SPECIAL ISSUES IN SEMICONDUCTOR MANUFACTURING

VI

EECS 290W Class Projects Reports, Spring 1995

Professor:

Costas J. Spanos

Students:

Jone Chen, Roawen Chen, Charles Fields, Herb Huang, Anna Ison, Xinhui Niu, Donggun Park, Jiang Tao, Manolis Terrovitis, and Amy Wang

Memorandum No. UCB/ERL M95/64

8 August 1995

SPECIAL ISSUES IN SEMICONDUCTOR MANUFACTURING

VI

EECS 290W Class Projects Reports, Spring 1995

Professor:

Costas J. Spanos

Students:

Jone Chen, Roawen Chen, Charles Fields, Herb Huang,
Anna Ison, Xinhui Niu, Donggun Park, Jiang Tao,
Manolis Terrovitis, and Amy Wang

Memorandum No. UCB/ERL M95/64

8 August 1995

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Preface

This is the sixth annual edition of the 290W report. This edition includes descriptions of projects completed during the Spring semester of 1995, in the context of the graduate course "Special Issues in Semiconductor Manufacturing". Ten students have participated, and according to the course requirements, these students worked with me on their projects during the last six weeks of the semester.

Each of the presented projects covers at least one novel aspect of semiconductor manufacturing. The first four projects deal with statistical modeling of processing steps: The first discusses the characterization of e-beam lithography, the second with thin film deposition, the third studies ion implantation damage, and the fourth models a novel plasma immersion ion implantation process.

The second group, consisting of four projects, deals with statistical process control and its extension for multiple real-time data streams, the introduction and statistical filtering of Optical Emission Spectroscopy, and the modeling of slow temporal variability in plasma etching. The last project in this group discusses the development of novel in-situ metrology for photolithography control during the definition of a poly gate stack.

The third group consist of two projects that deal with IC design for manufacturability issues. The first of these projects discusses the application of robust design principles, while the second analyzes the electrical mismatch of devices fabricated in the Berkeley Microfabrication Laboratory.

It is my hope that these reports will add to our understanding of semiconductor manufacturing. My thanks go to the 290W students 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.

Costas J. Spanos

August, 1995

Table of Contents

1. Characterization of Factors Affecting Electron-Beam Lithography using Regression *Page 7*

Charles H. Fields

2. A Model for the Formation of Thin Film Anodized Aluminum *Page 17*

Amy Wang

3. Ion Implant Damage Study Using a Factorial Design *Page 25*

Donggun Park

4. Modeling Plasma Immersion Ion Implantation by a Response Surface *Page 33*

Jiang Tao

5. Multiple Data Streams in Real-Time Multivariate Statistical Process Control *Page 39*

Herb Huang

6. Real Time SPC for Plasma Etching Using Optical Emission Spectroscopy *Page 49*

Roawen Chen

7. Modeling Temporal Variability on a Lot to Lot Basis in Manufacturing Equipment *Page 65*

Anna Ison

8. In-Situ Poly Gate Photoresist Metrology and Control *Page 77*

Xinhui Niu

9. Application of the Robust Design Method for IC Design Improvement *Page 91*

Jone Chen

10. Transistor Matching Properties of the UC Berkeley CMOS Process *Page 101*

Manolis Terrovitis

Characterization of factors affecting electron-beam lithography using regression analysis

Charles H. Fields

This paper explores the results of a 3^5 full factorial experiment undertaken to analyze the factors that affect the ultimate resolution achievable through electron-beam (e-beam) lithography. The exposure tool used for this experiment was a Jeol 6400 Scanning Electron Microscope (SEM) that has been modified to perform e-beam lithography. The five variables investigated are electron energy expressed in kV, exposure dose in $\mu C/cm^2$, photoresist thickness, line spacing pitch, and drawn linewidth. The results of the experiment were analyzed using regression analysis. The significance of each parameter and cross-interactions was analyzed and a model was fit to significant terms. The results of the analysis are summarized in the conclusion section of this paper.

