sub-micron range (0.1 μm - 0.5 μm). There have been several advances in photolithography, and some of the more promising techniques for achieving these smaller feature sizes are: phase-shifting optical techniques, shorter wavelength optical techniques, electron beam direct writing, and x-ray lithography. I have chosen to investigate phase-shifting techniques because unlike the other technologies, it does not require a new exposure tool. The phase-shifting effect was first investigated by M.D. Levenson in 1982 at IBM, but it is only recently that significant efforts have gone into developing this technology. The phase-shift effect is usually achieved through the use of specially manufactured phase-shift masks, but a group at SHARP corporation of Japan [1] has developed a technique which uses the photoresist on the wafer to create the phase-shifter.

2.0 Phase-Shift on Substrate (POST) Concepts

The principle behind the POST technique is illustrated in Figure 1, while the basic process sequence is illustrated in Figure 2. The process can be summarized as follows:

1. Partial exposure of the resist using a conventional mask.

2. Development to remove the exposed part of the resist layer.

3. Flood exposure without a mask using the resist phase-shifter created during the first exposure.

4. A second develop cycle to define the resist pattern occurring at the boundary of the mask pattern from the first exposure.

The depth, **d**, of removal necessary to create the phase-shifter is given by the following expression:

$$d = \lambda(2m-1)/2(n-1)$$

m = natural number (1,2,3,...)

λ = exposure wavelength

n = refractive index of the resist

With the POST process, the project team at SHARP was able to define 0.15 μm wide lines with a 0.5 μm pitch using a conventional photo mask with 0.5 μm lines and spaces. They used an i-line stepper with a lens NA of 0.54 to make the first exposure, and an i-line stepper with a lens NA of 0.45 to do the second exposure. The higher NA gives better resolution, while the lower NA gives better depth of focus.

3.0 Experimental Design

The experiment was designed using orthogonal array matrices from the Robust Design method founded by Dr. Genichi Taguchi [2]. Ordinarily, this method is used only when the chosen control factors are known not to interact with each other. However, several of the standard arrays can be used to estimate factor interactions, and Taguchi has designed standard linear graphs and interaction tables (Figure 3.) which make it easy to determine factor column assignments in order to avoid confounding the main factor effects with interaction effects. I have chosen to investigate the effects of four control factors at two levels, and three interaction effects at the same levels (Table I). The orthogonal array which is best suited for this experiment is the L_8 standard orthogonal array (Table II).

4.0 Data Analysis

The critical parameter for achieving the phase-shift effect is the thickness, **d**, of the resist that is removed during the first exposure. The goal of this experiment is to optimize the wafer-to-wafer thickness removal uniformity. The first step in analyzing the data is to calculate the signal-to-noise (S/N) ratio for the thickness data η,

$$\eta = 10 \log_{10} (\mu^2/\sigma^2)$$

where:

$$\mu = 1/25 \sum \sum \tau_{ij}$$

$$\sigma^2 = 1/24 \sum \sum (\tau_{ij} - \mu)^2$$

τ = measured thickness

i = number of wafers measured

j = measurement sites on the wafer (T, C, F, L, R)

After the S/N ratios have been calculated for each of the factors, the results can be plotted to determine optimum (maximum) values for the S/N (in dB). The relative effect of each factor can be determined by analysis of variance (ANOVA). ANOVA also gives information about the error variance and prediction error variance. The interaction effects can be estimated by determining the average responses for a particular combination of levels for the factors. The results are plotted to determine the nonparallelism of the factor effects (Figure 4).

5.0 Experimental Procedure

Five bare silicon wafers were coated for each of the experimental runs (the soft bake will vary according to the experimental array). The initial resist thickness was measured at 5 points on the wafer (T, C, F, L, R) using the nanospec. The wafers were then exposed on the gcaws, baked at the time and temperature indicated in the experimental array, and then developed using the standard develop cycle. The residual resist thickness was measured again at five points on the wafer to calculate resist removal thickness **d**.

6.0 Results

The ANOVA results for the experiment are shown in Table III. The interaction effects are plotted in Figure 5.

7.0 Conclusions

Conditions That Maximize S/N (optimum):

Softbake Temperature	70°C	Level 1
Softbake Time	60 secs	Level 2
PEB Temperature	110°C	Level 2
PEB Time	60 secs	Level 2

Major Factor Effect

Softbake time is responsible for 47.63% of the total variation

Interactions Exist

Between Factors A and C (softbake temperature and PEB temperature)

Between Factors B and C (softbake time and PEB temperature)

Further Work

Verify optimum conditions; this can be accomplished by taking SEM photos

8.0 References

[1] H. Tabuchi et al., "Novel 0.2μm i-Line Lithography by Phase-Shifting on the Substrate (POST)", IEDM Technical Digest, 1991, pp. 63-66.

[2] M. Phadke, Quality Engineering Using Robust Design. New Jersey: Prentice Hall, 1989.

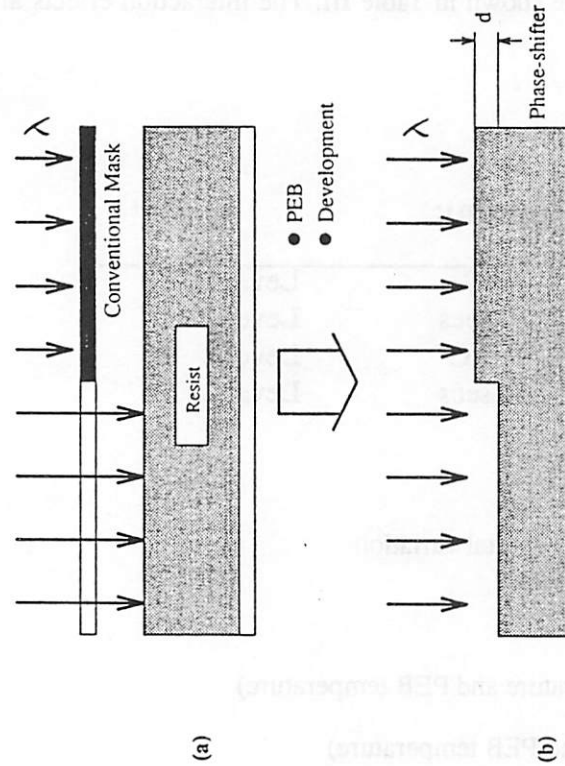


Figure 2. POST Process Sequence
(a) First exposure with conventional masks
(b) Second exposure without masks
(c) Fine pattern formed along shifter edge

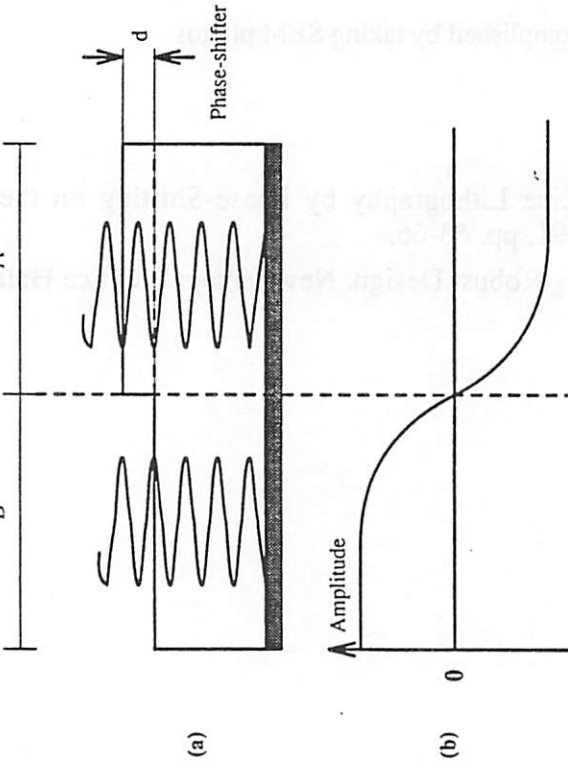


Figure 1. Principle of POST
(a) Phase-shifting on the substrate during second exposure
(b) Amplitude distributions
(c) Intensity distribution (decreased at resist step edge)

Interaction Table for L_8

Column	Column 1	2	3	4	5	6	7
1	(1)	3	2	5	4	7	6
2		(2)	1	6	7	4	5
3			(3)	7	6	5	4
4				(4)	1	2	3
5					(5)	3	2
6						(6)	1
7							(7)

Linear Graphs for L_8

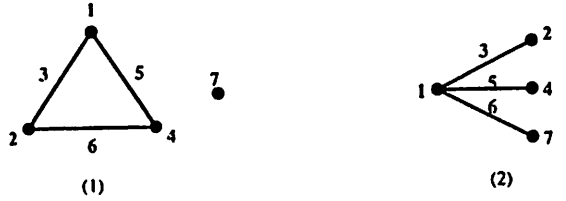


Figure 3.

Calculating Interaction Responses

		Level of Factor B	
		B_1	B_2
Level of Factor A	A_1	$\frac{y1+y2}{2}$	$\frac{y5+y6}{2}$
	A_2	$\frac{y3+y4}{2}$	$\frac{y7+y8}{2}$

Plot of 2 Factor Interaction

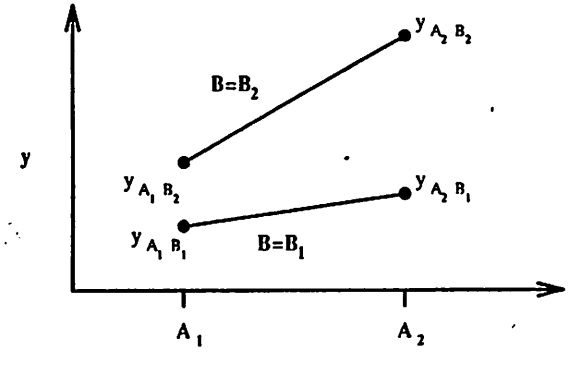


Figure 4.

Table I. Factors and Their Levels

Factor	Levels	
	1	2
A. Softbake Temperature (°C)	70	90
B. Softbake Time (secs)	30	60
C. Post Exposure Bake Temperature (°C)	90	110
D. Post Exposure Bake Time (secs)	30	60

Table II. Orthogonal Array

Expt. No.	Column Number and Factor Assignment						
	1 (A)	2 (B)	3 (AXB)	4 (C)	5 (AXC)	6 (BXC)	7 (D)
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

Table III. Analysis of POST Data

Factor	Average η by Level (dB)		Degree of Freedom	Sum of Squares	Mean Square	F
	1	2				
A. Softbake Temp.	13.93	13.40	1	0.56†	0.56	
B. Softbake Time	8.63	18.70	1	202.81	202.81	4.3
C. PEB Temp.	11.68	15.66	1	31.68	31.68	
D. PEB Time	13.05	14.29	1	3.07	3.07	
Error			3	187.63†	62.54	
Total			4	425.75	106.43	
(Error)			(4)	(188.19)	(47.05)	

Overall mean η = 13.67 dBam.
† Indicates the sum of squares added together to form the pooled error sum of squares shown in the parantheses.

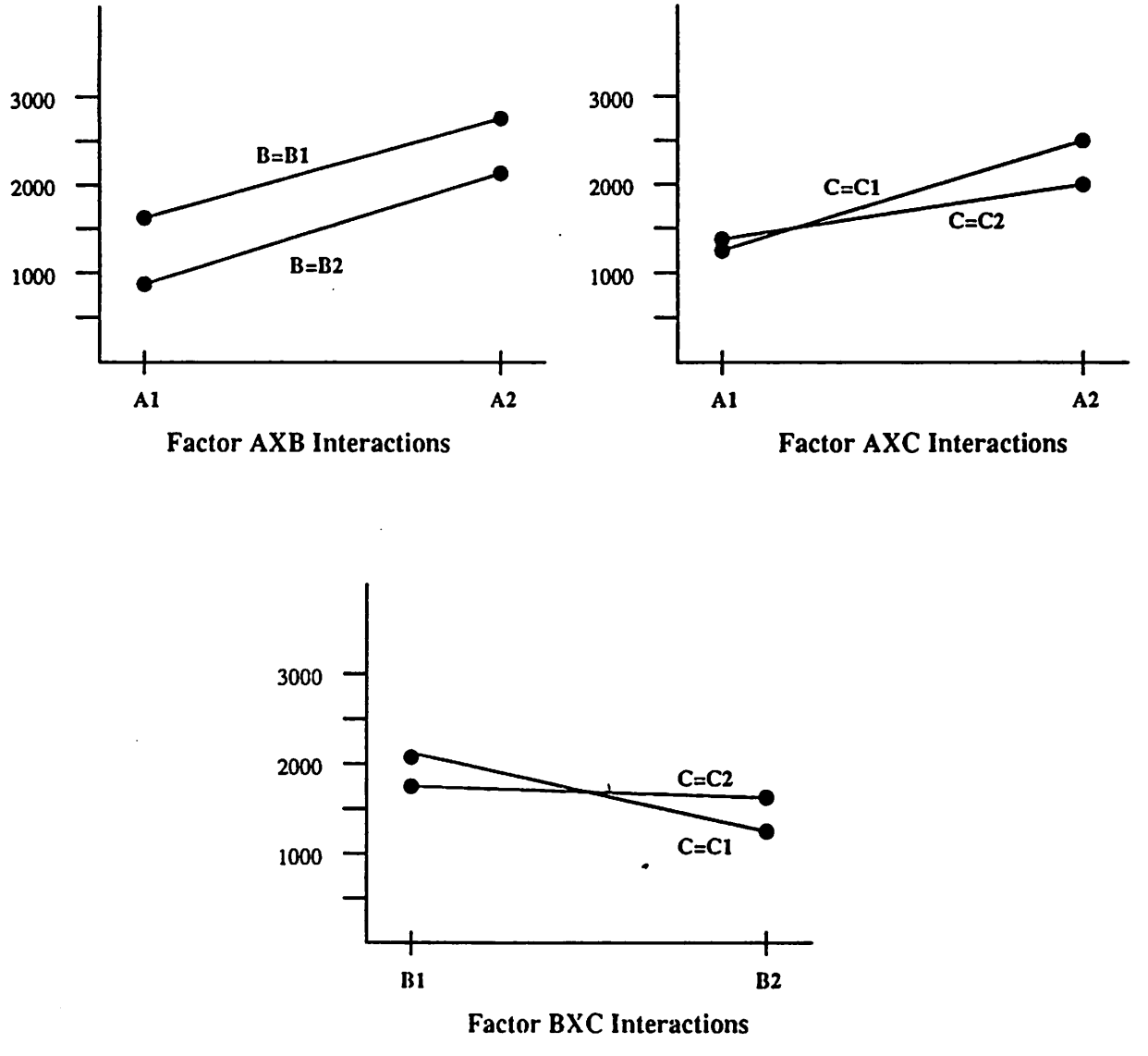

3000
2000
1000
B=B1
B=B2
A1
A2
Factor AXB Interactions
3000
2000
1000
C=C1
C=C2
A1
A2
Factor AXC Interactions
3000
2000
1000
C=C2
C=C1
B1
B2
Factor BXC Interactions

Figure 5.

Improving LPCVD Thin Oxide Quality by Using Robust Design Methodology

Joseph C. King

A designed experiment using a standard orthogonal array was used to improve the quality of thin LPCVD oxide. Important factors affecting the uniformity of the film thickness were identified and their levels for obtaining optimal oxide quality were decided. Verification experiment confirmed the result of the analysis and the prediction of the model.

1.0 Introduction

Integrated circuit MOS devices usually use thermally-grown thin oxide as the gate dielectrics because of its high quality and controllability. However, the quality of thermal oxide strongly depends on the substrate, which makes thermal oxide unsuitable in certain cases where high quality substrate is not available, like thin film transistors built on polycrystalline silicon or amorphous silicon. Oxide grown by low-pressure chemical vapor deposition has been considered as an alternative because its quality is virtually independent of that of the substrate. On the other hand, when combined with conventional thermal oxide, CVD oxide even shows certain superior properties which are not obtainable in thermal oxide[1, 2].

The LPCVD oxide in IC processes is most commonly used as thick (2000-5000Å) isolation layers, therefore, films when medium across-wafer uniformity (10%) and high deposition rate (200Å/min) are desired. However, in thin gate dielectric (50-150Å), fast growth rate is not necessary because the deposition time in very short, but high degree of uniformity is very important since the thickness changes the threshold voltages of devices, and directly affects the circuit performance.

In this project, we use Taguchi's orthogonal array to improve the LPCVD process for thin oxide deposition in the Microfabrication Laboratory. Based on the generic recipe of the Tylan 12 LPCVD furnace and operation condition formally set by Jack Lee[1], we use a designed experiment to optimize the process for high oxide quality.

2.0 Methodology

A standard $L_9(3^4)$ orthogonal array was used to explore and improve the deposition process. First, important variables and their levels were decided after a thorough check of the process conditions. Then the appropriate orthogonal array was chosen and a series of experiment runs with different variables levels was planned.

After completing all experiment runs and collecting the data, we use analysis of means (ANOM) and analysis of variance (ANOVA) to find out the important variables and their optimum levels. We can then build a simple linear model based on the analyzed data. This model is used to estimate the result of the predicted optimal conditions.

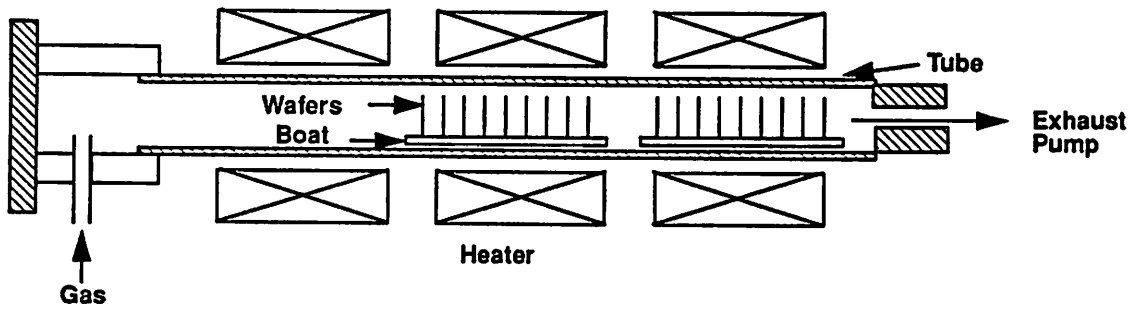
Finally, a verification experiment was run to check the accuracy of the prediction, and to draw some conclusions about the process under study.

3.0 Implementation

3.1 Furnace

The experiment was run on the Tylan12 furnace, which is a hot wall low pressure CVD tube in the Berkeley Microfabrication Laboratory. This furnace is depicted in Figure 1.

FIGURE 1. Schematics of the LPCVD tube.



The reaction temperature is maintained by heaters in 3 different zones through feedback control. Gases used are silane (S_iH_4) and oxygen (O_2). Phosphine (PH_3) can also be used to dope the S_iO_2 but is not used in this experiment. The whole process is controlled by a computer after the recipe is loaded and the proper values of the parameters are set.

3.2 Important Variables and Their Levels

The important factors which can be directly controlled are the deposition temperature and the gas flow rates. From experience, we know that a low silane/oxygen ratio should be used to obtain controllable and repeatable deposition rate and good uniformity. The deposition process is very sensitive to the temperature but only a small range of temperature can be used since too low a temperature will result in poor electrical properties and these properties are not easily measurable. Although the temperature of different zones can be specified differently, we usually use the same temperature for all the three zones.

After choosing the temperature, the silane flow rate and the oxygen flow rate as the variables, we found we can still have another factor to use the $L_9(3^4)$ orthogonal array. The next important factor which is usually believed to have certain effect on the uniformity and deposition rate is the orientation of the wafers in the furnace (facing the inside of the tube or outside). The final experimental matrix is shown in Table 1. The levels marked with asterisk (*) are the formally used operating conditions. Because the factor C (wafer orientation) can only have two levels (in or out), we repeat the "out" condition and this replication can be used to obtain an independent estimate of the experimental error.

TABLE 1. The experiment matrix

Factor	Levels		
	1	2	3
A:Temperature (C)	440	450*	460
B: Silane Flow (sccm)	0.5	1*	2
C: Wafer Orientation	Facing Out*	Facing Out*	Facing In
D: Oxygen Flow (sccm)	70	90*	110

3.3 Experiment Runs

Table 2 shows the experiment runs and their individual settings. The run number is the actual sequence in which each run is executed to reduce confounding with the aging effect of the tube. In each run, eight test bare silicon wafers are loaded into the tube and the respective recipe is loaded. The main deposition time is chosen to be 10 minutes but the whole process takes about two hours, therefore achieving high deposition rate is not very important in this process.

TABLE 2. The experiment runs and results

	Column Number and Factor Assigned					Observations		
Expt. No.	Run No.	1 Oxygen flow (D)	2 Temp (A)	3 Silane flow (B)	4 Orientation (C)	Q_1 (dB)	Q_2 (dB)	Q_3 (dBam)
1	5	1	1	1	1	32.16	20.60	22.43
2	1	1	2	2	2	30.55	20.86	23.73
3	8	1	3	3	3	27.16	21.87	25.10
4	3	2	1	2	3	30.34	19.73	22.98
5	7	2	2	3	1	27.70	20.49	24.35
6	2	2	3	1	2	25.96	18.77	23.04
7	9	3	1	3	2	28.00	21.33	24.19
8	6	3	2	1	3	26.76	21.67	22.75
9	4	3	3	2	1	24.06	19.24	23.43

After each run is done, we measure the deposited oxide film thickness using the ellipsometer which is good for measuring thin layers. Thickness values of the film for five locations are recorded and a total of 40 values are collected in each run.