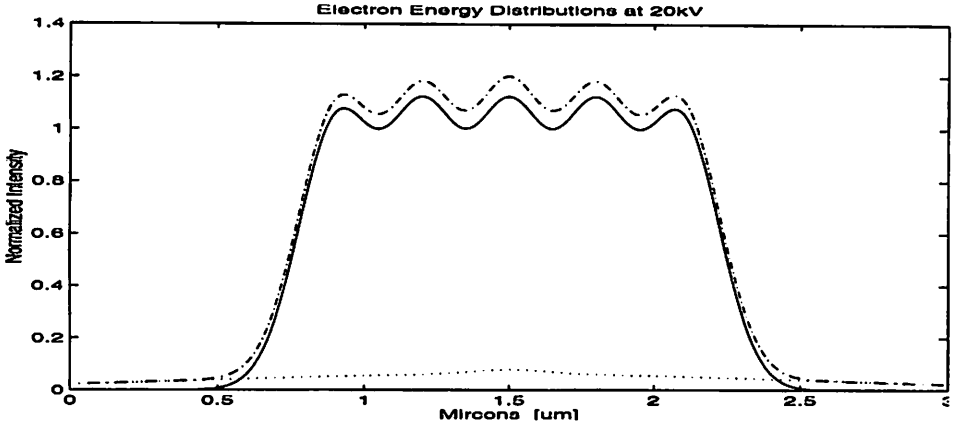
1.0 Introduction

Electron-beam lithography offers the capability of patterning sub-quarter-micron features. In fact, features as small as 10 nanometers have been patterned with the use of e-beam lithography using lift-off and special pattern transfer techniques [2,3,4,5,6,7,8]. Indeed, this form of lithography has many interesting possibilities for the future of lithography in the semiconductor manufacturing industry.

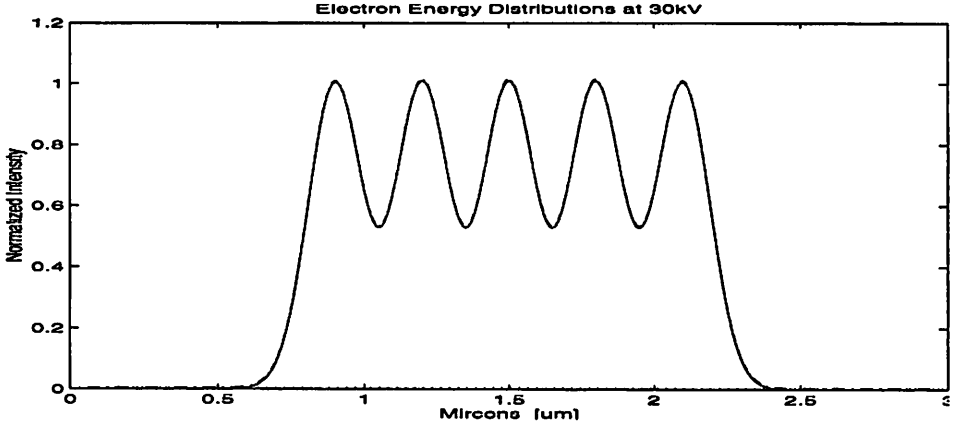
One might wonder why e-beam lithography is not more widely used in today's semiconductor manufacturing facilities. The reason is that there are still many obstacles to overcome before a production-ready system would be feasible. Electron beam lithography differs from conventional optical lithography in many areas. One difference is the photoresist used in e-beam lithography: Poly-methylmethacrylete (PMMA). While PMMA is capable of good resolution (on the order of 5nm [1]), it suffers from poor plasma etch resistance which limits its effectiveness in pattern transfer.