The observations Q_1,Q_2 and Q_3 indicate the within-wafer uniformity, between-wafer uniformity and the deposition rate, which are calculated by using the following equations;

$$\mu_i = \frac{1}{5}\sum_{j=1}^{5}\tau_{ij} \qquad \sigma_i^2 = \frac{1}{4}\sum_{j=1}^{5}(\tau_{ij}-\mu_i) \qquad \text{(within each wafer)}$$

$$\mu = \frac{1}{40}\sum_{ij}\tau_{ij} \qquad \sigma^2 = \frac{1}{39}\sum_{ij}(\tau_{ij}-\mu)^2 \qquad \text{(Within each run)}$$

$$Q_1 = 10 \cdot \frac{\sum_{i=1}^{8}\log\left(\frac{\mu_i^2}{\sigma_i^2}\right)}{8} \qquad \text{With-in Wafer Uniformity}$$

$$Q_2 = 10 \cdot \log\left(\frac{\mu^2}{\sigma^2}\right) \qquad \text{Between Wafers Uniformity}$$

$$Q_3 = 20 \cdot \log\left(\frac{\mu}{10}\right) \qquad \text{Deposition Rate}$$

4.0 Results

4.1 Analysis of Data

The result of each experiment shown in Table 2 is analyzed using analysis of mean (ANOM) and analysis of variance (ANOVA). Tables 3 to 5 show the result of analysis and Figure 2 plots the effect of the four factors to the within-wafer uniformity, between-wafer uniformity and deposition rate, respectively. Also shown in Figure 2 are the 2-sigma (95%) confidence levels.

4.1.1 Within-Wafer Uniformity

From Table 3, we can see that the temperature and oxygen flow rate have strong effect on the uniformity within a single wafer. Low temperature and low oxygen flow rates result in high uniformity, while silane flow rate and wafer orientation are not significant factors.

4.1.2 Between-Wafer Uniformity

The uniformity between wafers shows insignificant dependence on the silane flow rate and the oxygen flow rate. The noise level can be estimated by looking at the effect of wafer orientation (factor C) since level 1 and 2 for factor C are essentially the same and any difference should be due to noise.

TABLE 3. Average within-wafer uniformity (Q_1) by factor levels (dB)

Factor	Ave Q_1 by Level (dB)			Degree of Freedom	Sum of Squares	Mean Square	F
	1	2	3				
A: Temperature	30.17	28.34	25.73	2	29.87	14.94	59.75
B: Silane Flow	28.29	28.32	27.62	2	0.94	0.47	
C: Orientation	27.97	28.17	28.09	2	0.06	0.03	
D: Oxygen Flow	29.96	28.00	26.27	2	20.38	10.19	40.75
Error				0	0.00		
Total				8	51.25	6.41	
(Error)				4	1.00	0.25	

*Overall mean $Q_1 = 28.08$

TABLE 4. Average between-wafer uniformity (Q2) by factor levels (dB)

Factor	Ave Q_2 by Level (dB)			Degree of Freedom	Sum of Squares	Mean Square	F
	1	2	3				
A: Temperature	20.55	21.01	19.96	2	1.65	0.83	
B: Silane Flow	20.35	19.94	21.22	2	2.57	1.28	
C: Orientation	20.11	20.31	21.09	2	1.60	0.80	
D: Oxygen Flow	21.11	19.66	20.74	2	3.39	1.69	
Error				0	0.00		
Total				8	9.21	1.15	
(Error)				(8)	(9.21)	(1.15)	

*Overall mean $Q_2 = 20.50$

4.1.3 Deposition Rate

While reaching a high deposition rate is not a goal of this study, we still monitored the average deposition rate and found that the silane flow rate is the only important factor. To gain high deposition rate, higher silane flow rate should be used and that is quite reasonable because the deposition rate is actually limited by the supply of silicon atoms.

TABLE 5. Average deposition rage by factor levels (dBam)

Factor	Ave Q_3 by Level(dBam)			Degree of Freedom	Sum of Squares	Mean Square	F
	1	2	3				
A: Temperature	23.20	23.61	23.86	2	0.66	0.33	
B: Silane Flow	22.74	23.38	24.55	2	5.03	2.52	16.01
C: Orientation	23.40	23.65	23.61	2	0.11	0.05	
D: Oxygen Flow	23.75	23.46	23.46	2	0.18	0.09	
Error				0	0.00		
Total				8	5.98	0.75	
(Error)				6	0.94	0.16	

*Overall mean $Q_3 = 23.56$

4.2 Model Building and Confirmation Runs

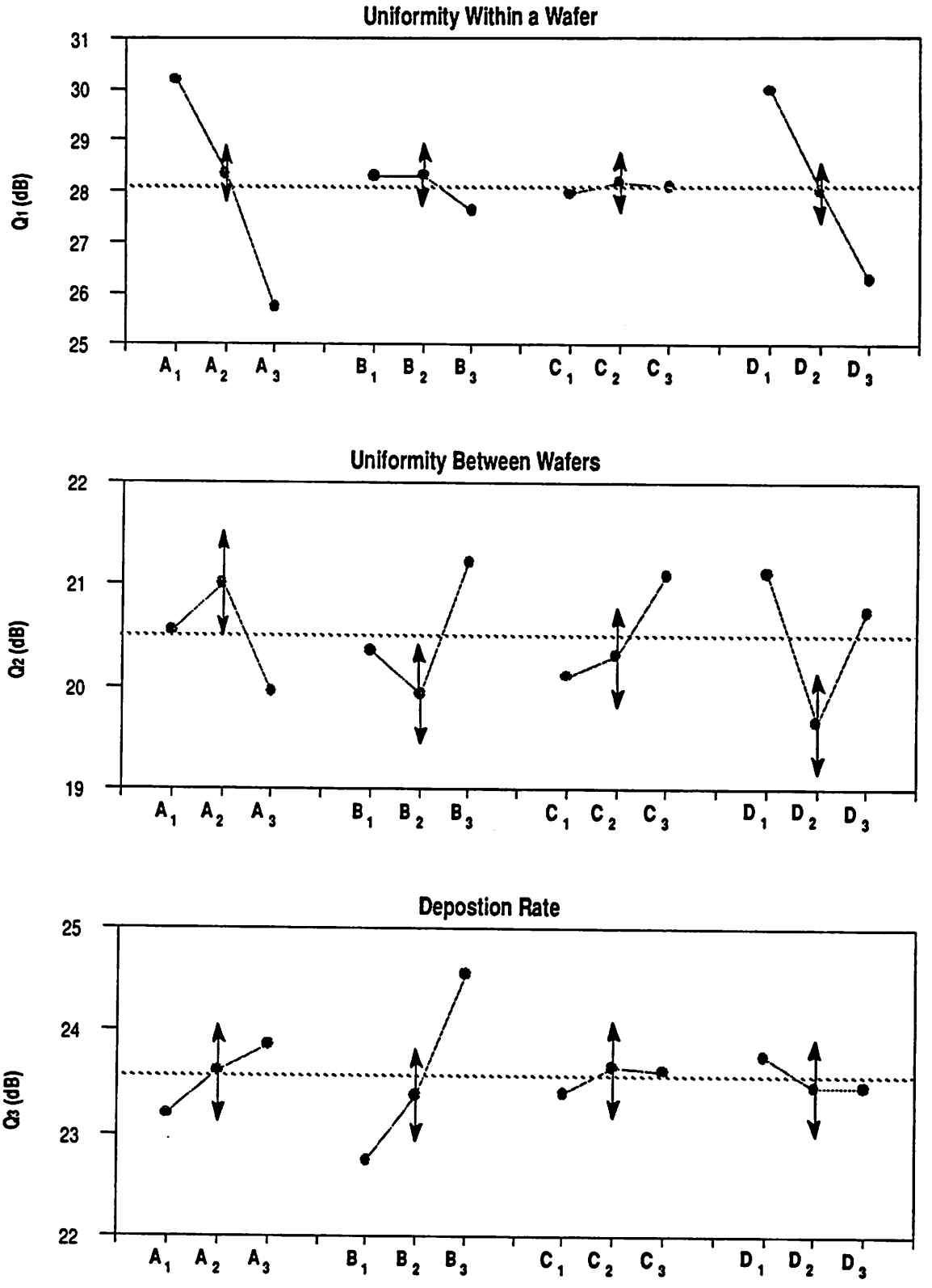
We can build a simple linear model based on the analysis above to find out the optimal operating condition and predict the result. The error in the prediction can also be obtained by using the estimated error in the analysis above. In choosing the optimal operation conditions, we can get highest uniformity by choosing the combination of $A_1B_3D_1$ and still have reasonable deposition rate. Because high deposition rate is not necessarily, no trade-off has to be made in choosing the conditions.

Confirmation experiment runs were executed twice for the determined optimal operating point (temperature = 440, silane flow rate = 2, oxygen flow rate = 110). The result is within the range of the prediction based on the model. Table 6 summarized the result of the prediction and the confirmation runs.

TABLE 6. The result of the verification runs

Experiment	Q_1 (dB)	Q_2 (dB)	Q_3 (dBam)
Experiment 1	32.96	20.98	24.86
Experiment 2	32.58	21.12	24.30
Average	32.77	21.05	24.58
Predicted	32.05 +/- 1.03	20.96 +/- 1.85	24.19 +/- 0.82

FIGURE 2. The plots of the effects.



5.0 Conclusions

A statistical experimental design using the robust design methodology is applied to thin oxide LPCVD process. Based on the result of the experiment, we found that the average uniformity within a wafer is affected by temperature and oxygen flow rate and the average uniformity between wafers in a single run is affected by the flow rate of silane and oxygen. The orientation of the wafer in the boat is actually not an important factor. Confirmation runs were done after the analysis and model building, and were consistent with the model predictions.

6.0 Reference

[1] J. Lee, C. Hegarty, and C. Hu, "Electrical Characteristics of MOSFET's Using Low-Pressure Chemical-Vapor-Deposited Oxide," *IEEE Electron Device Lett.*, vol. 9, no. 7, p. 324, Jul. 1988.

[2] J. Ahn, W. Ting, and D. Kwong, "High-Quality MOSFET's with Ultrathin LPCVD Gate SiO_2," *IEEE Electron Device Lett.*, vol. 13, no.4, p.186, Apr. 1992.

[3] M. S. Phadke, "Quality Engineering Using Robust Design", Prentice-Hall, 1989.

Spatial defect statistics & In-situ monitoring of contamination

Sean Patrick Cunningham

Airborne particulates in processing equipment can cause catastrophic yield loss in semiconductor products. A particle defect simulator is developed along with statistical routines to test the goodness of fit of various hypothesized distributions of particles on the simulated defect maps. These routines make use of the quadrat method for analyzing spatial dispersion. The negative binomial distribution is found to fit simulated data provided quadrat sizes are made small enough. In addition, a 2^{5-1} experiment is presented in which the particle count in a plasma etcher is measured for different settings of etch rate, gas flow, chamber cleanliness, and polysilicon type. The results of this experiment are inconclusive, but future experiments are discussed based on the shortcomings of this experiment.

1.0 Introduction

This report presents research in semiconductor yield modeling. Specifically, this report documents two related efforts to understand catastrophic yield caused by airborne particulates in processing equipment. While semiconductor processing is performed in a clean environment, there is still potential for particles to land on wafers and destroy circuits. With advances in cassette containers for wafers such as SMIF boxes and other efforts to reduce environmental cleanliness, the problem of airborne particles is slowly being reduced. However, within processing equipment, there is the potential for particles to fall on wafers. Unlike the randomness of environmental particles, equipment particles may be expected to fall in patterns which may be thought of as signatures. Paz and Lawson [8] discovered a radial dependence in defect patterns for diffusion in LSI processes, a result which has been replicated frequently. Whether this sort of patterning effect exists for other processing equipment is important to production planners, chip designers, and process controllers.

Work has been done describing the statistics of dispersion, and much of this originally came from forestry and urban operations research applications. The inferential question of what process caused a particular data set to occur is very important to each of these fields. Section 3 provides background regarding the origin of some of the common yield models used. These models can be descriptive of the *result* of a given process, but they are inadequate for determining the causes. For instance, the yield of a process may be modeled as a negative binomial random quantity[1], but this does *not* imply the existence of a negative binomial random generator for yield. A given clustering model may be used to describe the result without disclosing the cause of that result. In fact, preliminary results discussed in this report show that the negative binomial model may be used to describe a simulated defect generating process which evolves without regard to the assumptions of that model.

Non-functional chips are observable, but often the defects which cause them are not. However, using a laser driven particle counter installed on the exhaust vent of a given piece of processing equipment, the

1. The term random quantity is equivalent to the term random variable. The term random quantity is used throughout this report.

defect process may be better understood. Particle counts are important, and Section 4 documents the result of an experiment which investigated particle counts; however, the time-series of data may also yield insight into the defect mechanisms of processing equipment.

Section 5 discusses the work done for this report in the larger context of future research opportunities in the field of equipment based yield modeling.

2.0 Spatial defect simulator

A spatial defect simulator is developed. Inferential statistical tests are then applied to the simulated defect maps to determine appropriate distributions to describe the dispersion of particles. The motivation for developing this simulator is outlined in a brief review of catastrophic yield research.

2.1 Yield models

The clustering of defects on wafers is a well documented phenomenon. Cunningham [3] gives a good history of yield models as they have evolved from simple, pessimistic Poisson models to more elaborate models. People learned that the yield of LSI chips was chronically underestimated by the Poisson model. This led to a flurry of activity in formulating modified models for defect density and yield prediction. Many of these involved convolving a Poisson kernel against some other distribution $f(\lambda)$ as shown in equation (1).

$$P(k) = \int_0^\infty e^{-\lambda}\frac{\lambda^k}{k!} f(\lambda)\, d\lambda \tag{1}$$

Here P(k) is the probability of having k defects on a die[2], where λ is the average defect density in particles per area and $f(\lambda)$ can be thought of as a probability distribution of defect densities from which the current defect density is chosen. Stapper [11] claims that the wafer to wafer variation of the defect density accounts for a large share of the non-Poisson behavior of yield; this may be appropriately modeled by (1) assuming $f(\lambda)$ describes a distribution from which each wafer takes its density.

The role of $f(\lambda)$ is not well defined in the literature. It is used for wafer to wafer variations, lot to lot variations, and even within wafer variations [8]. With each of these phenomena present, using a single mixing distribution to describe them all leads to poor understanding of each. More accurate yield prediction requires a more careful separation of which variation owes itself to between lot, between wafer, and within wafer effects.

Friedman and Albin [5] recognize the clustering effect at the within wafer level. They publish one of 400 wafer maps which clearly displays the effects of clustering. This clustering phenomenon appears where a group of adjacent dice fail on one wafer. This phenomenon is thought to be more likely near the edge of the wafer, possibly owing to handling. The authors use a Neyman Type-A distribution to describe the number of non-functioning chips in a sample, shown in (2).

$$P(k) = \sum_{j=1}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} e^{-j\varphi}\frac{(j\varphi)^k}{k!} \tag{2}$$

This is a compound Poisson process: clusters arrive according to a Poisson process, and each cluster contains a number of particles distributed as a Poisson unknown quantity.

2. If the existence of one or more defects on a die implied a faulty die, then P(0) is an expression for yield.

Stapper [10] proposes the negative binomial as another distribution which describes clustering phenomena. The negative binomial distribution may be derived from (1) when f(λ) is a Gamma distribution. The negative binomial is shown in (3).

$$P(k) = \frac{\Gamma(\alpha+k)}{k!\,\Gamma(\alpha)} \frac{(\lambda/\alpha)^k}{(1+\lambda/\alpha)^{\alpha+k}} \tag{3}$$

Here, α is interpreted as a cluster parameter. As α decreases, the degree of clustering increases. The negative binomial distribution is used extensively in the literature owing to the good empirical fit obtained through its use. Other clustering distributions are discussed by Rogers [9] which are variants of the Poisson process. Ferris-Prabhu [4] proposes an empirical modification of the Poisson process, this time adding a clustering exponent to the Poisson as in (4).

$$Y(A) = e^{-\lambda(A_0)(A/A_0)^{1-b}} \tag{4}$$

Here, b indicates an empirically determined parameter which describes the clustering effect. The A_0 refers to an existing product die size, and A refers to a new product die size. Ferris-Prabhu notes that, as die size increases, a larger value of b will account for more clustering, and hence higher yield.

These yield models seek to describe a complex, multi-variate, stochastic phenomenon with simple parametric models. The result is that these models tend to oversimplify reality. For describing yield it may be acceptable to use some simple clustering distributions to estimate the effects of die size changes and future yields. However, for prescribing remedies to improve yield, these models are not adequate. There is simply not enough causal information in these models.

2.2 Simulator

With this caveat in mind, we developed a wafer cluster simulator based on the center-satellite method discussed in Meyer and Pradhan [7].

The center-satellite method requiring the specification of four distributions regarding the placement of defects on a wafer. The generalized distribution describes the number of clusters. The cluster distribution is a spatial distribution describing where on the wafer the clusters fall. The generalizing distribution describes the number of defects in a cluster. Finally, the dispersion distribution is a spatial distribution describing the shape of the cluster. For example, the Neyman Type-A distribution uses the Poisson distribution for both the generalized and generalizing distributions. The cluster and dispersion distributions are uniform random processes.

Using the S language [1], a cluster simulator has been developed according to this center-satellite method. Using an algorithm from Stapper [12], clusters are created in 2x2 squares. These clusters are constructed point by point by comparing a uniform (0,1) random quantity against the probability density of a two-dimensional Gaussian distribution at the point. Where the density of the distribution is greater than the generated random quantity, a defect point is placed. The number of points in the grid mesh may be specified, and a 100x100 grid has been found to be a compromise between resolution of the cluster pattern and speed of the routine. In addition, after all grid points are placed, each point is perturbed by adding a uniform (-0.01, 0.01) random quantity in each of the x- and y-directions.

These symmetrically generated clusters are then stretched and squeezed along the x- and y-axes by dividing the current x- and y-positions of each of the defect points by uniform (0.2, 2) random quantities. These resulting clusters have an elliptical shape. Stapper [12] further suggests rotating these clusters in the plane, but this feature is not yet available here.

These clusters are then placed on an 8x8 square. The cluster size is smaller than the 8x8 square, so it is possible to determine where the clustering has occurred. Using a routine similar to the cluster generator, this 8x8 square can have background noise; this background of defects is created on the square by comparing a uniform (0, 1) random quantity against a prespecified value. This places points according to a binomial process: each point has a prespecified probability becoming a defect. A true Poisson process would simply place points at random anywhere in the grid without regard to where previous points were generated. For the background process this is computationally feasible; however, for creating the clusters this revision in the algorithm dramatically increased running time. The parameters of the model are *listed by* distribution in Table 1. One additional controllable parameter is the map size, which has been taken as 8x8 throughout; the choice of cluster squares as one-sixteenth of total are is arbitrary, and this ratio is likely one of the most important parameters of this model. Some examples of plots from this algorithm are shown in Figure 1[3].

Table 1: Control Parameters for the Defect Simulator

Distribution	Parameters	Default value
Generalized	Cluster count	Uniform(0,4)
Cluster	Cluster centerpoint	Uniform((0,8), (0,8))
Generalizing	Gaussian distribution constant	0.8
	Cluster grid density	100x100
	Cluster square size	2x2
Dispersion	Gaussian distribution sigma	0.4
	x-translation scalar	Uniform(0.2, 1)
	y-translation scalar	Uniform(0.2, 1)

Once the point pattern is mapped the statistical inference of a descriptive distribution may commence. The quadrat method discussed in Rogers [8] is used. The quadrat method requires the square be partitioned and the number of points in each partition be summed. A routine has been written to partition the square into smaller squares and construct a histogram of the defect count against the frequency of each count.

Given a histogram of frequency counts, the empirical results are tested against the hypothesized distribution. This is accomplished using a χ^2 goodness of fit test. The χ^2 statistic is shown in (5).

$$\chi^2 = \sum_{r=0}^{w} \frac{[f_r - NP_0(r)]^2}{NP_0(r)} \tag{5}$$

Here, f_r refers to the frequency of quadrats with r defects, N is the total number of defects, $P_0(r)$ is the probability of a quadrat containing r defects under the hypothesized distribution, and w+1 is the total number of frequency points. If the data and the hypothesized distribution are close, the bracketed term may be considered noise. If we assume this to be normally distributed, the total expression is the sum of squared normals of mean zero and variance one: this is a χ^2 statistic. It is compared to the χ^2 value at significance level α and w degrees of freedom. If the statistic deviates from the χ^2 distribution, the data is said to deviate from the hypothesized distribution.

The number of frequency points used, or frequency classes, may not be equal to the number of frequency points. That is, Rogers suggests aggregating frequency points such that each frequency class has at least five points. However, this reduces the degrees of freedom of the test, and Rogers notes that the empir-

3. The Figures are located at the end of this chapter.

ical data is most likely to deviate in the tails of the distribution where the frequencies are lowest. Another pitfall is in the specification of the hypothesized distribution; degrees of freedom can be expended by using a non-central χ^2 test which will also estimate the parameters of the hypothesized distribution. To avoid this second pitfall, moment estimators are used to estimate the parameters of the distributions. These moment estimators are shown in Table 2. In the Poisson case, the moment estimator is also the maximum likelihood estimator; for the other cases, finding the maximum likelihood estimator requires an iterative procedure which uses the moment estimators as a starting point.

Quadrat analysis depends on quadrat areas. Stapper [10] notes that the cluster parameter α of the negative binomial distribution depends on the chosen area of the quadrat. As quadrat area tends toward zero, the likelihood of significant clustering decreases. If the quadrat area were small enough, the distribution would tend toward a binomial process in which the quadrat analysis would detect no clustering. Quadrat analysis ignores the spatial relationship of the quadrats. Clusters may be spread over more than one quadrat, but the analysis does not take this into account. Preliminary analysis reveals that as quadrat area decreases, the parameters of the chosen distribution tend to approach a limit and fit the goodness of fit test.

Table 2: Cluster Distributions and Moment Estimators

Distribution	Parameters	Estimate	Estimate
Poisson	λ	$\lambda = m_1$	--
Neyman Type-A	λ, φ	$\lambda = \frac{m_1^2}{m_2 - m_1}$	$\varphi = \frac{m_2 - m_1}{m_1}$
Negative Binomial	λ, α	$\lambda = m_1$	$\alpha = \frac{m_1^2}{m_2 - m_1}$

An example is shown in Figure 2. There are four clusters on the map as well as some background noise. When the grid is partitioned some clusters are split into more than one quadrat. For this example, the cluster parameter α approaches 0.45 as the quadrat area is reduced. Also, the map fits the hypothesized negative binomial better as the quadrat size decreases.

Quadrat analysis is descriptive of the degree to which a wafer map departs from Poisson statistics. However, the analysis is not powerful enough to make prescriptions about how to improve processes. While it is possible to infer a distribution for the data, it may be more difficult to develop a generator which yields a given distribution consistently based on the negative binomial or other descriptive clustering distribution.

3.0 In-situ Monitoring Experiment

While work has been done to extract defect density from final wafer probe yield, we seek an understanding of the defect mechanisms in particular processing equipment. Using a particle counting monitor, it is possible to gain insight into the particle behavior in one machine.

3.1 Experiment

This experiment was performed on the LAM etching machine in the Berkeley Microfab. A laser-driven particle counter was attached to the exhaust system of the etcher such that particles larger than 0.38 microns tripped the beam and were registered in one of five size bins.

A factorial experiment was chosen for two reasons. First, the causes of particulate contamination in the plasma etcher were not clear in advance of the experiment. One purpose of this experiment was to identify these causes. Second, given our ignorance of defect mechanisms in the plasma etcher, we chose not to rule out interaction effects. The two-level factorial design considers interactions.

The 2^3 factorial experiment was performed on the first day, and the results warranted further experimentation. Additional runs were made such that the combined experiment corresponded to a 16 run, 2^{5-1} design with the day as a blocking variable. The full experimental design is shown in Table 3 *in the order of* the runs.:

Table 3: The full experimental design in the order of the runs

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Day
1	-	+	+	+	+
2	+	-	-	-	+
3	+	-	+	+	+
4	-	-	-	+	+
5	+	+	+	-	+
6	+	+	-	-	+
7	-	-	+	-	+
8	-	+	-	+	+
9	+	-	+	-	-
10	+	+	-	-	-
11	-	-	+	+	-
12	-	-	-	-	-
13	+	+	+	+	-
14	+	-	-	+	-
15	-	+	+	-	-
16	-	+	-	+	-

Table 4:

Factor	+ setting	- setting
Etch Rate	5000 Å/minute	3000 Å/minute
Gas Flows	150, 200, 20 sccm	100, 50, 10 sccm
Pre-clean	Yes	No
Wafer set	Old	New
Day	First	Second

The etch rate is measured in Å/sec. It was assumed to follow the equation of May, *et al.*[6], which calculated etch rate dependent upon power, pressure, electrode gap, and the three gas flows CCl_4, He, and O_2. The pressure for this experiment was maintained at 250 mtorr, and the electrode gap was maintained at 1.5 cm throughout, so the etch rate was essentially a surrogate for the power. The gas flow is measured in standard cubic centimeters per second; the high gas flow corresponds to 150 sccm of CCl_4, 200 sccm of He, and 20 sccm of O_2; the low gas flow corresponds to 100 sccm of CCl_4, 50 sccm of He, and 10 sccm of O_2. The pre-clean is an indicator for whether the run follows a standard double cleaning step. The wafer set is an additional factor necessitated by running the experiment on two sets of wafers. While both sets of

wafers had surface polysilicon layers of at least 8000 Å, each set was grown on a different day. The day indicates which of the first or second days the experiments were run.

Each experimental run on the first day consisted of five wafers etched for 90 seconds. The particle counter aggregated time into 15 second intervals; a run began at the start of the first interval following the first wafer starting into the chamber, and a run ended at the end of the interval during which the fifth wafer left the chamber. Hence, some non-processing time was included in the monitored window.

The order was partially randomized in that the pre-clean steps were left in alternating order so that the non-clean steps would have only one previous run before it. According to the technician, the performance of the etcher was sometimes seen to degrade as early as 8-10 wafers into the process.

3.2 Results

The particle counts were much lower than expected. Based on the documentation received with the particle counter, particle counts as high as 50 per minute were expected. However, the highest particle count during any run was seven in a 15 second span. The low counts reduced the effectiveness of time-series methods for analyzing the data. However, for detecting the *relative* particle counts for different runs, the magnitude of the counts was sufficient.

The experimental runs were not all the same length. The data has been normalized in each case to particle count per 60 time intervals, or 15 minutes. The results are shown in Table 4.

Table 5: Experimental Results

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Day	Particle Count	Number of Intervals	Count per 15 min.
1	-	+	+	+	+	46	58	48
2	+	-	-	-	+	73	107	41
3	+	-	+	+	+	73	63	70
4	-	-	-	+	+	106	58	110
5	+	+	+	-	+	80	75	64
6	+	+	-	-	+	76	61	75
7	-	-	+	-	+	94	62	91
8	-	+	-	+	+	88	57	93
9	+	-	+	-	-	41	37	66
10	+	+	-	-	-	43	37	70
11	-	-	+	+	-	35	36	58
12	-	-	-	-	-	41	37	66
13	+	+	+	+	-	45	36	75
14	+	-	-	+	-	60	41	88
15	-	+	+	-	-	84	37	136
16	-	+	-	+	-	40	36	67

Plots of the time-series of each of these runs are shown in Figure 3 in the appendix to this report. Time-series for each day's experiments are shown in Figure 4. Putting the data into canonical form, the effects of each of the variables and any interactions may be calculated. This is shown in Table 5.

Table 6: Experimental Analysis of Effects

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Day	Revised Count	Effect	1-2 Level	3-5 Level
4	-	-	-	-	+	110	76.13	avg.	12345
14	+	-	-	-	-	88	-10.50	1	2345
16	-	+	-	-	-	67	4.75	2	1345
8	+	+	-	-	+	93	4.50	12	345
11	-	-	+	-	-	58	-0.25	3	1245
3	+	-	+	-	+	70	-4.00	13	245
1	-	+	+	-	+	48	4.75	23	145
13	+	+	+	-	-	75	-12.50	45	123
12	-	-	-	+	-	66	0.00	4	1245
2	+	-	-	+	+	41	-21.25	14	235
6	-	+	-	+	+	75	15.50	24	135
10	+	+	-	+	-	70	-11.25	35	124
7	-	-	+	+	+	91	26.50	34	125
9	+	-	+	+	-	66	-12.75	25	134
15	-	+	+	+	-	136	-3.50	15	234
5	+	+	+	+	+	64	-4.25	5	1234

Table 7:

Factor	+ setting	- setting
Etch Rate	5000 Å/minute	3000 Å/minute
Gas Flows	150, 200, 20 sccm	100, 50, 10 sccm
Pre-clean	Yes	No
Wafer set	Old	New
Day	First	Second

A normal probability plot of the effects yielded a nearly straight line, suggesting that the effects here are simply noise. The inconclusiveness of this experiment may be further appreciated by looking at the results of each day separately, as shown in Tables 6-7.

Table 8: The Results of First Day Experiment

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Revised Count	Effect	First Effect
4	-	-	-	-	110	74.0	avg.
8	+	-	-	+	41	-14.0	E
3	-	+	-	+	75	-8.0	G
1	+	+	-	-	93	31.0	EG

Table 8: The Results of First Day Experiment

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Revised Count	Effect	First Effect
2	-	-	+	+	91	-11.5	P
6	+	-	+	-	70	11.5	EP
7	-	+	+	-	48	-16.5	GP
5	+	+	+	+	64	-12.5	EGP

Table 9: The Results of Second Day Experiment

Run	Etch Rate	Gas Flow	Pre-clean	Wafer set	Revised Count	Effect	First Effect
4	-	-	-	+	66	78.3	avg.
8	+	-	-	-	88	-7.0	E
3	-	+	-	-	67	17.5	G
1	+	+	-	+	70	-22.0	EG
2	-	-	+	-	58	11.0	P
6	+	-	+	+	66	-19.5	EP
7	-	+	+	+	136	26.0	GP
5	+	+	+	-	75	-12.5	EGP

What is especially noteworthy here is that significant effects from the first day have reversed signs in the second day. For instance, the interaction of gas flow with etch rate is the most positive effect on first day, but most negative on second day. No strong statement about the effects of etch rate, gas flow, or pre-cleaning may be made on the basis of this experiment.

3.3 Discussion

Given the inconclusiveness of this experiment, a discussion of improvements is in order. Foremost, the three largest effects in this experiment were the two-level interactions between wafer set and the equipment factors. In the future, the wafer set variable should be eliminated by using wafers grown together.

Another improvement would be to run all experiments for the same number of wafers. In addition, it may be advantageous to collect particle counts only when wafers are being processed in the chamber; the time-series is noisy enough that the small peaks in particle count during the time intervals when wafers were being removed from the chamber may be hidden.

Finally, the experimental settings taken from May, *et al.* [6] were at the limits of the range for which their model was validated. New experimental runs limiting the range of the factors might be more conclusive; the plasma was not sustained well during some of the low power, low gas flow runs owing to the deficient amount of He. Also, etch uniformity was not considered in this experiment; a new experimental design should take this into account, since the only factors which might have mattered here depended on the wafer more than the processing equipment.

To further understand how particle counts relate to the condition of the equipment, the machine was passively monitored continuously for one month. As it can be sees in this figure (page 74), there is a definite relationship between maintenance and cleaning events and particle counts. However, particle counts seem to be controlled by additional, uncharacterized effects, as it is evident from the unexplained count reduction midway through the monitoring experiment. Clearly, more analysis is needed.

4.0 Future research

This report outlines two of three parts of an equipment-based approach to yield modeling. The first part is the analysis and statistical inference of wafer maps. The quadrat method has been discussed and found inadequate for analysis beyond measuring departures from Poisson statistics. However, work in understanding spatial processes has been done, and such methods as nearest-neighbor analysis and random Markov fields may be fruitful in providing prescriptive inferences [2].

The second part is the analysis of particle counts on specific equipment types. Currently, the LAM plasma etcher in the Berkeley Microlab has been equipped with a laser sensor. Although the results of the experiment described here were inconclusive, it may be that future experiments will yield better results. In addition, sensors could be installed on other pieces of equipment in the Microlab to learn about particulate contamination from other processes.

The third part is the linking step of empirical wafer mapping. Correlating particle counts to actual wafer maps for different equipment types could lead to the discovery of processing equipment signatures. However, wafer mapping is an expensive, time-consuming process, and has thus far been ignored in this analysis.

Combining these three analyses, a spatial yield model based on equipment characteristics observable through particle counts may be constructed. For process control purposes, specification limits for acceptable particle counts may be set for different equipment types and different processes. For production planning purposes, in-line catastrophic yield predictions based on observable parameters may prove valuable in deciding whether to continue processing lots through the fab. For wafer probe purposes, knowing the likely distribution of faulty chips could be useful for optimization of probe patterns.

5.0 Conclusion

This report documents two efforts to understand semiconductor yield issues. A simulator of wafer defect maps is motivated and developed. Statistical tests are developed to characterize the resulting distribution. An experiment to determine the factors impacting on airborne particle generation is performed for a plasma etching machine. Future efforts in this area are discussed.

6.0 Acknowledgments

The in-situ monitoring experiment was performed at the Berkeley Microlab by Debra Hebert. The particle counter, a PM-150 Particle Trend Monitor, was generously loaned to the Berkeley Microlab by High Yield Technology in Sunnyvale, Calif.

7.0 References

[1] Becker, RA, et al. *The S Language*. Wadsworth & Brooks, (Pacific Cove: 1988).

[2] Cliff, AD and JK Ord. *Spatial Processes: Models & Applications*. Pion Ltd., (London: 1981).

[3] Cunningham, JA. *The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing*. IEEE Transactions on Semiconductor Manufacturing, 3(2), May 1990, 60-71.

[4] Ferris-Prabhu, AV. *A Cluster-Modified Poisson Model for Estimating Defect Density and Yield.* IEEE Transactions on Semiconductor Manufacturing, **3**(2), May 1990, 54-59.

[5] Friedman, DJ and SL Albin. *Clustered Defects in IC Fabrication: Impact on Process Control Charts.* IEEE Transactions on Semiconductor Manufacturing, **4**(1), Feb. 1991, 36-42.

[6] May, GS, J Huang, and CJ Spanos. *Statistical Experimental Design in Plasma Etch Modeling.* IEEE Transactions on Semiconductor Manufacturing, **4**(2), May 1991, 83-98.

[7] Meyer, FJ and DK Pradhan. *Modeling Defect Spatial Distributions.* IEEE Transactions on Computers, **38**(4), April 1989, 538-546.

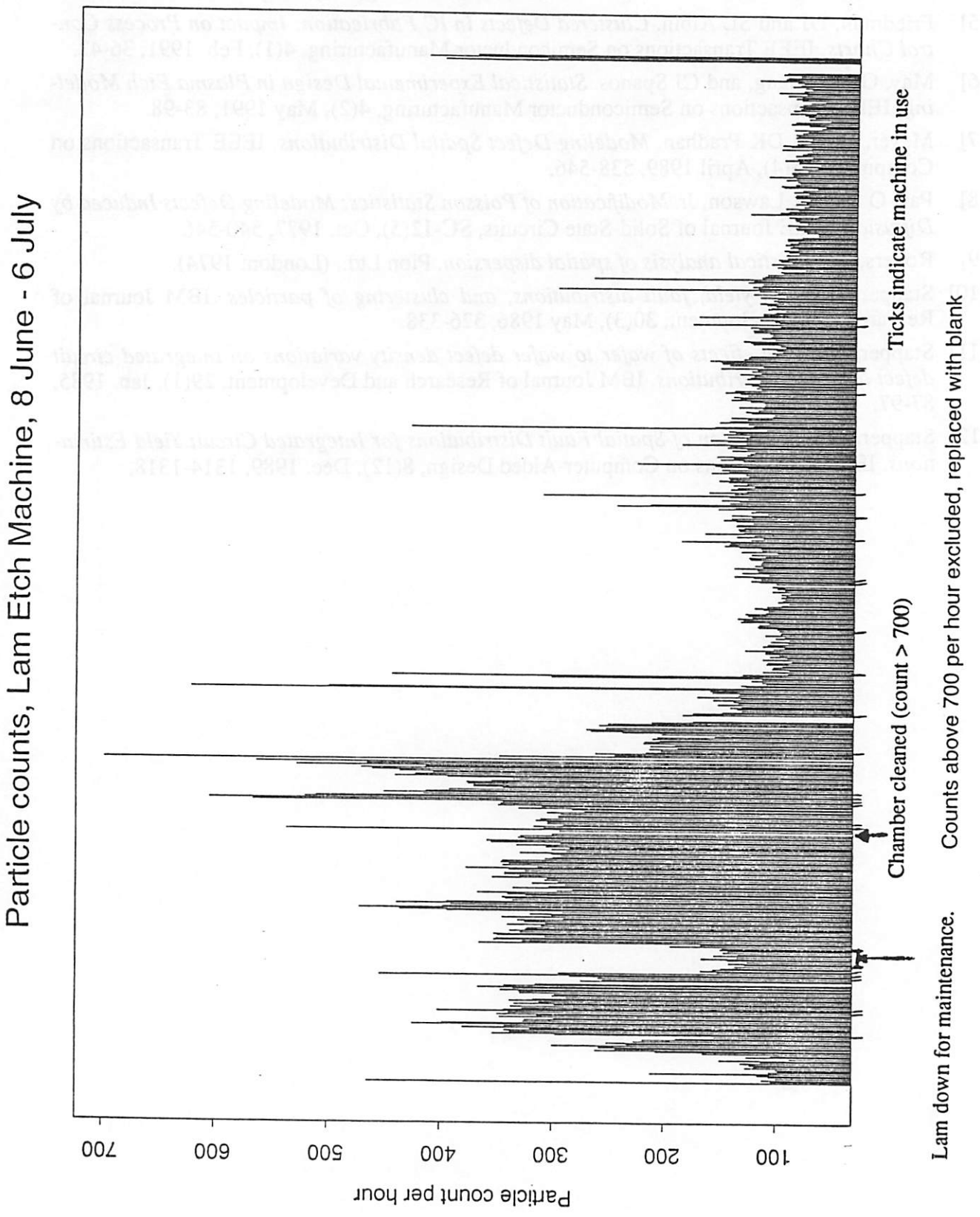
[8] Paz, O and TR Lawson, Jr. *Modification of Poisson Statistics: Modeling Defects Induced by Diffusion.* IEEE Journal of Solid-State Circuits, **SC-12**(5), Oct. 1977, 540-546.

[9] Rogers, A. *Statistical analysis of spatial dispersion.* Pion Ltd., (London: 1974).

[10] Stapper, CH. *On yield, fault distributions, and clustering of particles.* IBM Journal of Research and Development, **30**(3), May 1986, 326-338.

[11] Stapper, CH. *The effects of wafer to wafer defect density variations on integrated circuit defect and fault distributions.* IBM Journal of Research and Development, **29**(1), Jan. 1985, 87-97.

[12] Stapper, CH. *Simulation of Spatial Fault Distributions for Integrated Circuit Yield Estimations.* IEEE Transactions on Computer-Aided Design, **8**(12), Dec. 1989, 1314-1318.


Particle counts, Lam Etch Machine, 8 June - 6 July
Particle count per hour
100
200
300
400
500
600
700
Ticks indicate machine in use
Chamber cleaned (count > 700)
Counts above 700 per hour excluded, replaced with blank
Lam down for maintenance.

Using Stochastic Functions for Modeling Computer-Based Experiments

Zeina Daoud

In this report we present a computer-based experiment used for improving the manufacturability of integrated circuit designs. This experiment consists of simulating in SPICE the performance of an IC design, while varying several of its design parameters. Since this is a computer-based (simulated) experiment, it cannot be analyzed with the classical statistical methods. To cope with this problem we have employed a stochastic function that has been shown to be suitable for experiments whose replication errors are spatially correlated throughout the experimental space.

1.0 Introduction

Optimizing a complex circuit with respect to many design parameters often requires a large number of computer simulations. Modeling a simulator's output would allow designers to explore more fully the design space with fewer computer runs. However, the output of computer-based experiments is deterministically replicated with the same inputs, thus calling for modeling techniques distinct from "traditional" statistical design of experiments.

Sacks et al [1] suggest modeling that deterministic output as a realization of a stochastic process, to account for the lack of random independent error in computer-based experiments. For this project, I propose to explore modeling the output of HSPICE [2] circuit simulator using stochastic functions. The model is designed using the results of a previously studied Taguchi experiment [3]. The resulting model's predictions are compared to the actual simulator's outputs.

Section 2 contains a general description of the methodology and some background on the models used. In section 3, the details of the implementation are discussed. The results are presented in section 4, as well as some modifications to the model needed to accommodate a small experiment.

2.0 Methodology

Circuit performances are determined by several controllable and uncontrollable parameters. The approach followed to model these performances is presented below.

Step 1: Choose a circuit and circuit performances to optimize.

Step 2: Choose variable parameters whose effects on circuit performances we want to explore.

Step 3: Postulate a model for the performances.

A special family of functions is used to model the output of deterministic computer experiments [1]. The output y(x) of a computer experiment is modeled by Y(x) consisting of a regression term and a stochastic term Z(x).

$$Y(x) = \sum_i \beta_i f_i + Z(x) \tag{1}$$

For several reasons discussed in [1] and [4], it has been shown that a constant regression term often gives a simpler and equally accurate representation of the model. So the model adopted for this project is

$$Y(x) = \beta + Z(x) \tag{2}$$

where β is a constant and Z(x) is a stochastic term with a mean of zero and a covariance V(x,w) between Z(x) and Z(w):

$$V(x,w) = Cov(Z(x), Z(w)) = \sigma^2 R(x,w) \tag{3}$$

R(x,w) is the correlation function defined by:

$$R(x,w) = \prod_i exp(-\theta_i \times |x_i - w_i|^{P_i}) \tag{4}$$

The correlation constants θ and p are unknowns that will be estimated in step 5.

Step 4: Design and perform the computer experiment, and gather the data.

Step 5: Use the data to fit the model:

Let $y = (y_1, ..., y_n)$ denote the observed output performances of the experimental runs with n inputs $s_1, ..., s_n$. It can be shown [1] that the best linear predictor of the performance y(x) at an untried input x is:

$$\hat{y}(x) = \hat{\beta} + r'_x R^{-1}(y - \hat{\beta} l) \tag{5}$$

where l is an n x 1 vector of 1's; R is the correlation matrix $R = [R(s_i, s_j)]$ between inputs of the experiment; r_x is the correlation between an arbitrary input x and the inputs s of the experiment ($r_x = [R(x,s_i)]$), and

$$\hat{\beta} = (l' R^{-1} l)^{-1} l' R^{-1} y \tag{6}$$

To compute these predicted values, the correlation parameters θ and p of the correlation matrix must be estimated. Maximum likelihood is used, assuming that y has a normal distribution (see [1] for details). The optimization simplifies down to numerically maximizing

$$-n \log(\hat{\sigma}^2) - \log(det(R)) \tag{7}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}(y - \hat{\beta} l)^T R^{-1} (y - \hat{\beta} l) \tag{8}$$

Step 6: Check the model's accuracy of prediction on *untried* inputs. It is worthy to note from the form of the predicted response $\hat{y}$, that the model's prediction will match exactly the experimental values used to

create the model. Therefore, the fit of the model can only be estimated on untried inputs, or points not included in the original experiment.

These steps describe the general methodology of this problem. The next section describes the details of the circuit and the experiment used.

3.0 Implementation

In a previous project [3], we have studied an adder bit slice circuit and explored the effects of certain parameters on its performance, using Taguchi's Robust Design Method. Partial results of this study are used for modeling the simulator's output using a stochastic function.