Another factor which limits the use of e-beam lithography is proximity effects. It is well known that e-beam lithography suffers from proximity effects due to electron scattering, which affect the critical dimensions (CDs) of the printed linewidths. This is due to elastic and inelastic collisions of the electrons with the resist and substrate atoms. These collisions act to distort the drawn patterns by scattering electrons both in the forward and backward directions. Computer simulations of electron trajectories support the theory that the electron distributions of both the forward and back scattered electrons are nearly gaussian. The forward and backward distributions can be characterized by separate standard deviations denoted σ_f and σ_b respectively. Figure 1 shows the distribution of the forwardscattered, backscattered and the total electron energy distributions for single pass lines drawn at the three electron energies studied in this paper. This figure illustrates the fact that electron beam energy will affect both the forward and the backscattered dose distributions and therefore the total dose distribution deposited in the resist. The forward-

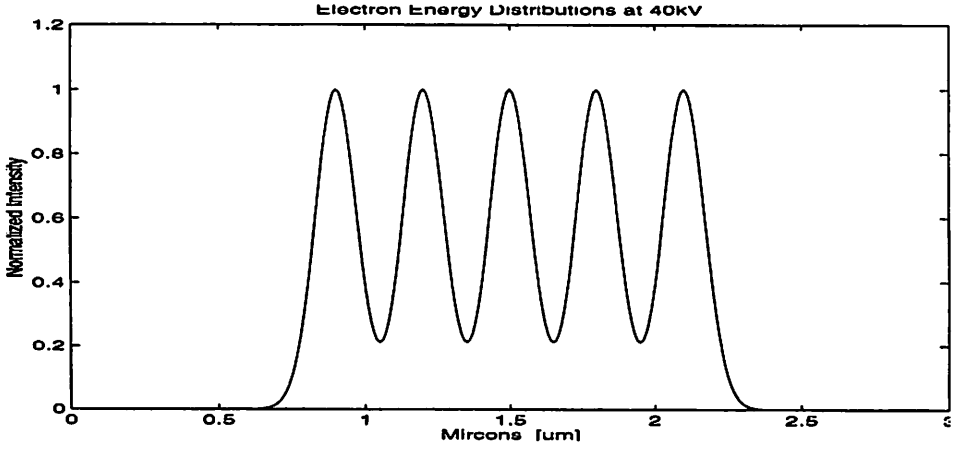
scattered gaussian half-width (σ_f) decreases at higher energies and the backscattered half-width (σ_b) increases. We should therefore expect to achieve higher resolution with higher electron energies. This figure also illustrates the fact that the line spacing pitch will have an effect on the resolution of closely spaced lines. This variation in exposure dose for closely spaced features is known as inter-proximity effect



a) Energy Distribution at 20kV



b) Energy Distribution at 30kV



c) Energy Distribution at 40kV

Figure 1. Forward-solid, Backward-Dotted, Total-dashed

This experiment is designed to study both types of proximity effects: Intra-proximity and inter-proximity. The major factor affecting the intra-proximity effect is the choice of electron beam writing energy and exposure dose. This experiment will study three different energies (20, 30, and 40kV) as well as three dose levels. Inter-proximity effects are seen as CD variations between closely spaced patterns. To characterize this variation, I propose to study three different line spacing pitches.

Process technologists who develop e-beam processes are interested in how certain variable process parameters will affect the lithographic performance capabilities. Process variables such as resist thickness and dose might be expected to affect the profiles of the patterns developed in the resist. This proximity effect will tend to be more noticeable for smaller drawn linewidth dimensions.

2.0 Methodology

A 3^5 full factorial experiment was conducted on the Jeol 6400 Modified SEM located in the Berkeley Microfabrication Laboratory. The five variables of writing energy (kV), resist thickness, drawn linewidth, linewidth pitch, and exposure dose were explored. Four of the 5 variables were run at three levels. A matrix of doses was run over a wide range to permit feature of different sizes and pitch to all clear at some dose. Three linewidth measurements were then taken at three doses around the critical dose for the correct linewidth development.

Constraints of time and effort would preclude exposing 3^5=243 different wafers. Fortunately, due to the nature of e-beam lithography it is possible to automate the variation of exposure dose, linewidth, and pitch so that several samples could be printed under varying conditions on the same substrate. The only time constraint was on the measurement of the resulting linewidth dimensions.