The set up of the experiment is briefly reviewed here. Since the performance of the ripple-carry adder is restricted by the speed of the carry-out bit, the performance chosen for optimization is the speed of the carry-out. The parameters of interest are:

- topology: either a transmission gate adder or a full static adder

- width of the carry input and output buffers: set to 8, 10 or 12 microns for n type transistors and 20, 22 and 24 microns for p type transistors.

- length of the carry output buffers: set to 1.8, 2 or 2.2 microns.

Taguchi's L_{18} orthogonal array shown in Appendix I is chosen for the experimental design. Eighteen simulation runs are performed using HSPICE circuit simulator and measurements are gathered for the performance of interest, at those points in the design space. A listing of the collected data is also shown in Appendix I.

In the next step, the stochastic model presented in section 2 is postulated as a representation of the speed of the circuit and the data is used to fit the model. The correlation coefficients θ and p, as well as σ^2 and β are obtained by numerical minimization using the Han-Powell constraint minimization technique.

The performance model obtained is checked at untried input vectors x and compared to the actual simulator's output for these given inputs. The inputs chosen, shown in Appendix II, are a set of parameter combinations not included in the original orthogonal array design.

4.0 Results

Appendix II shows the actual results of HSPICE simulations for a given set of inputs, called the confirmation set, used for checking the model's accuracy. The confirmation set is entirely disjoint from the set of experimental points, as noted above. The model's prediction for every point in the confirmation set is compared to the actual circuit simulator result for that given input.

In an initial attempt to model the speed of the adder bit-slice's carry-out as obtained by HSPICE, the full model described in section 2 was used. Eight spatial correlation factors are needed to study the variation of four parameters: four values of θ's and four values of p's. Han-Powell optimization technique was used to determine values for θ and p by solving the problem described by Eq. (7). Table 4 in Appendix III shows the optimized values of the spatial correlation factors, starting from initial guesses of 1.0 and 1.1. Each set of optimized values of θ and p defines a unique model of the speed of the circuit. As noted above, the model fits exactly on the experimental points and must be validated on the confirmation set. The mod-

el's predictions are tabulated and the graphs of the model's predictions versus the actual HSPICE results are shown in Appendix III. The solid line y = x on the graphs represents an ideal model where the prediction would exactly match the actual response. It is clear from the graphs that the model developed is far from ideal, and the discrepancy between the predictions and the expected results is large. We believe that the lack of fit of the model is due to the fact that the problem at hand is under-determined. The small experiment used (few data points) may be insufficient to determine eight values of the spatial correlation factors that define the stochastic model.

Motivated by this speculation, a slight modification to the original stochastic model is made to accommodate a smaller experiment. Instead of solving for eight coefficients, let the four values of p constant and optimize for the values of θ only. The reason for locking the values of p is that, due to the form of the correlation matrix R, the p coefficients are most sensitive to the difference in the orders of magnitude of the input parameters. A value of 1.0 is chosen for the p_i's. The values of θ_i are still obtained as before by numerical optimization, with initial guesses of 1.0 or 1.1. The constant values of p's and the optimized values of θ define a stochastic model for the speed. The optimized values of θ and the modified model's predictions for the confirmation set are shown in Appendix IV. The graphs of the predicted versus expected values of the delay (for the confirmation set) display a noticeable improvement of the model, as the points lie close to the y = x diagonal. This result confirms the idea that the original problem is under-determined, and that a way to adapt the method to a small experiment is to reduce the number of unknown spatial correlation factors that must be determined.

5.0 Conclusion

In this report, some background was presented for modeling the output of computer-based experiments using stochastic functions. Stochastic functions are used to account for the lack of random independent error in this type of experiments. An application of this technique to circuit optimization was shown. Modeling the output of a circuit simulator allows the designer to explore a larger number of variable parameters at different levels, using fewer computer runs.

For this stochastic modeling method, the number of unknowns is twice the dimension of the input space. If a small number of data points are used to fit the model, the model must be altered to avoid an under-determinate problem. One such modification to the model was discussed that lead to major improvements in the model's prediction capability.

6.0 References

[1] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, "Design and analysis of computer experiments," *Statist. Sci*, Vol. 4, pp. 409-435, November 1989.

[2] Meta-Software, Inc., "HSPICE Users Manual H9001," 1990.

[3] Z. Daoud, C. J. Spanos, "Application of the Robust Design Technique to IC design for manufacturability," *IC Design for Manufacturability I*, Memorandum No. UCB/ERL M92/17 February 1991.

[4] M. Bernardo, R. Buck, L. Liu, W. Nazaret, J. Sacks, W. Welch, "Integrated Circuit Design Optimization Using a Sequential Strategy," *IEEE Transactions on Computer-Aided Design*, Vol. 11, No. 3, March 1992.

7.0 Appendix I

The matrix experiment corresponding to four input parameters, one at two levels and three at three levels utilizes the L_{18} orthogonal array, shown in the table 2 below. Also presented are the results of HSPICE circuit simulations for the carry-out delay.

Table 1: Definition of Parameter Levels

factors	level 1	level 2	level 3
topology	trans. gate	full static	
width_out	W_0	$W_0 + i$	$W_0 + 2i$
length_out	L_0	$L_0 - 0.2$	$L_0 - 0.2$
width_in	W_0	$W_0 + i$	$W_0 + 2i$

Table 2: Experiment Matrix and HSPICE Delay Results

trial	topology	width_out	length_out	width_in	delay (ns)
1	1	1	1	1	2.366
2	1	1	2	2	2.152
3	1	1	3	3	2.297
4	1	2	1	1	2.203
5	1	2	2	2	1.997
6	1	2	3	3	2.126
7	1	3	1	2	1.992
8	1	3	2	3	1.822
9	1	3	3	1	2.193
10	2	1	1	3	3.296
11	2	1	2	1	3.315
12	2	1	3	2	3.517
13	2	2	1	2	2.944
14	2	2	2	3	2.733
15	2	2	3	1	3.226
16	2	3	1	3	2.630
17	2	3	2	1	2.700
18	2	3	3	2	2.838

8.0 Appendix II

The confirmation set is a set of inputs different from the ones in the experiment matrix, used to check the accuracy of the model. Table 3 shows the confirmation set and the actual HSPICE results of the delay as simulated under these conditions.

Table 3: Confirmation Set and HPICE Delay Results

conf. run	topology	width_out	length_out	width_in	delay (ns)
19	1	1	1	2	2.261
20	1	1	2	3	2.079
21	1	1	3	2	2.371
22	1	2	1	3	2.024
23	1	2	2	1	2.101
24	1	2	3	2	2.200
25	1	3	1	3	1.916
26	1	3	2	1	1.996
27	1	3	3	3	2.007
28	2	1	1	1	3.480
29	2	1	2	3	3.137
30	2	1	3	3	3.453
31	2	2	1	3	2.875
32	2	2	2	2	2.798
33	2	2	3	3	3.017
34	2	3	1	1	2.844
35	2	3	2	2	2.570
36	2	3	3	3	2.762

9.0 Appendix III

The numerical optimization used to solve for eight correlation coefficient values, depends on initial guesses for these values. Each set of values defines a unique model whose prediction on the confirmation set is shown in Table 5.

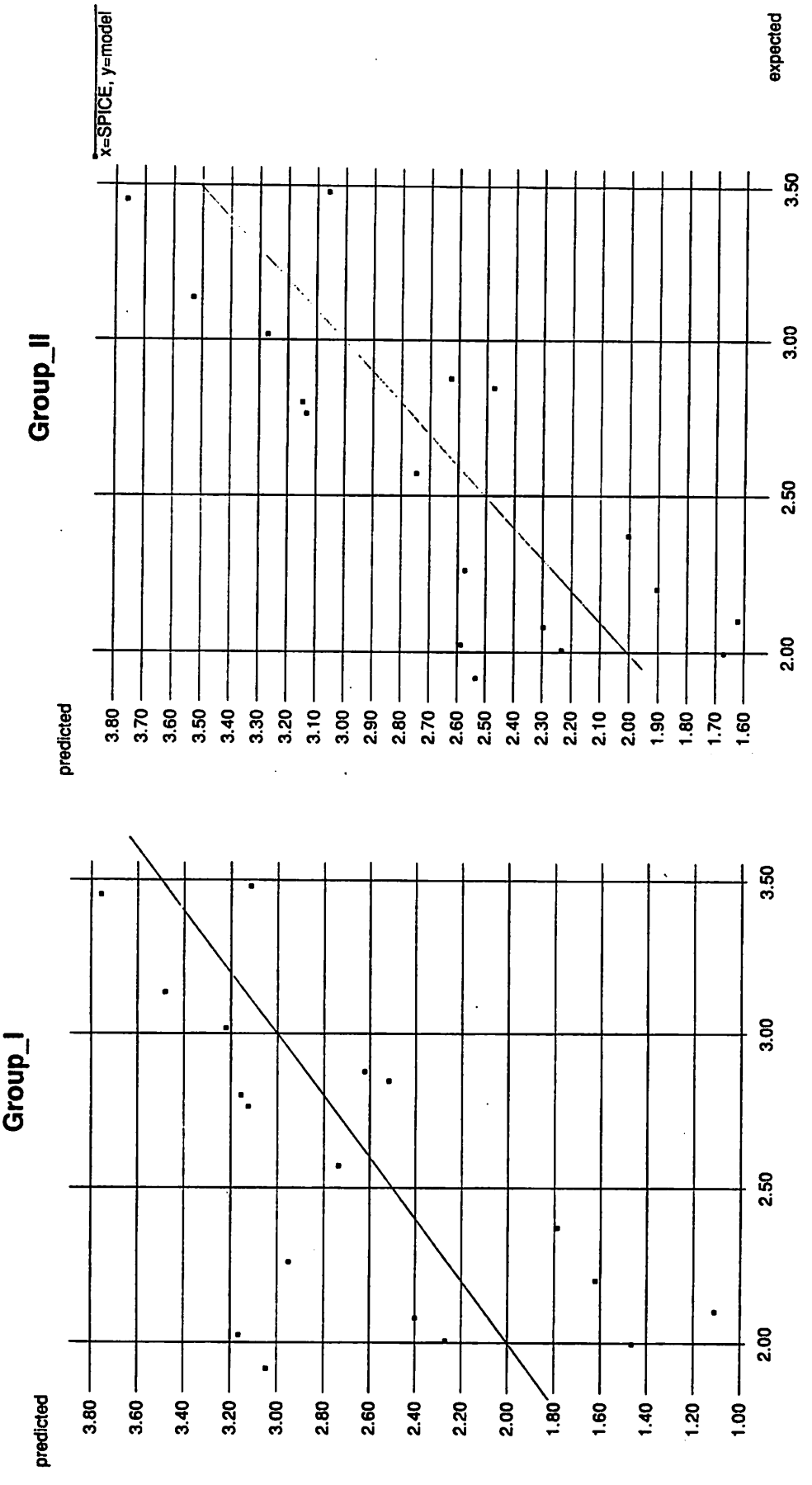
Table 4: Optimized Values of the Correlation Coefficients

correlation coefficients	initial guesses: 1.0 (Group I)	initial guesses: 1.1 (Group II)
p_1	1.00	1.100
p_2	0.455	0.252
p_3	2.849	2.466
p_4	0.099	0.038
θ_1	2.574	1.685
θ_2	0.099	0.095
θ_3	1.338	0.708
θ_4	0.099	0.051

Table 5: Models Predictions on the Confirmation set (in ns)

conf. run	actual delay	Group I	Group II
19	2.261	2.951	2.573
20	2.079	2.401	2.296
21	2.371	1.786	2.003
22	2.024	3.163	2.587
23	2.101	1.110	1.623
24	2.200	1.622	1.904
25	1.916	3.046	2.5334
26	1.996	1.464	1.671
27	2.007	2.267	2.233
28	3.480	3.115	3.059
29	3.137	3.480	3.529
30	3.453	3.757	3.761
31	2.875	2.622	2.627
32	2.798	3.156	3.147
33	3.017	3.221	3.270
34	2.844	2.516	2.473
35	2.570	2.735	2.746
36	2.762	3.123	3.133

The same information is displayed graphically in the next two pages: predicted versus expected values of delay for each group of optimized correlation coefficients, with the y = x diagonal for reference.


Group_II
x=SPICE, y=model
predicted
expected
3.80
3.70
3.60
3.50
3.40
3.30
3.20
3.10
3.00
2.90
2.80
2.70
2.60
2.50
2.40
2.30
2.20
2.10
2.00
1.90
1.80
1.70
1.60
2.00
2.50
3.00
3.50
Group_I
predicted
3.80
3.60
3.40
3.20
3.00
2.80
2.60
2.40
2.20
2.00
1.80
1.60
1.40
1.20
1.00
2.00
2.50
3.00
3.50

10.0 Appendix IV

The model is altered to accommodate a small experiment. Values of pi's are set to 1.0 and θ's are found by numerical optimization.

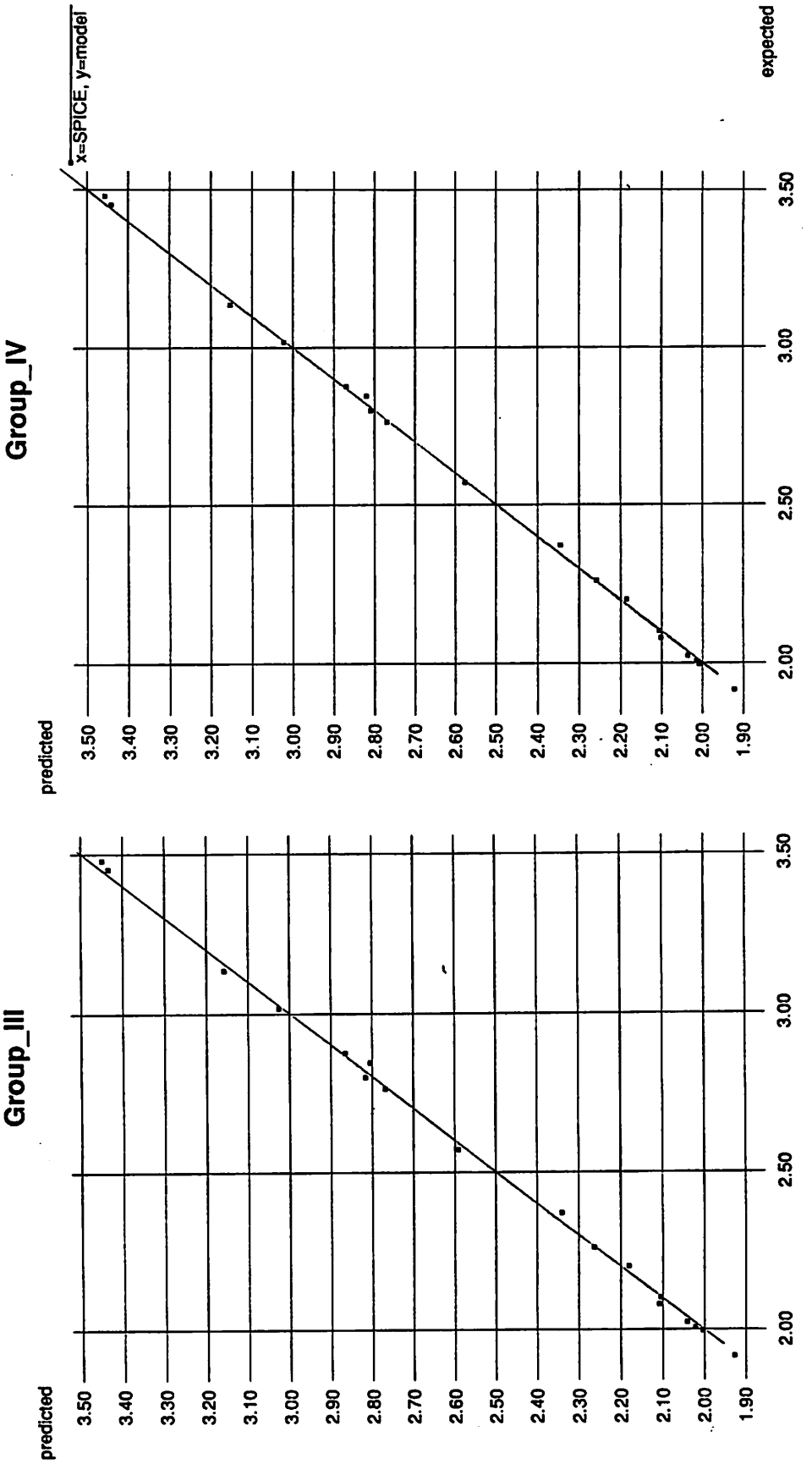
Table 6: Optimized Values of Θ

optimized	initial guesses: 1.0 (Group III)	initial guesses: 1.1 (Group IV)
θ1	3.083	3.265
θ2	0.058	0.029
θ3	0.347	0.173
θ4	0.028	0.017

Table 7: Modified Models Predictions on the Confirmation Set (in ns)

conf. run	actual delay	Group III	Group IV
19	2.261	2.263	2.259
20	2.079	2.108	2.102
21	2.371	2.340	2.345
22	2.024	2.041	2.036
23	2.101	2.104	2.105
24	2.200	2.181	2.185
25	1.916	1.927	1.922
26	1.996	2.003	2.008
27	2.007	2.022	2.015
28	3.480	3.448	3.457
29	3.137	3.157	3.154
30	3.453	3.434	3.441
31	2.875	2.865	2.869
32	2.798	2.815	2.809
33	3.017	3.025	3.022
34	2.844	2.804	2.820
35	2.570	2.590	2.576
36	2.762	2.768	2.769

The same information is displayed graphically in the next two pages: predicted versus expected values of delay for each group of optimized correlation coefficients, with the y = x diagonal for reference.


Group_IV
x=SPICE, y=model
predicted
expected
3.50
3.40
3.30
3.20
3.10
3.00
2.90
2.80
2.70
2.60
2.50
2.40
2.30
2.20
2.10
2.00
1.90
2.00
2.50
3.00
3.50
Group_III
predicted
3.50
3.40
3.30
3.20
3.10
3.00
2.90
2.80
2.70
2.60
2.50
2.40
2.30
2.20
2.10
2.00
1.90
2.00
2.50
3.00
3.50

Extraction of Bleach Parameters from Peak Reflectivity Measurements

David M. Newmark

This paper describes the relationship between the peak reflectance measurements used to characterize positive photoresist, the photoactive compound (PAC) concentration and photoresist bleaching parameters which describe physical properties of the resist. The validity of the model for PAC concentration is explored by developing an empirical model for the fraction of PAC remaining in the resist as a function of wafer track settings. Modified photoresist bleaching parameters are used as inputs to SAMPLE, an optical lithography simulation program, to predict the output reflectance as a function of input reflectance and thickness. The predicted output peak reflectance is compared to experimental measurements. The difference between predicted and measured values is attributed to lack of knowledge regarding the change of absorption of the non-bleachable component of the photoresist as a function of wavelength.

1.0 Introduction

Theoretical models for resist exposure and development were first introduced by Dill in 1975 [1]. They provide a convenient way to describe the photoresist exposure and development processes. One problem with these models is that they require careful extraction of parameters under circumstances which may be somewhat different from the actual processing conditions of the wafers. In addition, processing conditions continually drift over time. Thus, it is difficult to use theoretical models to monitor equipment in a manufacturing environment.

Therefore, manufacturing engineers tend to rely on empirical models obtained using factorial experiments. Although such models are accurate, they offer no insight into the process, and engineers are often reluctant to accept empirical models based on measurements which do not have a solid theoretical base. In addition to convincing process engineers of the validity of specific measurements, tieing theoretical modeling to measurements made during production has the added benefit of allowing in-situ monitoring of a process with theoretical models by coupling manufacturing models to simulation tools.

Several methods have been introduced to monitor photoresist in a manufacturing environment. One potential technique is to directly measure, using the appropriate wavelengths, the absorbance of the photoresist as described by Watts [2]. An alternate technique, which uses peak reflectance to infer absorbance, was introduced by Ling and Spanos [3]. The problem with these methods is that the relationship between the absorbance or peak reflectivity measurements and the physical parameters of the resist is not well understood. The goal of this project is to investigate the relationship of PAC concentration and Dill's bleaching parameters to peak reflectance. This relationship will be tested by using a model to predict output reflectance through SAMPLE.

This paper first describes Dill's positive photoresist bleaching model and the use of peak reflectance as a means to monitor photoresist. Based on this background information, the link between peak reflectance and the fraction of PAC remaining in the photoresist after the resist spin-coat and bake process is explained. This result leads quickly to a method for calculating new bleach parameters. These parameters can then be used by SAMPLE to predict the post-exposure peak reflectance. Finally, a model for the frac-

tion of PAC remaining as a function of wafer track parameters is developed along with a comparison between the output peak reflectance predicted by SAMPLE and experimental output peak reflectance measurements.

2.0 Methodology

2.1 Photoresist Bleaching Model

In 1975, Dill published a classic paper in which he presented a model for photoresist bleaching and development [1]. This model has been incorporated into a variety of simulators, such as SAMPLE [4] and PROLITH [5]. The basic model has been extended, for example, to include additional effects such as post-exposure bake [6].

Dill's model is formed from a physical basis, but the actual parameters of the model are substantially different from the photochemical constants used by manufacturers of photoresist. The physical basis comes from the assumed relationship of the parameters in the model to the physical process of absorption of light by the photoactive inhibitor in which the photoactive compound is destroyed under exposure to light. In Dill's model, the process is described by three parameters: "A", an exposure absorption term; "B", an exposure-independent term; and "C", an optical sensitivity term.

Traditionally, the A, B, and C parameters are extracted by measuring the exposure time versus transmittance curve for the resist on a quartz substrate. A typical curve is shown in Fig. 1 for AZ1350J resist. As discussed by Dill, Equations (1), (2), and (3) are used to find A, B, and C, respectively. T(0) is the transmission at exposure time equal to 0. $T(\infty)$ is the transmission of the fully bleached resist, and d is the thickness of the resist. The A, B, and C parameters are wavelength dependent so the transmission versus exposure time curves must be measured at the exposure wavelength. However, the results obtained for the model parameters can be used for photoresist of any thickness.

$$A = \frac{1}{d}\left(\log\left[\frac{T(\infty)}{T(0)}\right]\right) \tag{1}$$

$$B = -\frac{1}{d}\log T(\infty) \tag{2}$$

$$C = \frac{A+B}{AI_0 T(0)\,[1-T(0)]}\frac{dT(0)}{dt} \tag{3}$$

Optical Transmittance Curve

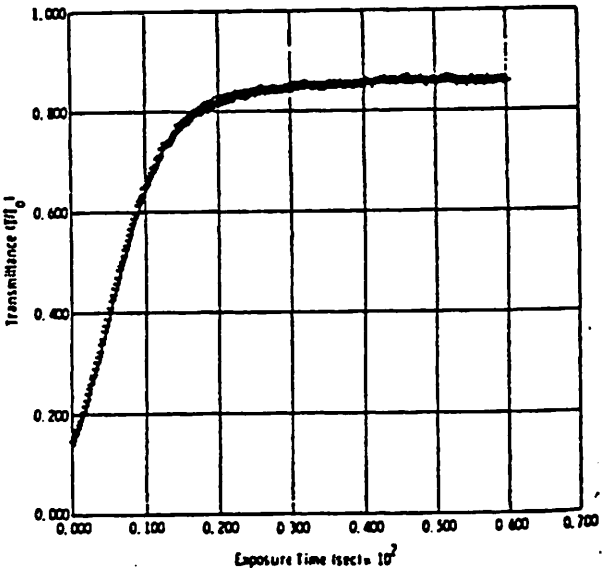


FIGURE 1. Optical transmittance of a 2.2um film of AZ1350J photoresist as a function of exposure time.

Once the A, B, and C parameters are extracted for a given photoresist, the working equations, (4) and (5), are used to find the fraction of inhibitor concentration remaining after the exposure:

$$\frac{\partial}{\partial x} I(x,t) = -I(x,t)\,[AM(x,t) + B] \tag{4}$$

$$\frac{\partial}{\partial t} M(x,t) = -I(x,t) M(x,t)\, C \tag{5}$$

where $M(x,t)$ is the fraction of inhibitor remaining at depth x after exposure time t and $I(x,t)$ is the light intensity at depth x in the film after exposure time t.

2.2 Reflectance as a Means to Monitor Photoresist

For process development, Dill's model provides an excellent way to characterize new resist processes; however, for special manufacturing techniques, such as feed-forward control, running quartz wafers to monitor the photoresist is prohibitively expensive and impractical. Thus, techniques which measure absorbance [2] or peak reflectance [3] have been used to monitor the photoresist.

The peak reflectance measurement is particularly useful for several reasons. First, the authors in [3] assert that it is directly proportional to PAC concentration. They show that peak reflectance is directly proportional to absorbance, and since

$$\alpha = AM + B \tag{6}$$

peak reflectance must be proportional to M or the PAC concentration. A and B are material constants of the resist. Second, due to the interaction of several material properties which depend on wavelength, the peak reflectance is almost constant from about 380nm to 430nm for KTI 820 resist. The combination of these properties implies that reflectance is proportional to M for these wavelengths.

Unfortunately, this assertion is not quite accurate since A and B depend on wavelength as illustrated in Fig. 2. The relationship is further emphasized by rewriting (6)

$$\alpha = A(\lambda) M(x) + B(\lambda) \tag{7}$$

In other words, although peak reflectance is almost constant over the measurement wavelengths and directly proportional to the absorbance, it is not directly proportional to the PAC concentration since A(λ) changes with wavelength. This presents a serious problem, since peak reflectance is, by definition, measured at wavelengths that shift in order to track thickness variations.

2.3 Calculating the Fraction of PAC Remaining and Dill's A Parameter

Based on these observations, a method for removing the dependence of A(λ) is developed in order to find a more accurate measure of the fraction of PAC remaining in the photoresist. The key idea behind this technique is that the fraction of PAC remaining in the resist is a constant regardless of the measurement wavelength; thus, it is possible to extract a relative measure of the remaining PAC using tabulated values for A(λ) and B(λ). From this information, the experimental value for A at the exposure wavelength of 365nm is calculated. *It is important to note at this point that this technique assumes B and C can be measured at the exposure wavelength and do not vary with resist process parameters.* Note that Mack has already established that *both* A and B vary with oven prebake temperature and time. However, for most resist systems, the fraction of the absorption due to A is much greater than that due to B before exposure, so ignoring variations in B is reasonable for most resists near the nominal exposure wavelength.

The method for calculating A based on input peak reflectance and thickness is illustrated in Fig. 3. In the first step, the wavelength is varied from 380nm to 430nm to find the wavelength which gives maximum reflection for the given thickness of photoresist spun on 980Å of oxide on silicon. For this calculation, the index of refraction of the photoresist has no absorption component since maximum (peak) reflection is mainly determined by the real part of the index. The change in the index of refraction of silicon is accommodated by using a table lookup function to find the index for a given wavelength. This calculated reflection is normalized to the reflection from a bare silicon wafer to mimic the Nanospec reflectance measurement. The wavelength for maximum reflection will be referred to as the measurement wavelength.[4]

In the second step, the complex index of refraction is increased until the reflection from the silicon surface is equal to the measured reflection. At this point, we know the index of refraction of the photoresist, and use

$$\alpha' = \frac{4\pi k}{\lambda} \tag{8}$$

to find the absorbance of the resist at the measurement wavelength. Since Mack has extracted the A and B parameters from 300nm to 500nm for KTI 820 resist [5] (see Fig. 2), the absorbance at the measurement wavelength can be extracted. Note that M is by definition equal to 1 before the exposure.

$$\alpha' = A(\lambda) f_A + B(\lambda) \tag{9}$$

Solve for f_A to obtain (10).

4. REFLOP, written by Prof. Oldham, is used for the reflectance calculations.

$$f_A = \frac{\alpha' - B(\lambda)}{A(\lambda)} \tag{10}$$

Physically, f_A represents the fraction of photoactive compound remaining after the resist spin-coat and bake process. Since the total amount of photoactive compound is constant, the fraction of PAC destroyed applies regardless of the wavelength. Thus, we can use

$$A' = A(365nm) f_A \tag{11}$$

to find A', a new value for A which now incorporates the knowledge about the reflectance of the wafer. The remaining bleach parameters, B and C, are assumed to be independent of the resist processing conditions, and their value is measured at the exposure wavelength of 365nm.

2.4 Calculating post-exposure Peak Reflectance

To test the theory that reflectance can also be used to monitor the PAC concentration at the output of the stepper by measuring output reflectance, the extracted A parameter in conjunction with constant values for B and C are used to model the post-exposure peak reflectance. Fig. 4 illustrates the procedure used to obtain the post-exposure peak reflectance. Initially, SAMPLE is run using the A parameter derived above in conjunction with the assumed values for B and C. SAMPLE returns the M(x) matrix which gives the fraction of PAC remaining in the resist after exposure and post-exposure bake. M(x) is specified at approximately 200 locations, or layers, in the resist. At each layer, the absorbance of the bleached resist is calculated using the formula

$$\alpha = A(\lambda) f_A M(x) + B(\lambda) \tag{12}$$

and the k value for each layer is determined from,

$$k = \frac{\lambda \alpha}{4\pi} \tag{13}$$

The refractive index for the photoresist is then simply,

$$n = n - ik \tag{14}$$

where n= 1.68 for KTI 820. The thickness of each layer is constant, and the thickness and refractive index of each layer provides sufficient information to calculate the peak reflectance from all 200 dielectric layers on silicon.

2.5 Implementation

A C program has been written to implement the methods outlined above for calculating the fraction of PAC remaining before exposure, Dill's A parameter, and the output reflectance. The program essentially implements the block diagrams shown in Figs. 3 and 4. The initial thickness, reflectance, and dose for an arbitrary number of wafers are specified in an input file. The program runs REFLOP to calculate the reflectance due to the dielectric stack and SAMPLE to find the M matrix after exposure. The PAC fraction, Dill's A parameter, and the output reflectance are written to the standard output after calculations for all wafers are completed.

3.0 Results

To test the model for the fraction of photoactive compound remaining before exposure, data from a factorial experiment on the wafer track is used to find the PAC fraction and model it based on the wafer

track settings. The SAMPLE reflectance model is tested by comparing the post-exposure peak reflectance predicted by the model with the results of a factorial experiment on the EATON wafer track [7].

3.1 Model for PAC Fraction

The fraction of PAC remaining in the photoresist before exposure can be modeled as a function of the variables on the wafer track. The data, the model, and the associated residual plots are shown in Appendix A. Although the terms of the model appear significant at the <1% level, the R^2 for the model is only 0.64. The residuals appear to be IIND. Since the fit of the model is somewhat questionable, a histogram of the data is plotted in Fig. 5. It shows that the fraction of PAC remaining is nearly Gaussian. The mean is in the center of the distribution and 67% of the measurements are located within +-1σ of the mean. This indicates that the variation in the PAC concentration may be purely random. The histograms for the measured values of thickness and reflectance are shown for reference in Figs. 6 and 7. These data do not appear to be Gaussian.

In theory, the fraction of PAC remaining should be related to the bake temperature and time, since Mack has shown theoretically and experimentally that A and B vary logarithmically with these variables for an oven prebake [6]. Transforming the data by looking at the logarithm or exponential of the PAC fraction does not improve the model. The problem in this case may be that the hot plate bake has a negligible affect on the photoactive compound concentration. The range of the factorial would have to be expanded to distinguish this effect from the experimental noise.

3.2 Modeling Post-Exposure Peak Reflectance

Assuming the PAC fraction remaining in the resist is more than just a measure of noise, the output reflectance predicted by SAMPLE, based on the modified values of Dill's A parameter, is compared with experimental reflectance measurements. The residuals plotted versus measurement wavelength and run number are shown in Fig. 8 and Fig. 9 respectively. For comparison, a plot of the reflectance residuals for a model in which A, B, and C are constant is shown in Fig. 10 and Fig. 11. Although, the mean square of the residuals can be minimized by modifying the dose, the variance will still be much greater than the experimental error of the reflectance measurement. The consideration of the residual plots versus wavelength led to an exploration of the physical cause for the dependence of SAMPLE's prediction error with wavelength.

After reviewing the equations used to calculate the post-exposure peak reflectance, the dependence of output peak reflectance on A(365nm) and B(λ) was examined. A(365nm) affects the output reflectance through a change in M(x) while M(x) and B(λ) affect the output reflectance directly through a change in the absorbance as shown in Equation (7). More precisely, modifying A(365nm) +-20% from its nominal value of 1.017 causes a 2% change in the output reflectance. This effect is demonstrated more clearly for the experimental data in the plot of Fig. 12. The difference in predicted reflectance with A equal to 1.017 and A varying from 0.7 to 1.0 according to Equation (11) is plotted versus run number. The maximum change in reflection is 2.5%, which confirms that output peak reflectance does not change significantly with large shifts in the initial concentration of photoactive compound. (This change in A should have a much larger effect on CD.) On the other hand, B(λ) has a significant affect on the output peak reflectance. For example, if B changes from 0.041 to 0.088 for a given resist thickness and reflectance, the output reflectance changes by 5%. As long as B(λ) is known, the changes are not a problem since tabulated values for B can be used. Unfortunately, B is not well characterized since the transmission calculated from values of B used in PROLITH [5] corresponds to the ideal (no absorption) thin film transmission curve as shown in Fig. 13. Therefore, the wavelength dependence on the residuals can be explained by an uncharacterized change in B(λ).

4.0 Conclusion

Peak reflectance is shown to be related to the fraction of photoactive compound present in a resist. The evaluation of the theoretical relationship has been automated in order to calculate the fraction of PAC remaining in the photoresist after the resist spin-coat and bake process. Based on the transformation of

peak reflectance to PAC concentration, a model is derived which relates the spin speed, spin time, bake temperature, and bake time to the fraction of PAC remaining.

The fraction of PAC remaining is used to calculate a new A parameter for the photoresist which is then utilized in SAMPLE to predict output reflectance. An analysis of the residuals of the predicted output reflectance compared to the measured output reflectance led to the discovery that an uncharacterized change in $B(\lambda)$ could account for about 5% of the difference between measured and predicted values of output reflectance. In addition to a wavelength dependence, B also probably varies with resist processing. Furthermore, the relatively subtle change in output reflectance with A suggests that reflectance will catch process deviations which significantly affect PAC concentration, such as variations in dose and thickness. The changes due to altered initial resist properties may only be evident after the development of the exposed photoresist.

5.0 References

[1] Frederick H. Dill, William P. Hornberger, Peter S. Hauge, and Jane M. Shaw, "Characterization of Positive Photoresist," *IEEE Transactions on Electron Devices*, Vol. ED-22, No. 7, July 1975.

[2] Mike Watts, Thiloma Perera, Bob Ozarski, Dave Meyers, Raul Tam, "Photoresist as its own Process Monitor," *Solid State Technology*, July, 1980, pp. 59-65.

[3] Zhi-min Ling and Costas J. Spanos, "A Novel Approach for Film Reflectance Measurement and its Application for the Control of a Photolithography Workcell," *ECS Conference Proceedings*, Washington, Sept. 1990.

[4] W. G. Oldham, et al, "A General Simulator for VLSI Lithography and Etching Processes: Part I - Application to Projection Lithography," *Trans. Electron Dev.*, Vol. ED-26, No. 4, April, 1979, pp. 717-722.

[5] Chris A. Mack, "PROLITH: a comprehensive optical lithography model," *SPIE Vol. 538, Optical Microlithography IV* (1985), pp. 207-220.

[6] Chris A. Mack, "Modeling the Effects of Prebake on Positive Resist Processing," *Kodak Microelectronics Seminar Interface*, 1985.

[7] Sovarong Leang and Costas J. Spanos, "Statistically Based Feedback Control of Photoresist Applications," *ASMC Proceedings*, Boston, Oct. 21, 1991.

[1]

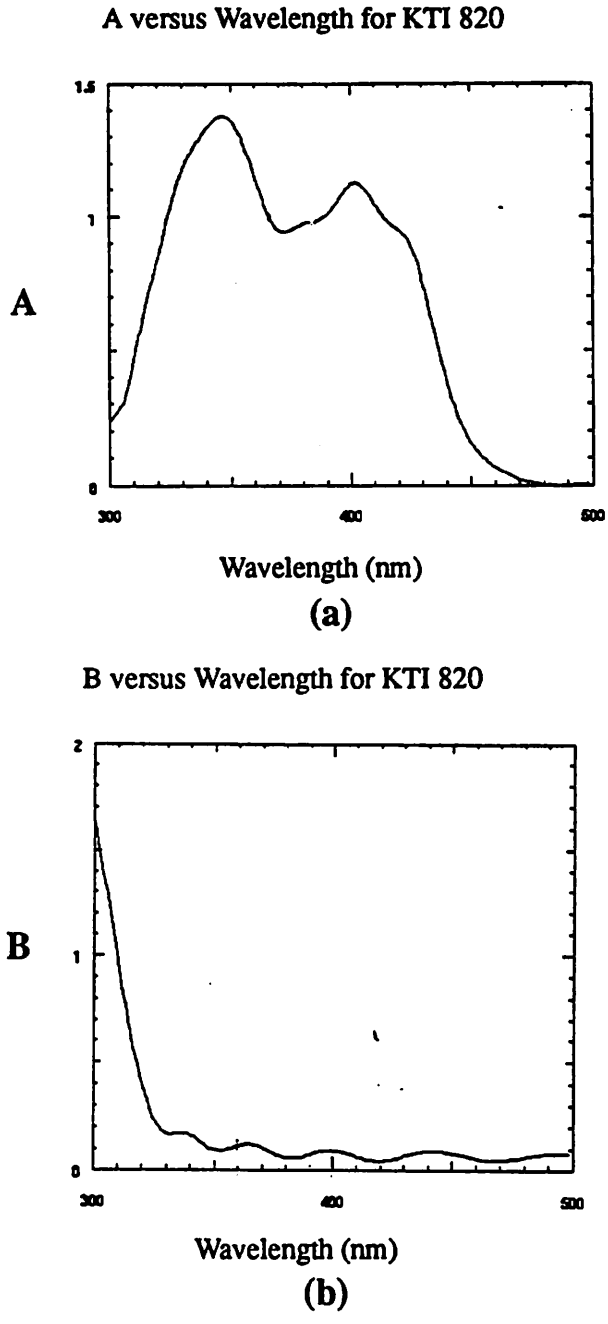


FIGURE 2. (a) Bleachable photoresist constant, A, versus wavelength for KTI 820 photoresist. (b) Non-bleachable photoresist constant, B, versus wavelength for KTI 820. Data obtained from PROLITH [5].

Calculation of Fraction of PAC Remaining

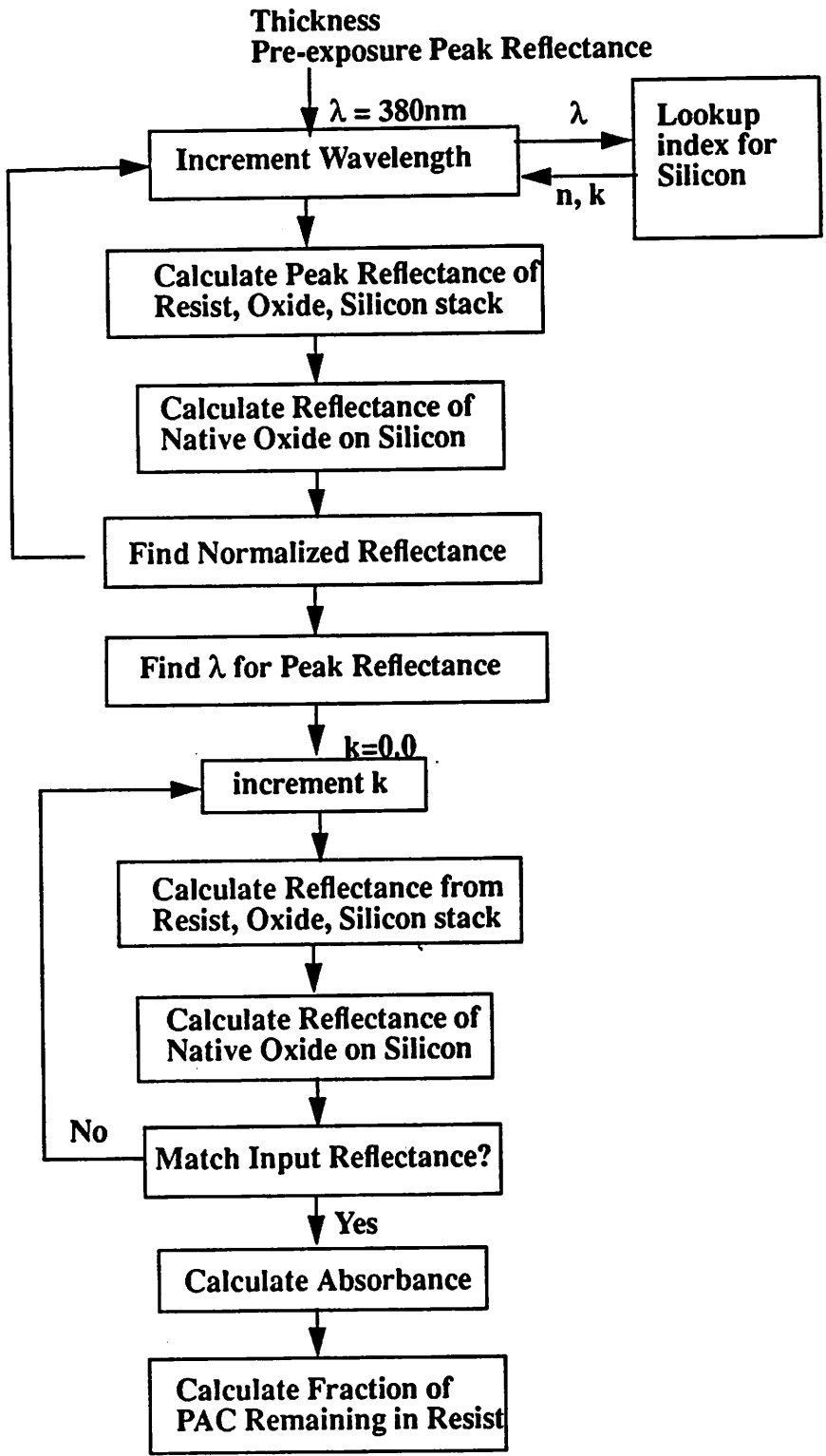


FIGURE 3. Block diagram which illustrates the calculation of the fraction of PAC remaining in resist.