It is important in any experiment to preserve the initial assumption that our results will be IIND. Therefore special care must be taken during the processing; i.e. exposure, development, Au coating, and measurement. To insure the II portion of our IIND assumption, it is necessary that all process steps are repeated identically under identical conditions for the different resist thickness samples.

3.0 Implementation

Three wafers were coated with different thicknesses of 950K molarity PMMA. The thickness of each wafer was then measured on a Nanospec ellipsometer in four locations on the periphery of the wafer as well as the center. The average values of the thickness were measured to be 1670, 2233, and 3023A with standard deviations of 17.8, 21.8, and 9.7A respectively. Each wafer was then cleaved into 10mm chips that were used for the e-beam exposure.

The experiment was conducted using a test matrix pattern designed using CAD software. The test pattern consisted of 10μm long lines drawn at three different line spacing pitches. Figure 2a shows the basic test pattern consisting of lines drawn at equal line space (L/S) pitch, one-half that spatial frequency (1/2 pitch), and again at one-fourth the spatial frequency of the original (1/4 pitch). This pattern was then repeated at 5 different linewidth dimensions: 0.2μm, 0.15μm, 0.1μm, 75nm, and 50nm. Although this variable is now at five levels, the last two smallest features are near the resolution limit of the system and are not expected to clear. Figure 2b shows the

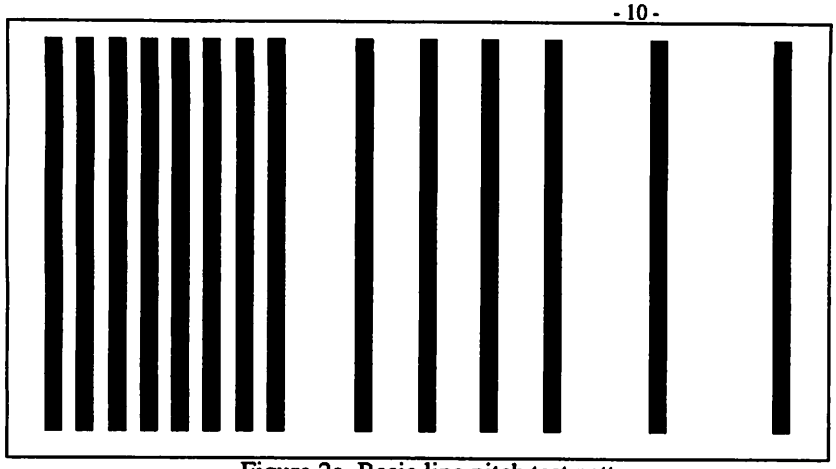


Figure 2a. Basic line pitch test pattern

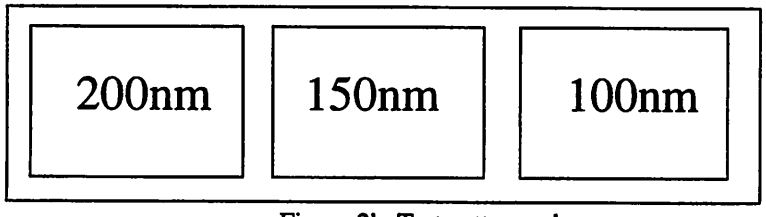


Figure 2b. Test pattern subgroup

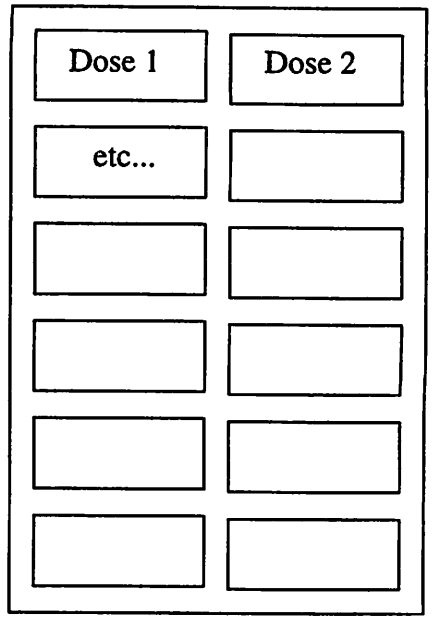


Figure 2c. Full test pattern

test pattern subgroup for the three largest linewidths. The final test pattern written consisted of the subgroup test pattern in figure 2b written 12 times (2 columns, 6 rows) at varying dose with 20μC/cm^2 steps; see figure 2c. With a step size of 20μC/cm^2, there could be a measurement error of at most 10μC/cm^2 which is at most a 1% error and at best an error of 0.2% from the resulting data set.