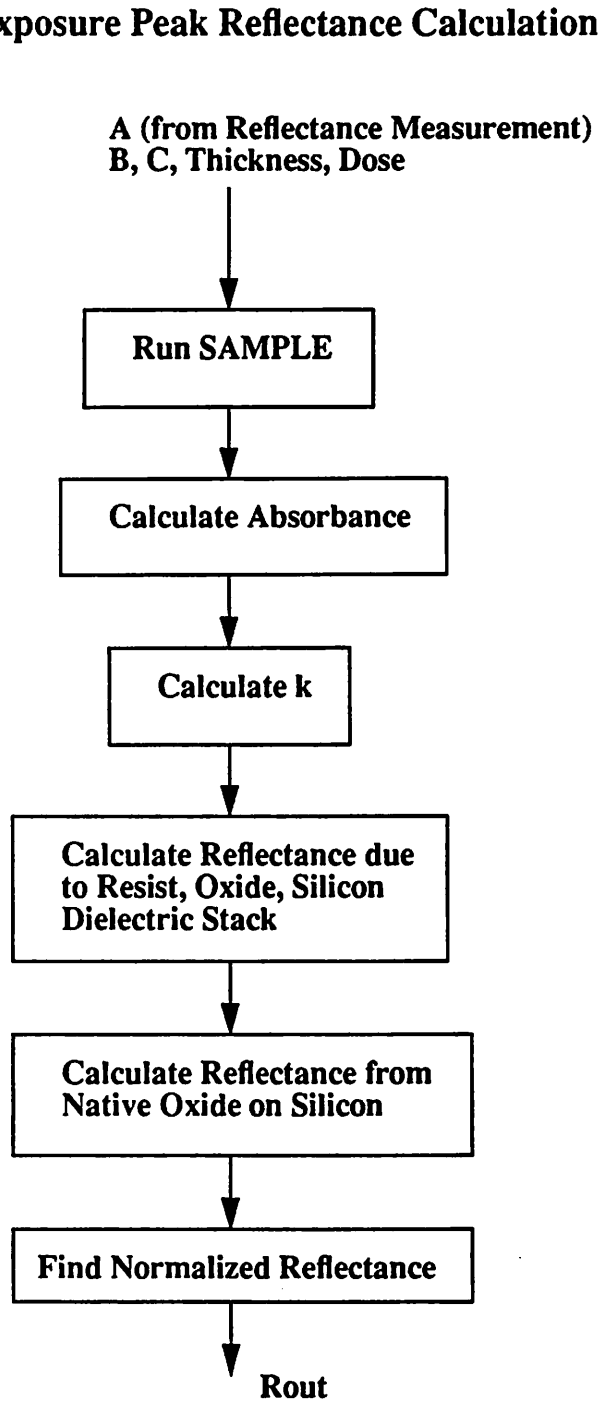
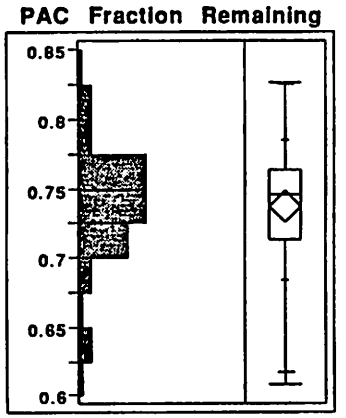


FIGURE 4. Calculation of output reflectance using SAMPLE.



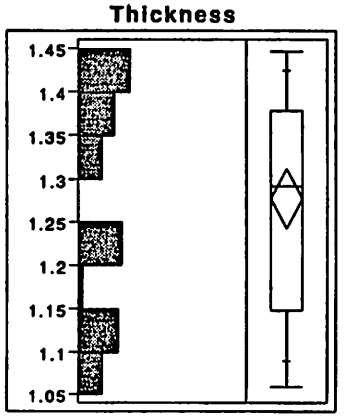
Quantiles

maximum	100.0%	0.82641
	99.5%	0.82641
	97.5%	0.82537
	90.0%	0.78544
quartile	75.0%	0.76292
median	50.0%	0.74557
quartile	25.0%	0.71307
	10.0%	0.68369
	2.5%	0.61767
	0.5%	0.60809
minimum	0.0%	0.60809

Moments

Mean	0.73707
Std Dev	0.04397
Std Err Mean	0.00588
upper 95% Mean	0.74885
lower 95% Mean	0.72529
N	56.00000
Sum Wgts	56.00000

Figure 5. Histogram of PAC fraction remaining in the photoresist before exposure. The Gaussian nature of this data indicates that the variance is simply noise.



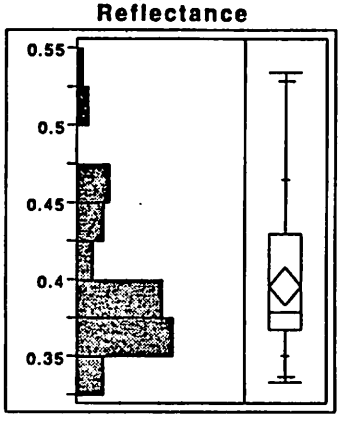
Quantiles

maximum	100.0%	1.4475
	99.5%	1.4475
	97.5%	1.4473
	90.0%	1.4251
quartile	75.0%	1.3784
median	50.0%	1.2914
quartile	25.0%	1.1483
	10.0%	1.0888
	2.5%	1.0599
	0.5%	1.0592
minimum	0.0%	1.0592

Moments

Mean	1.27744
Std Dev	0.12525
Std Err Mean	0.01674
upper 95% Mean	1.31098
lower 95% Mean	1.24389
N	56.00000
Sum Wgts	56.00000

Figure 6. Histogram of Thickness measurements from Eaton wafer track.



Quantiles

maximum	100.0%	0.53470
	99.5%	0.53470
	97.5%	0.52930
	90.0%	0.46541
quartile	75.0%	0.42958
median	50.0%	0.37885
quartile	25.0%	0.36708
	10.0%	0.35004
	2.5%	0.33731
	0.5%	0.33340
minimum	0.0%	0.33340

Moments

Mean	0.39633
Std Dev	0.04724
Std Err Mean	0.00631
upper 95% Mean	0.40898
lower 95% Mean	0.38367
N	56.00000
Sum Wgts	56.00000

Figure 7. Histogram of Reflectance measurements.

Residual Plots of SAMPLE Model for Output Reflectance for A Varying with Input Reflectance

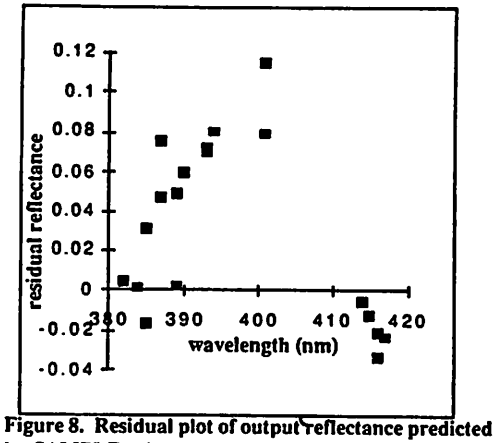


Figure 8. Residual plot of output reflectance predicted by SAMPLE minus measured output reflectance versus wavelength. Note trend in residuals.

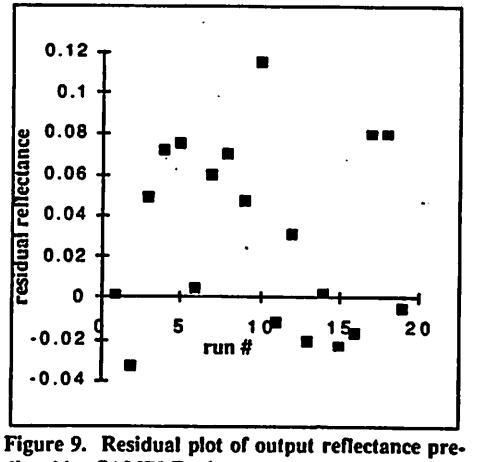


Figure 9. Residual plot of output reflectance predicted by SAMPLE minus measured output reflectance versus run number.

Residual Plots of SAMPLE Output Reflectance for A, B, and C Constant

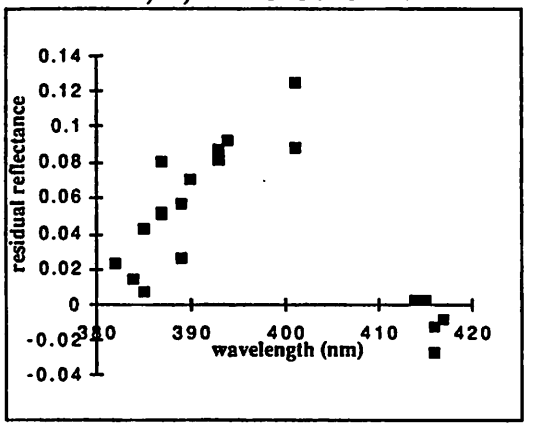


Figure 10. Residual plot of output reflectance predicted by SAMPLE minus measured output reflectance versus wavelength. Note obvious trend with wavelength.

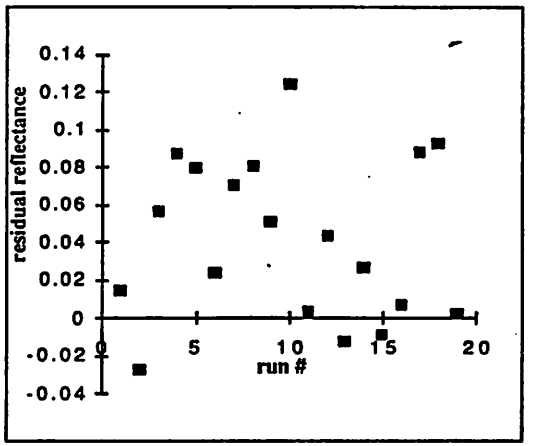


Figure 11. Residual plot of output reflectance predicted by SAMPLE minus measured output reflectance versus run number.

Plot of the Difference Between Predicted Output Reflectance for A Constant and Changing

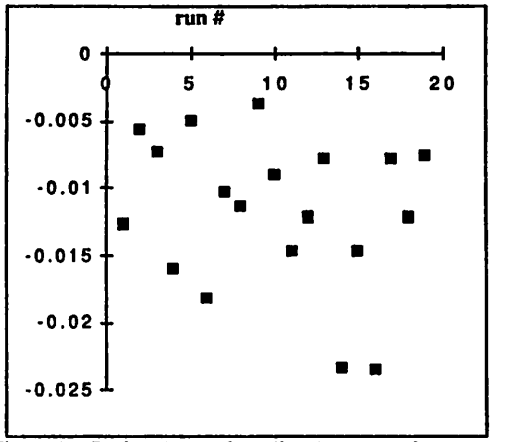


Figure 12. Residual plot of predicted output reflectance for A constant minus predicted output reflectance for A varying with input reflectance. Note that both models give the same result to within +-1%.

Transmission versus Wavelength for Ideal Thin Film Interference of Photoresist on Quartz and for B Values of KTI 820

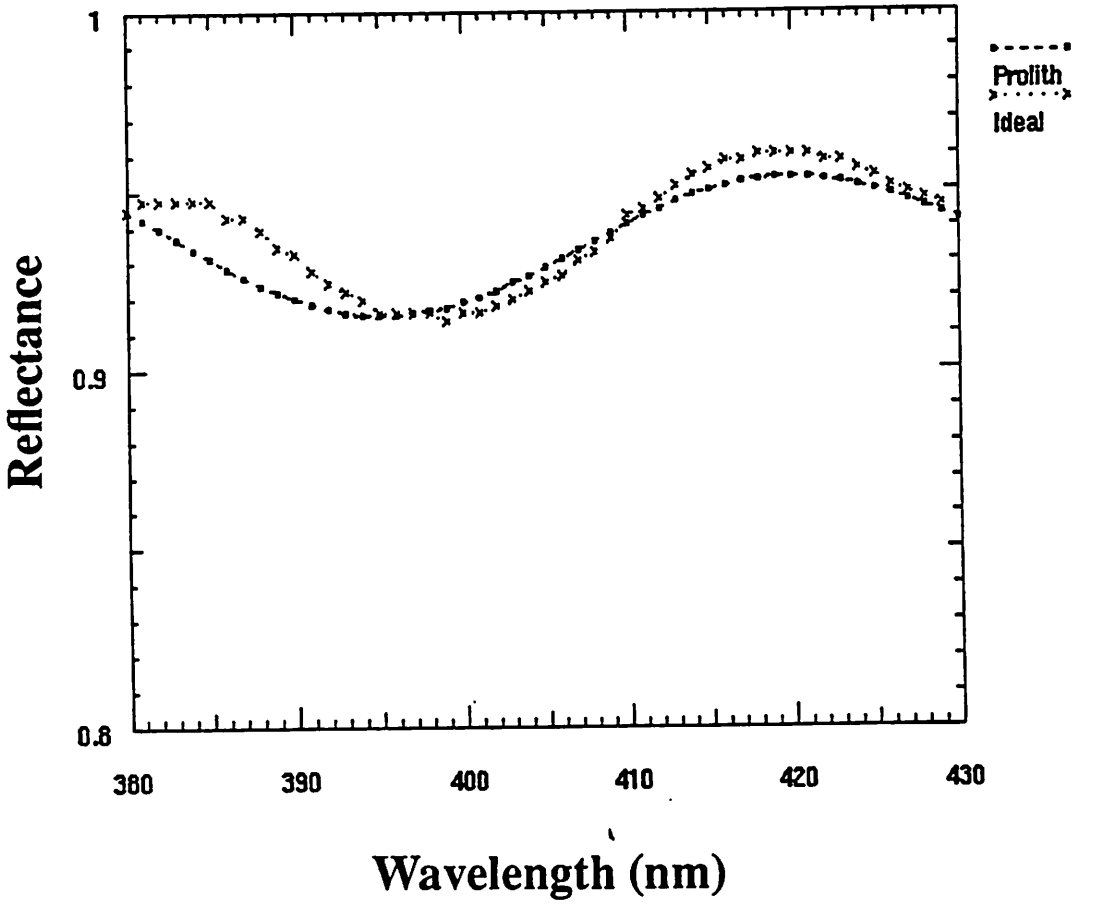


FIGURE 13. Plot of ideal thin film interference pattern from photoresist on quartz in which the photoresist has no absorption component. The transmission derived from Mack's values or B for KTI 820 are also plotted. The correspondence between these curves indicates that the value for B is dominated by the thin film interference of photoresist with quartz rather than the actual absorption constant of the photoresist.

Appendix A: Model for PAC fraction remaining in Photoresist

Response: fraction

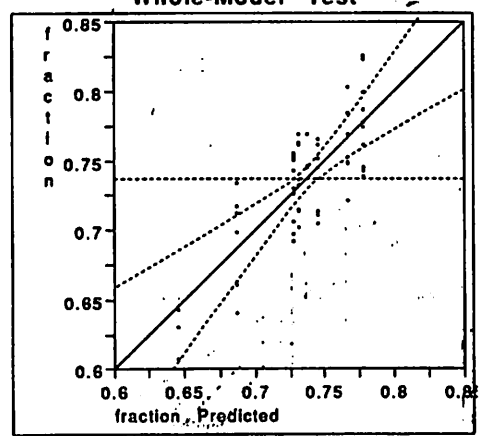
Summary of Fit

Rsquare	.6430638
Root Mean Square Error	.0281196
Mean of Response	0.73707
Observations (or Sum Wgts)	56

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
spin time	1	1	0.01550698	19.6113	0.0001
spin spe*spin tim	1	1	0.01517915	19.1967	0.0001
bake temp	1	1	0.01088440	13.7652	0.0005
spin spe*bake tem	1	1	0.01797138	22.7280	0.0000
spin tim*bake tem	1	1	0.00756036	9.5614	0.0033
Poly(spin speed,2)	2	2	0.04510253	28.5200	0.0000

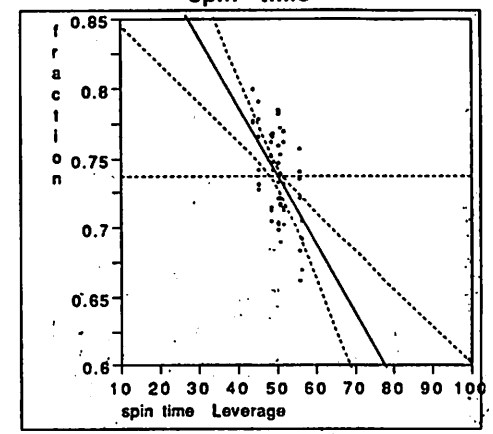
Whole-Model Test



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	0.06837947	0.009768	12.3540
Error	48	0.03795441	0.000791	Prob > F
C Total	55	0.10633387		0.0000

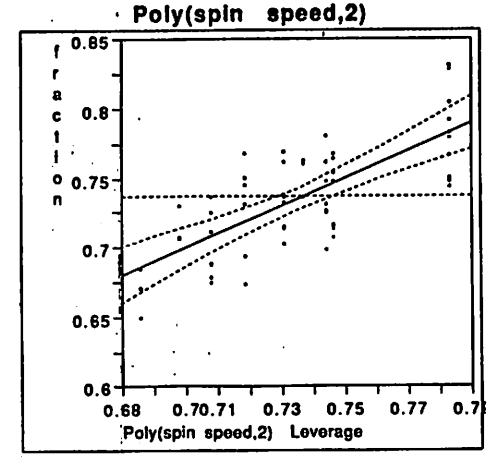
spin time



Effect Test

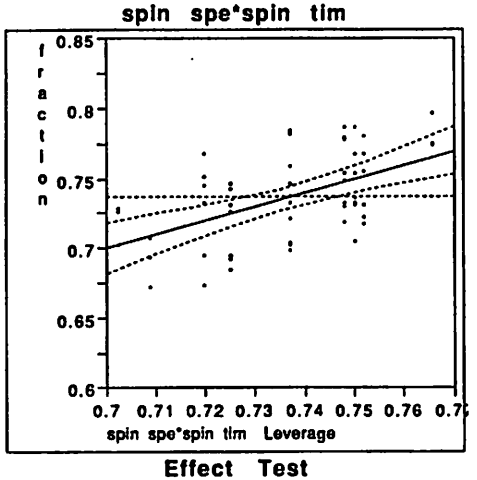
Sum of Squares	F Ratio	DF	Prob > F
0.01550698	19.6113	1	0.0001

Poly(spin speed,2)



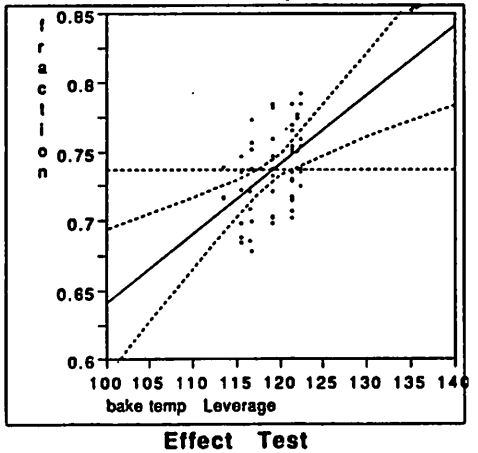
Effect Test

Sum of Squares	F Ratio	DF	Prob > F
0.04510253	28.5200	2	0.0000



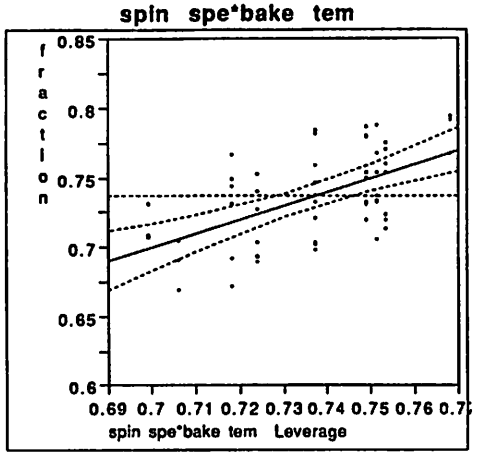
Effect Test

Sum of Squares	F Ratio	DF	Prob > F
0.01517915	19.1967	1	0.0001



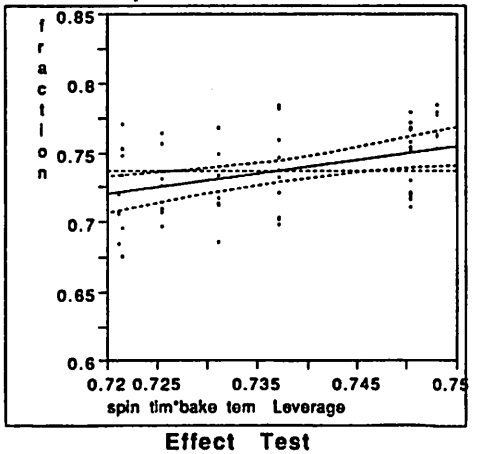
Effect Test

Sum of Squares	F Ratio	DF	Prob > F
0.01088440	13.7652	1	0.0005



Effect Test

Sum of Squares	F Ratio	DF	Prob > F
0.01797138	22.7280	1	0.0000



Effect Test

Sum of Squares	F Ratio	DF	Prob > F
0.00756036	9.5614	1	0.0033

A G2 Formulation of Queuing Effects due to Metrology in a Photolithography Workcell

Bart Bombay

The need for the improvement of photolithography workcell capability requires regular measurements of equipment performance. This report analyses an application of Gensym Corporation's G2 software to simulate wafer measurement scheduling problems.

1.0 Introduction

Recent developments in integrated circuit design call for the improvement of the performance of photolithography workcells. In order to accomplish this improvement, computer aided manufacturing techniques are being applied to the process, and these techniques require regular measurements of equipment performance. These measurements, however, are subject to the associated costs of additional hardware, time, and labor. Hence the industry is faced with the problem of implementing these measurements in a manner which will minimize the cost per unit product produced yet improve product quality. The most obvious goals are to increase product yield (decrease the fraction nonconforming) and improve product performance. Because measurements will slow down the manufacturing process, the desired implementation will attempt to minimize the impact of taking these measurements upon the product throughput, and thus attempt to maintain a satisfactory production level.

There are several issues which must be addressed in any formulation of this scheme. Specifications must be determined on how many wafers to measure, which wafers to measure, and how often to measure them. The types of measurements must be decided upon. The effects on work in progress inventory must be examined. And finally the production costs must be studied to determine the magnitude of any improvement in marginal cost versus marginal revenue.

The Berkeley Computer Aided Manufacturing (BCAM) group was recently presented with the opportunity to study a new software product from Gensym Corporation. The product, G2, is a flexible tool which uses an object oriented environment to simulate and control various types of systems. Of particular interest are this product's extended graphical capabilities which assist an operator in using the system.

For this study, the G2 software was used to simulate a photolithography workcell and to study the effects of introducing a measurement strategy into the workcell. The feasibility of using G2 as an interface to a control and monitoring system is also addressed.

2.0 Methodology

The G2 software possesses several appealing features. Among these are its graphics capabilities, its object oriented environment, its simulation ability, and its general flexibility. In order to introduce customers to the software, Gensym provides a two day course on the G2 system. This course proved effective in familiarizing new users with the general use of G2.

Although G2 is very flexible, considerable effort must be expended to program algorithms into the system. Also, the one second clock cycle of G2 is rather restrictive. These limitations prevent the use of G2 to implement the generalized BCAM control and monitoring system. However, one interesting feature of the

software is its ability to interface with C programs. This feature leads to the possibility of a more limited use of G2 as a graphical interface to the BCAM software, which is written primarily in C++. While such an implementation is attractive, the high cost of G2 precludes such a limited use. For this study a more self-contained application is chosen, namely a study of the queuing problems associated with introducing regular measurements into a photolithography workcell.

The design for this study focuses on the construction of a relatively simple model of the photolithography workcell timing (see figure on the next page). Wafers are processed by a machine and then placed into a storage area. From this area, wafers are either taken to an analytical station for measurement and then transferred to the following storage area, or they are transferred directly to the next storage area. The next processing station then takes its wafers from that storage area. The decisions about whether or not to measure any particular wafer are dependent upon production flow and control criteria.

Initial work to design a knowledge base with G2 includes the definition of several object types and icons, and preliminary connections among instances of these icons. Gensym also supplied a customer support visit which is effective in assisting users new to the G2 system. With such assistance, a basic design was implemented. This basic design may then be further refined with the introduction of an enhanced set of rules, more informative readouts, and more precise timing specifications

3.0 Results

This study compares two distinct algorithms for the scheduling of wafer measurements. These two methods are henceforth referenced as algorithm A and algorithm B and are described below.

3.1 Wafer Measurement Scheduling Algorithm A

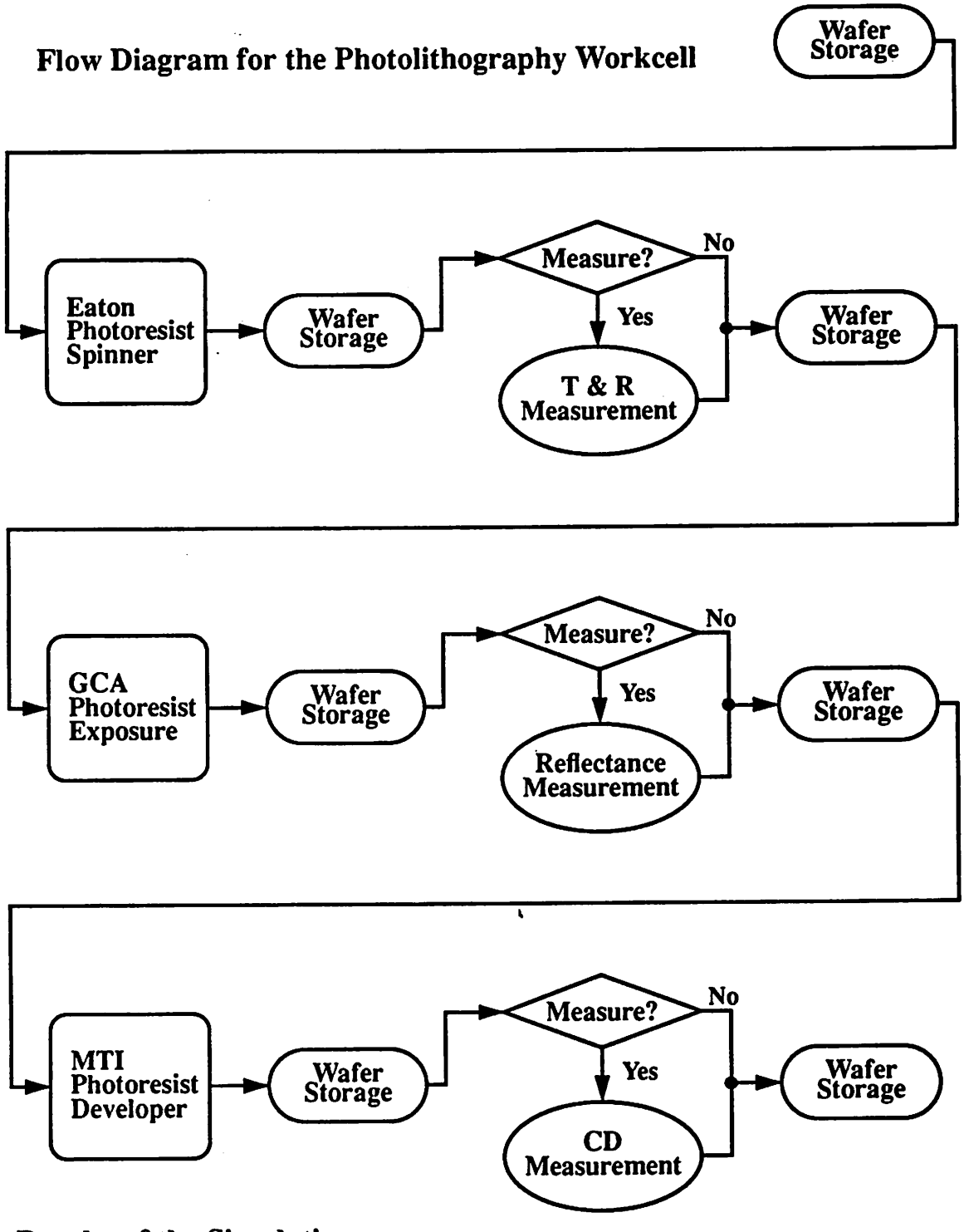
The first method, algorithm A, uses inventory based rules to decide the number of wafers from which measurements would be taken. Each storage area immediately preceding a processing station has a specific "low level". If the wafer count in the storage area falls below this low level, the deficit is immediately taken from the preceding storage area, and the wafers so taken do not get measured. As long as the count of the storage areas remains above or at the low level, each wafer will be subjected to measurement as it passes between storage areas.

Algorithm A proves to be successful in maintaining production levels since it foregoes measurements whenever the relevant intermediate wafer inventories fall below designated low levels. Because wafers require queuing before measurement, this formulation does, however, increase the overhead in progress inventory. Another drawback to this method is the variability of the frequency of measurement; during some time periods, many wafers are measured, while during other time periods few or no wafers are measured.

3.2 Wafer Measurement Scheduling Algorithm B

The second algorithm for wafer measurement, algorithm B, sets specific goals for the number of wafers to be measured at each step in the process. One out of every four wafers is subjected to measurement as it passes between storage areas. This algorithm is indifferent to the supply levels in the storage areas.

Algorithm B is successful at providing a steady stream of data, but results in a somewhat reduced production level. This method also results in a lower work in progress inventory than the first method, although still higher than a process without measurements.



3.3 Results of the Simulations

It should be noted that the simulations fail to give precise information. This deficiency is attributed to two factors. The first limitation of the system is that it discretizes time into one CPU-second intervals. Thus if the time were scaled to simulate five minutes of production every second, then the resolution of the process simulation would be limited to five minutes. The time scale chosen for the simulation is one wafer processing minute per CPU-second. This time scale yields sufficient resolution for the simulation, while providing results after a reasonable period of time. (At this scale, the simulation of a 24 hour workday

requires 24 minutes.) The second limitation of the system results from the structure of the G2 rule system. Creating complex rule patterns with G2, although undoubtedly possible, is time consuming, particularly with respect to making structural changes in the flow decision rules. Hence time requirements for the implementation of the desired simulations exceeded the resources allotted to the project. The project was therefore somewhat scaled back.

Several topics of interest are ignored in these simulations. An analysis of the profitability of the different algorithms is not performed. The simulation of the workcell is idealized. In an actual fabrication facility, the processing times of the various equipment change with varying conditions, including change of operators, and random noise. The equipment in an actual workcell also experiences periodic downtime due to failures and general maintenance. The relative time requirements of the processing steps may also change with different product lines. In addition, the changing operating conditions in a fabrication facility may require a dynamically changing measurement scheduling algorithm which can emphasize data collection for issues of interest, while reducing the emphasis on lesser issues. This report does not address these problems.

The results of the simulations are strongly dependent on the specific time requirements of the particular elements in the workcell, especially the time required to take measurements. Since these time requirements vary significantly for different technologies, the results of this project should only be interpreted in a relative manner.

In particular, many of the relevant measurements can now be implemented 'in situ' on the wafer track so that they have no impact on the wafer processing time. In such a case, measurements can easily be made on all wafers, providing valuable information to an appropriate process control and SPC system. Thus the only increase in cost comes from the purchase and maintenance of the new measurement equipment.

In the case that measurements are taken off the wafer track, the time for measurement is an important consideration. Some measurements require more time than others. (For instance in the Berkeley Microfabrication Laboratory, a manual critical dimension measurement may require tenfold the time required for a photoresist thickness measurement.) When faced with such circumstances, a successful scheduling scheme may reduce the frequency of measurement for those measurements which are time intensive.

For model based control schemes a measurement scheduling algorithm must ensure the maximization of the number of wafers which are measured at all stations. Thus wafers which were previously measured receive priority for future measurements in order to facilitate model building. For feed-forward control, every wafer (or at least samples from every lot) must be measured. For statistical quality control, an increase in the number of measurements taken will almost always be beneficial. Maximizing the frequency of measurements in the processing line will expedite the detection and diagnosis of equipment problems.

4.0 Example

Figure 2 displays a screen dump of the G2 formulation of a photolithography workcell.

5.0 Conclusions

The results of this project are highly dependent upon the configuration of the photolithography workcell. In general, any increase in measurement frequency is beneficial as long as it does not cause too great an increase in cost. 'In situ' measurements on wafers along the wafer track are extremely desirable, as they do not cause delays in the processing line.

Drawbacks of in-process measurements:

- Cost of the measurement equipment and its maintenance

- If a system is dependent on measurements, measurement equipment failure could cause interruptions in production.
- The time requirements of measurements can slow down production.
- Queuing requirements of measurement scheduling can increase the work in progress inventory.

Advantages of in-process measurements:

- Additional information for problem detection and diagnostic efforts
- Quantitative records of machine performance
- In-process measurements allow the implementation of a feed-forward control scheme to eliminate the propagation of disturbances and increase yield.
- Measurement data assists in the development of equipment models for various control and design purposes.

6.0 Future Work

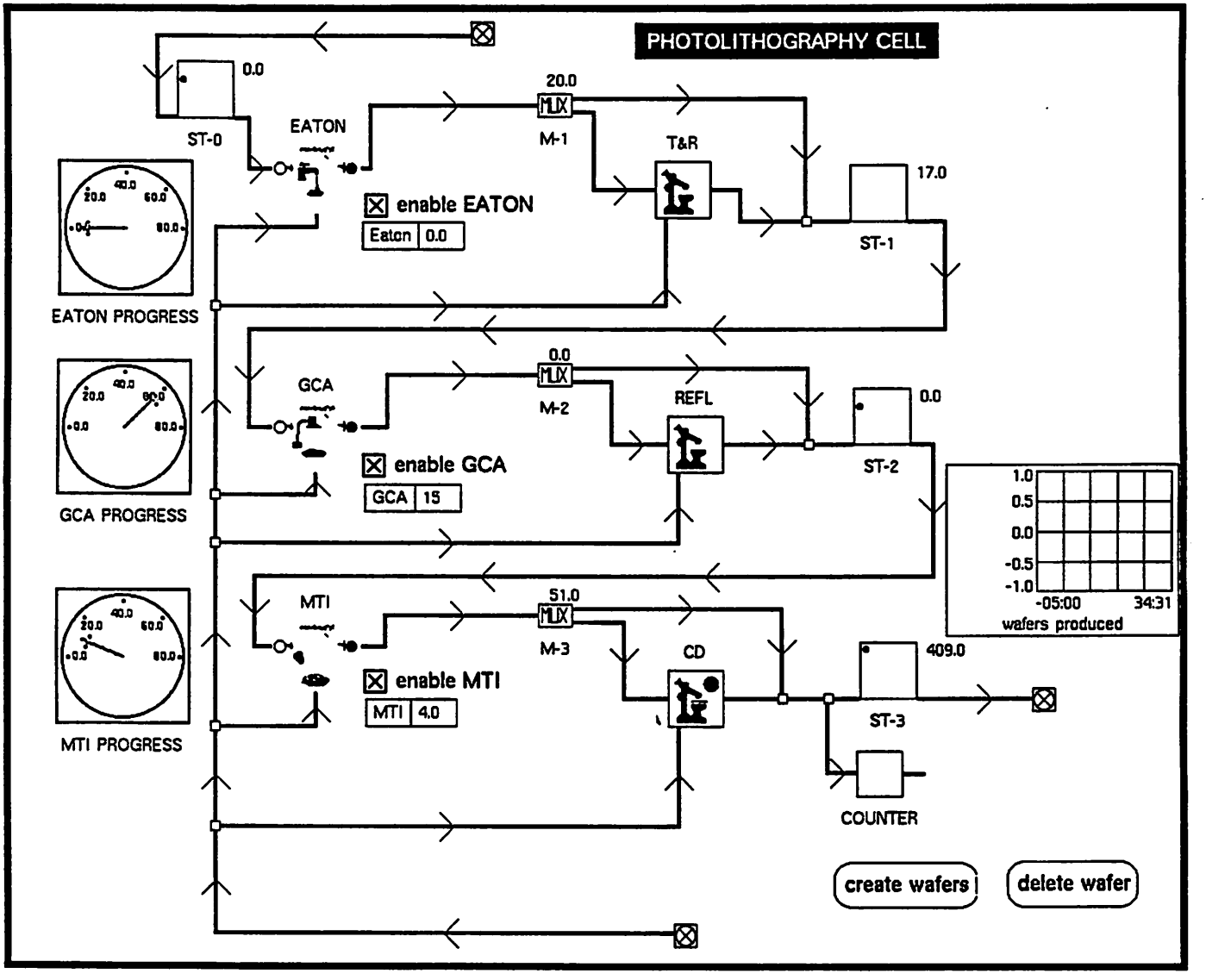
A comprehensive study requires more detailed models of photolithography equipment performance. The G2 representation of the photolithography cell might be expanded to include interfaces to fabrication and measurement equipment, as well as interfaces to C code to handle computationally intensive control and modeling computations. The system would then be able to handle many applications, including scheduling, model-based control, statistical quality control, diagnosis, recipe design for equipment operation, and database operations. In such a case, G2 would serve primarily as a graphical interface to a computer aided manufacturing system, and C code would provide the remainder of the functionality. This G2 formulation of a computer aided manufacturing system would, however, be limited to operations which require time discretization at a level no lower than one second intervals, as this is the maximum clock speed of the G2 system. The G2 representation could also be expanded to include multiple workcells and thereby simulate and control an entire production process. Such an implementation would interact well with G2's object oriented structure.

From our limited exposure to G2 we were impressed by its capability to produce an effective, animated pictorial summary of the process. This project also yielded the following suggestions for improving G2's applicability to integrated circuit manufacturing:

- G2 lacks the computational power required for advanced control and modeling purposes. The existing interface with the C programing language has not been tested by the author.
- The current 1-CPU-second system clock is too slow and inflexible.
- G2 is rather unwieldy for new users. Significant training is required before the user-interface of G2 feels natural, as many of the most common tasks require convoluted menu selections.
- The object-oriented framework can be improved. Currently there are limitations when updating structures. Specifically, instances must often be deleted and recreated whenever base structures change.

7.0 Acknowledgments

The author is grateful for the help received by Gensym's Vasu Subbiah during this project.


PHOTOLITHOGRAPHY CELL
0.0
ST-0
EATON
20.0
MIX
M-1
T&R
17.0
ST-1
enable EATON
Eaton 0.0
EATON PROGRESS
GCA
0.0
MIX
M-2
REFL
0.0
ST-2
enable GCA
GCA 15
GCA PROGRESS
1.0
0.5
0.0
-0.5
-1.0
-05:00
34:31
wafers produced
MTI
51.0
MIX
M-3
CD
409.0
ST-3
enable MTI
MTI 4.0
MTI PROGRESS
COUNTER
create wafers
delete wafer

Sidewall Slope Optimization for Phase Shifted Contact Cuts

John Helmsen

The goal of this project is to investigate the experimental space of the photoresist etching step in a contact cut manufacturing process which uses phase shifted masks. By using the simulation tools SPLAT, BLEACH and ETCH, the phase shift contact cut process was examined for its effect on the sidewall slope when four parameters are varied. The parameters are the two mask dimensions, the misalignment from the focal plane and the coherence of the light source. The simulation space is mapped on selected two dimensional surfaces in the four dimensional space. The first order effects of the parameters are also mapped to localize the points of minimum variation.

1.0 Introduction

The semiconductor manufacturing industry, in its attempts to achieve minimal feature sizes, has recently adopted the use of phase shift masks[2][3]. These masks differ from traditional masks by producing a diffraction pattern on the surface of the photoresist. Existing optical and exposure equipment may be used to produce smaller feature sizes. While use of these masks is therefore desirable, the exposure step of a process must be reexamined to determine its optimal regions of operation and sensitivity to optical parameters[2].

Due to the excessive cost of conducting of analyzing a process through conducting actual experiments, it is often instructive to map the parameter space through process simulation. This allows the experimenter to reduce the number of fabrication runs to describe the process, because the simulation can be used as an accurate initial guess. Experiments are still necessary, however, to confirm the simulated result because the simulator may have inaccuracies.

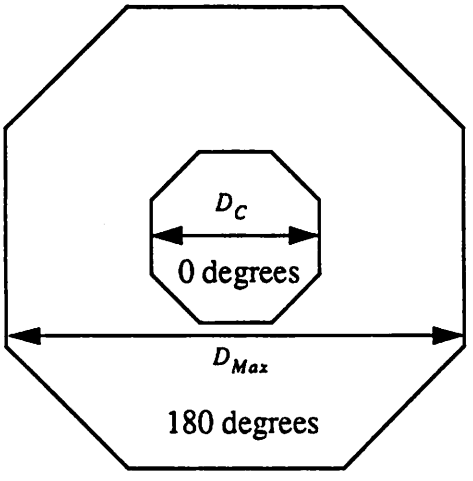
The specific process of photoresist etching has been chosen for examination, because the SAMPLE-3D [5]suite of simulators performs this particular simulation task effectively. Three of these simulators were made to work in conjunction with one another. The first is SPLAT [3], which generates the intensity contour on the surface of the photoresist from the mask and the optical parameters. The intensity contour is sent to BLEACH [5], which simulates the exposure of the resist and determines the etch rate throughout the exposed resist layer. Finally, the ETCH [6] program, simulates development of the photoresist during the etching process.

In Section 2, a full description of the photoresist etching process is given and the inputs and functions of the simulators are described. Section 3 details how the simulators were employed to simulate the process, and the manner in which the results were generated. The data and its analysis is presented in Section 4. Conclusions are presented in Section 5.

2.0 Photoresist Etching

2.1 The Mask

A phase shift mask structure has been proposed for forming contact cuts [2] and is shown below. This mask contains two octagonal transparent regions, one centered inside the other. The central region has a minor diameter of length D_C. This region is not phase shifted. The outer ring has a minor diameter of length D_{Max}. This region is phase shifted by 180°. The inner region and the outer ring are concentric. The two mask diameters are two of the parameters in the simulation space. This mask produces a stronger and thinner central spot than a non-phase shifted mask, provided the phase shift mask is of the proper dimensions. At the surface of the photoresist, the two phase shifted components destructively interfere to create a



Phase Shift Mask and Dimensions

ring of zero intensity. The interference at the center of the image reacts constructively and creates a central spot of exceptional magnitude. These are both desirable conditions for exposing photoresist for a contact cut, because the hole created will be thinner and have steeper sidewalls than a cut created by a normal mask. The disadvantage is constructive interference again occurs at twice the distance of the dark ring from the center. This sidelobe, although normally low in intensity, may partially develop the resist. It can interfere with nearby structures, so it is desirable to reduce its intensity when possible.

2.2 Creating the Image

Simulation of the optics is handled by the SPLAT program. It accepts as input a description of the mask and the dimensions of the area to be imaged. It also requires the following:

Table 1: SPLAT Parameters

Parameter	Meaning	Value
λ	Wavelength	435.8 nm
NA	Numerical Aperture	0.45
Focus	Distance from Focal Plane	Experimental Parameter
σ	Partial Coherence	Experimental Parameter

The wavelength is chosen to be in the g-line regime because a g-line resist is most appropriate for this experiment. The numerical aperture is a physical parameter based on the lens dimensions and index of

refraction, and was chosen to be 0.45 for consistency. The distance from the focal plane is measured in μm. Sigma represents the partial coherence of the imaging system. It may take values from 0 to 1. The focal distance and the partial coherence are the other two parameters, besides the mask parameters, in the experimental space. The output of the SPLAT program is a discretized representation of the intensity of the image as it appears at the surface of the photoresist.

2.3 Exposing the Resist

Table 2: BLEACH Parameters

Parameter	Meaning	Value
Resist	Photoresist	SNR-248
Dose	Exposure Dose	100 mJ/cm^2
Diffusion	Heat Diffusion Length	1 μm
Thickness	Thickness of Resist	0.7133 μm

Exposure of the resist is performed by the BLEACH simulator. A surface intensity contour is taken as input, along with a file that describes the parameters for the photoresist. The resist chosen is the SNR-248 model [1]. This is a g-line acid hardening resist, which is especially useful for phase shift masks, because a certain threshold of intensity is necessary to expose the resist. The etching of the resist due to the sidelobes is, therefore, less pronounced. The exposure dose and the resist thickness are held constant for all simulated exposures. The diffusion parameter is included due to the formation of standing waves in the resist. Because the energy reflects off of the substrate during exposure, alternating layers of high and low etch rates can form. If the photoresist is etched without a preceding diffusion step, the vertical sides will have a rippled character. A one-dimensional vertical gaussian diffusion is therefore performed so that the ripples are removed. BLEACH generates as output a three-dimensional array which contains the etch rates at regular points in the resist.