The field size of the 12 dose pattern was 95μm. This pattern was repeated twice at each kV setting on each of the three chips for a total of 24 different dose settings per kV per chip. This test pattern was used to expose each of the three wafer chips described above.

Beam setup and shape optimization is critical to the effective patterning of fine features. Therefore, to insure identical/independent portion of our IIND condition, it is necessary that all exposures are done under the same settings. All three chips were loaded on the same sample holder and exposed during a single system pump-down. Changing the beam energy will affect the shape of the electron column and therefore it is necessary to re-optimize the beam between changes in kV. All the patterns for a given kV and dose range were written on each of the three chips before the electron energy was set to a new value.

All three samples were developed at the same time under the identical conditions. The development was done for 60 seconds in a 3:1 mixture of Isopropyl Alcohol(IPA):Methyl Isobutyl Ketone (MIBK). The development was immediately followed by a 20 seconds rinse with IPA and then a final rinse for an additional 20 seconds with DI water. Following development, all three samples were coated with approximately 100A of Gold to aid in the linewidth measurements. The samples were then measured on the same SEM in which they were exposed.

4.0 Results and Discussion

Very few of the 50 and 75nm lines cleared (or remained for the case of equal L/S patterns). Few of these fine features printed and only at the highest kV setting. Therefore the data on these two drawn linewidths was not analyzed. This means that the variable of drawn line will be studied at three levels. This fact will be discussed later as relates to the results of the analysis. Neglecting this data, a total of 195 linewidth measurements were taken.

The dose to clear each feature varied greatly for different combinations of the other four variables. This resulted in measured linewidth data that is no longer perfectly orthogonal. Also there was not always three separate doses that resulted in patterns that cleared and remained in tact for every combination of variables. Nevertheless, the data can still be analyzed with the use of linear regression. Using the statistical software JMP, a model was fit to the measured linewidth data using linear regression. This model consisted of only the 5 first order terms. Table 1 details the fit of the model. Parameter estimates are given as well as the significance of each effect. It is clear that linear regression resulted in an excellent model fit to the data. It can also be seen, either from the t-ratio or the F-ratio, that the most significant parameter modeled here is the drawn linewidth with an F-ratio of 373.95. It is to be expected that the most important parameter in determining the dimension of the resulting line is what dimension it was drawn. Ideally, a model fit with drawn linewidth leverage should have a slope of one. From table 1 we see that the coefficient is 1.16. This can be attributed to intra-proximity effects since each line in this experiment was drawn with multiple electron beam passes per single drawn line. The backscattered electron exposures are cumulative for each pass so the larger lines that require more beam passes will tend to be overexposed and result in "blooming" of the lines.

The next most significant variable is the exposure dose with an F-ratio of 134.88. The exposure dose determines the total number of electrons that are deposited into the resist. As mentioned earlier, dose played a critical role in determining which features cleared and remained in tact. Electron energy was the next most important variable with an F-ratio of 111.60. At higher electron energies, electrons are more likely to penetrate through the resist and stop in the silicon substrate. These electrons do not provide exposure dose to the resist and therefore do not aid in the scissoring action in the positive photoresist. Finally we see that resist thickness is much less significant but still almost three times as important as linewidth pitch.