2.4 Etching the Resist

Table 3: ETCH Parameters

Parameter	Meaning	Value
Time	Development Time	6 seconds
N	Surface Discretization	20

The three dimensional structures that form when the photoresist is etched, are computed by the ETCH simulator. It takes as input the three-dimensional etch rate array produced by BLEACH. The surface in ETCH is represented by a triangular mesh. Its evolution is computed by solving a PDE which is discretized in space by the parameter N. N is the number of triangles in both the X and Y directions. The PDE is also discretized in time. The time step is variable and controlled internally. The limitation on the time step is that the distance traveled by the surface during one step must be less than 15% of the length of the side of the original triangles. This 15% condition causes ETCH to give highly accurate results. The photoresist development time for all simulations is 6 seconds. ETCH produces a list of triangles as its output. This list is the geometrical representation of the developed surface.

2.5 Sidewall Slope

The sidewall slope was selected as the measurement parameter. Because the intention of a contact cut processes is to make a small hole with straight sides, this is an accurate indicator of the effectiveness of the process. The sidewall slope is also easy to derive automatically. This makes it an especially effective measurement of process suitability when the number of simulation results is too large to be analyzed by the user.

3.0 Implementation

3.1 Inputs

Table 4: Experimental Parameters

Parameter	Meaning	Value	Step	# of Values
D_C	Inner Mask Diameter	0.8 μm to 1.6 μm	0.1 μm	9
D_{Max}	Outer Mask Diameter	1.3 μm to 2.61 μm	0.164 μm	9
Focus	Distance from Focal Plane	-5.0 μm to 5.0 μm	1.0 μm	11
σ	Partial Coherence	0.01 to 1	0.1	11

Mapping the entire space was not attempted due to the prohibitive number of simulations which must be performed. Certain coordinate parallel two dimensional planes were selected for analysis in the four dimensional simulation space. In each of these cases, two parameters were held constant, while the other two parameters were varied over their entire range. For any particular variable, the partitioning of the range is uniform, except when the partial coherence is equal to 0.01. This is done to avoid a divide by zero error in SPLAT. The following planes were analyzed:

Table 5: Examined Planes

Plane	D_C	D_{Max}	Focus	σ
1	Varies	Varies	0.0 μm	0.5
2	Varies	Varies	1.0 μm	0.5
3	Varies	Varies	-1.0 μm	0.5
4	Varies	Varies	0.0 μm	0.01
5	Varies	Varies	0.0 μm	0.3
6	Varies	Varies	0.0 μm	0.4
7	Varies	Varies	0.0 μm	0.6
8	Varies	Varies	-1.0 μm	0.4
9	1.10 μm	2.12 μm	Varies	Varies

3.2 Output

The result of each simulation was analyzed automatically to determine the slope of the contact cut sidewalls. The slope of the sidewall S_{Side} is determined (EQ 1) by the observed diameter of the contact cut at the middle of the resist D_{Mid}, the bottom of the resist D_{Bot} and the depth of the resist R_{Depth}.

$$S_{Side} = \frac{D_{Mid} - D_{Bot}}{R_{Depth}/2} \quad (1)$$

4.0 Results

The 9 planes that were used as input in Table 5, are plotted in Figures1through 9 respectively using the CONTOUR program [7].Figures 1 through 9 demonstrate that the mask dimensions that consistently give large sidewall slopes are D_C of 1.1 and D_{Max} of 2.12. Figures 2, 3 and 9 demonstrate that the experimental

space is symmetric about the focal plane. This effect was expected for this mask configuration [4]. The sidewall slopes increase as the coherence tends towards 0 in figures 4, 5, 6, 7 and 9. The effect of lower coherency on the sidewall slope is shown directly by the side views in Figures 13, 14 and 15. These figures are plotted at D_C of 1.1, D_{Max} of 2.12 and a Focus of 0. Figures 13, 14 and 15 have coherency of 0.01, 0.5 and 0.9 respectively. These figures also show that sidelobes become more pronounced for lower coherency. Figures 10 and 11 demonstrate the sensitivity of the experimental space to first order changes in the focus and coherence respectively. Figure 12 graphs the sensitivity of the space to a change in both the focus and sigma simultaneously. Examining the plots of Figures 10, 11 and 12, an important point is located. The change in the slope for D_C of 1.1 and D_{Max} of 2.28 is near 0 sensitivity for changes in both focus and sigma (where focus is about. 0 and sigma is about. 0.5.) The combined effect from a change in both focus and coherence is also minimal. This spot may be considered the least sensitive to changes in the process.

5.0 Conclusions

A parameter space for a phase shifted contact cut photolithography process has been investigated. The dimensions of the mask that give the best sidewall slope have been determined. The process space has been shown to be symmetrical about the focal plane, and the effects of coherence have been investigated. The most important future work is to confirm these results in the lab. Investigation of the process for other cost functions besides sidewall steepness may be performed.

6.0 References

[1] R.A. Ferguson, "Modeling and Simulation of Reaction Kinetics in Advanced Resist Processes for Optical Lithography", *Ph.D. Dissertation*, University of California, Berkeley, May 1991.

[2] M.D. Levenson and F.M. Schellenburg, "Phase-Shifting Mask Strategies: Isolated Bright Contacts", *Microlithography World*, May/June 1992.

[3] D.M. Newmark, "Computer Aided Design Tools for Phase-Shift Masks and Spatial Filtering", *Masters Report*, University of California, Berkeley, Dec. 1991.

[4] A.R. Neureuther, Personal Communication.

[5] E.W. Scheckler, K.K.H. Toh, D.M. Hoffstetter,and A.R. Neureuther, "3D Lithography, Etching, and Deposition Simulation (SAMPLE-3D)", *1991 Symposium on VLSI Technology, Digest of Technical Papers*, Oiso, Japan, May 28-30, 1991.

[6] E.W. Scheckler, "Algorithms for Three-Dimensional Simulation of Etching and Deposition Processes in Integrated Circuit Fabrication", *Ph.D. Dissertation*, University of California, Berkeley, Nov. 1991.

[7] E.W. Scheckler, *A User's Guide for SAMPLE-3D v1.0: Lithography Simulation*, University of California, Berkeley, Nov. 1991.

FIGURE 1.

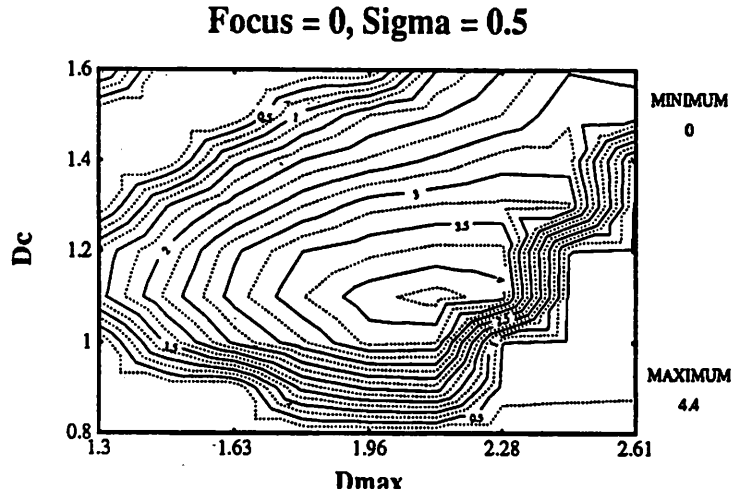


FIGURE 2.

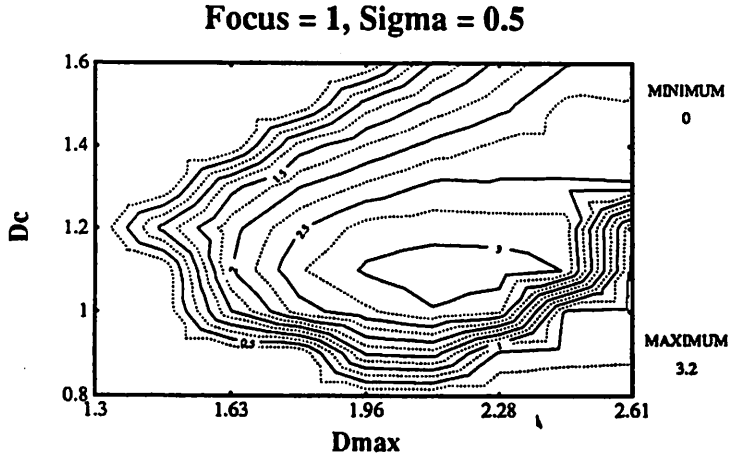


FIGURE 3.

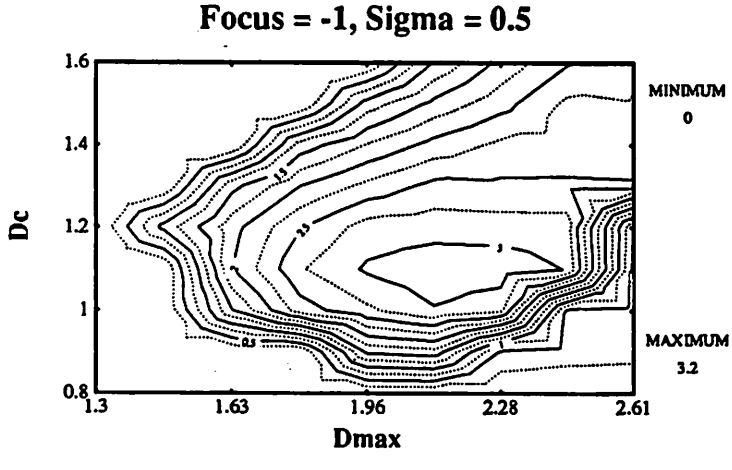


FIGURE 4.

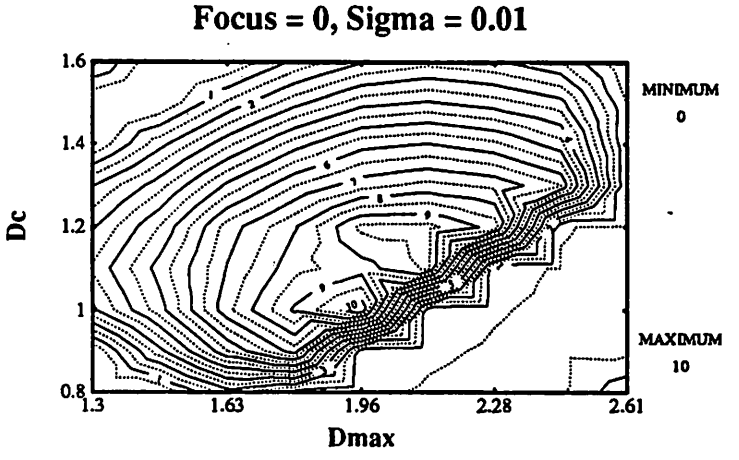


FIGURE 5.

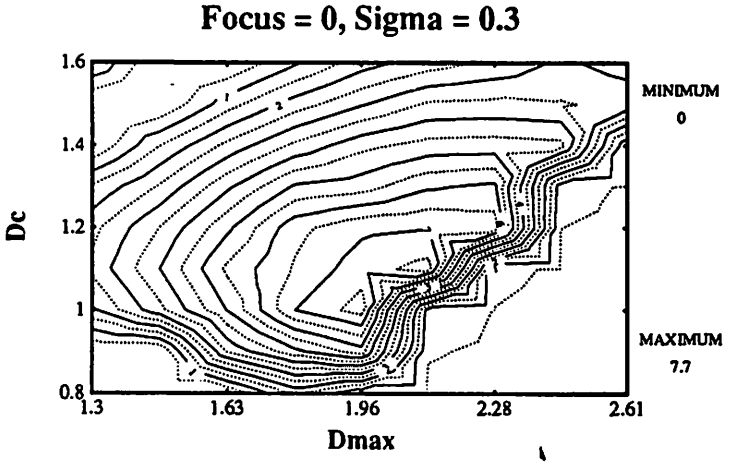


FIGURE 6.

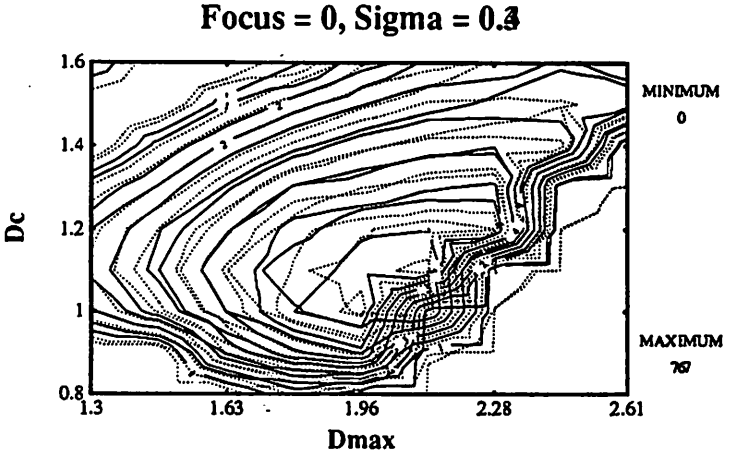


FIGURE 7.

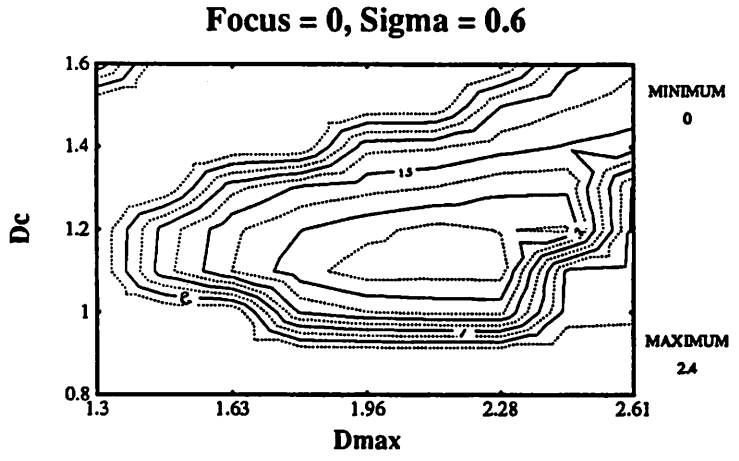


FIGURE 8.

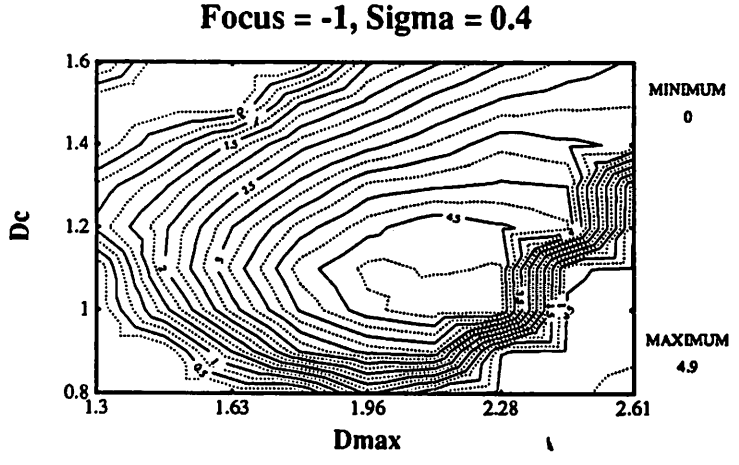


FIGURE 9.

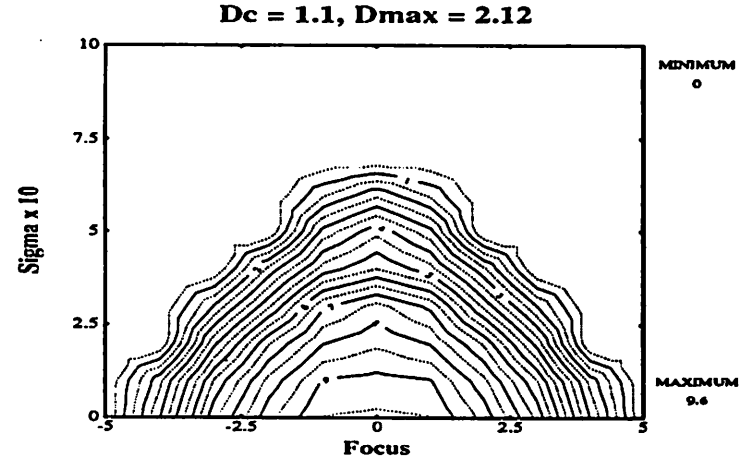


FIGURE 10.

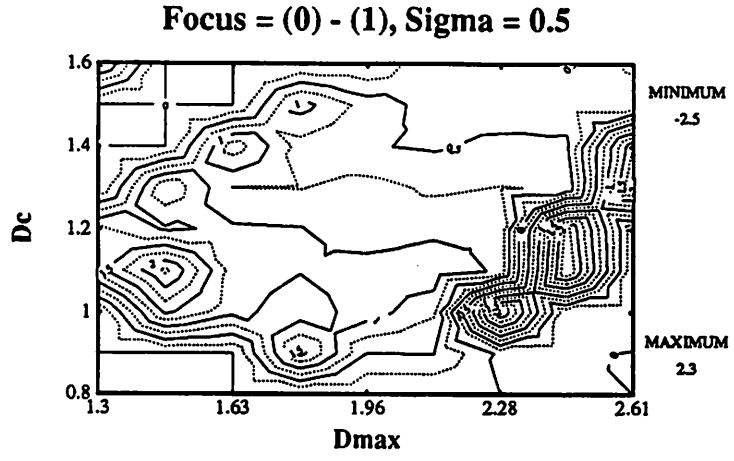


FIGURE 11.

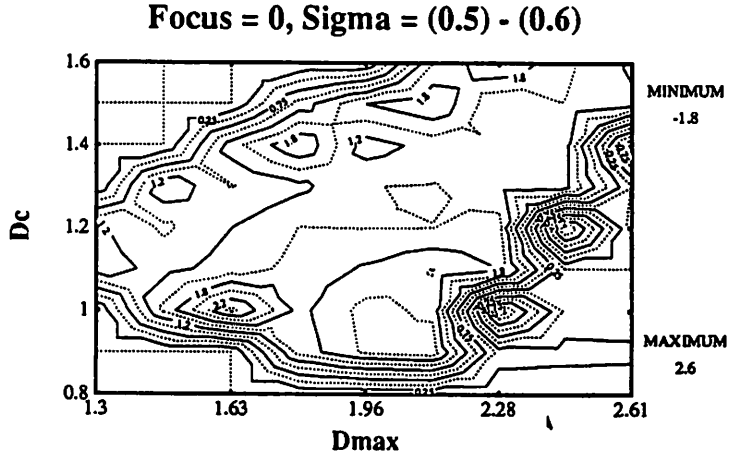


FIGURE 12.

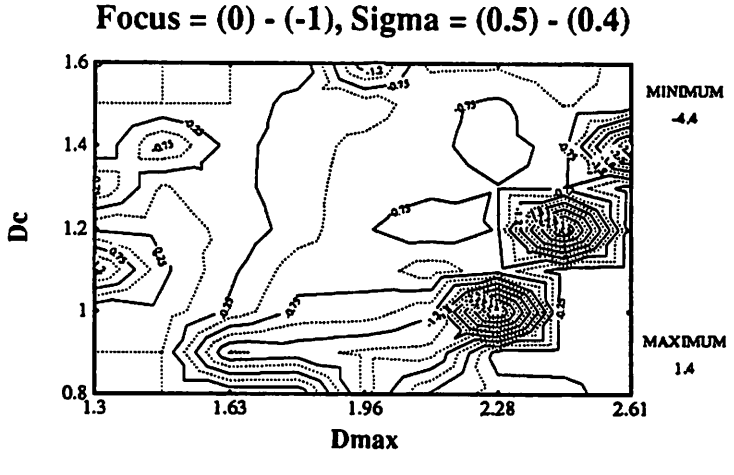


FIGURE 13.

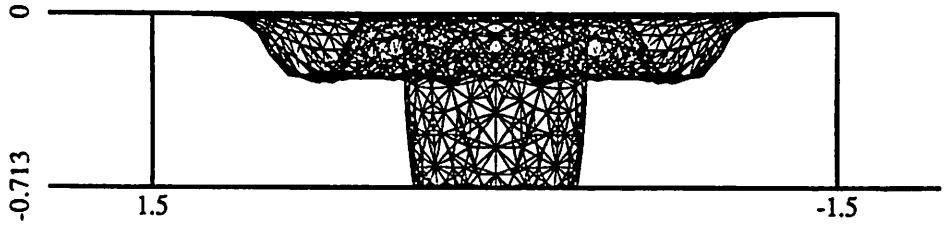


FIGURE 14.

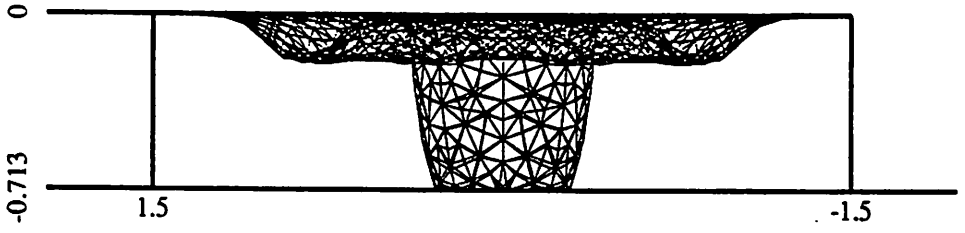


FIGURE 15.