Response: MeasLW

Summary of Fit

Rsquare	0.682403
Root Mean Square Error	0.029272
Mean of Response	0.162154
Observations (or Sum Wgts)	195

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0270409	0.0144	1.88	0.0620
kV	-0.005397	0.00051	-10.56	0.0000
tresist	-0.00003	0	-5.87	0.0000
DrawnLW	1.1602101	0.06	19.34	0.0000
Pitch	0.0355368	0.00963	3.69	0.0003
Dose	0.0006061	0.00005	11.61	0.0000

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
kV	1	1	0.09562502	111.6031	0.0000
tresist	1	1	0.02956206	34.5016	0.0000
DrawnLW	1	1	0.32041484	373.9533	0.0000
Pitch	1	1	0.01166108	13.6095	0.0003
Dose	1	1	0.11557207	134.8831	0.0000

Table 1. Linear regression model fit data

Figure 3 shows a plot of the whole-model test plotted along with the 95% confidence limit curves, the Analysis of Variance (ANOVA) table for the model, and finally a plot of the residuals of the model. The model has an F-ratio of 81.22; clearly an excellent fit.

Clear groupings of the data are present. These are seen as horizontal lines on the whole-model test and as diagonal lines on the plot of the residuals. This can be explained by the fact that during the measurement, the linewidth measurements were discretized into steps of 25nm due to the resolution limits of the SEM. The lines in the Whole-Model test are also in increments of 25nm.

Figure 4 plots the linear regression model and the data vs. 4 of the 5 experimental variables. The plot of linewidth vs. kV shows a decrease in developed CD as the energy is increased. This

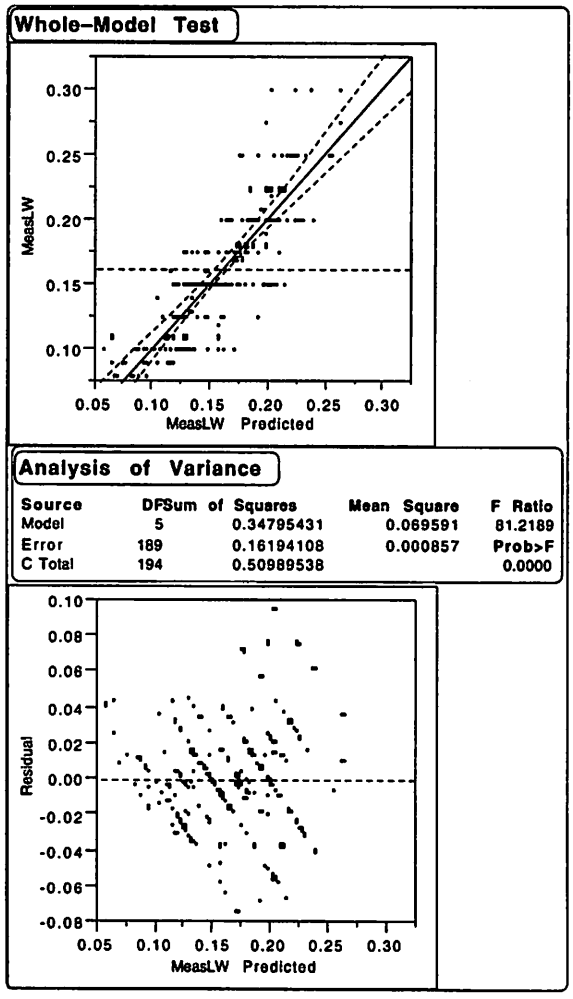


Figure 3. Linear regression model and plot of model residuals

provides us with a measure of the combined intra and inter-proximity effects. As the electron energy is increased, there is less scatter in the resist and substrate which leads to a more narrow energy deposition profile. The plot vs. the leverage of resist thickness shows a slight negative slope, indicating the minor effect of this variable. Extrapolating the linear fit, we find that it would require an extra ~0.6μm of resist to equal the effect of increasing the writing energy by only 10kV. The plot of measured vs. drawn linewidth shows a positive slope of almost one, as mentioned earlier. Finally the plot of model vs. line pitch provides a measure of the inter-proximity effect. This plot demonstrates the tendency for closely spaced lines to blur together. This is the result of the summation of the forward and backward scattered electrons from each of the nearby lines. This

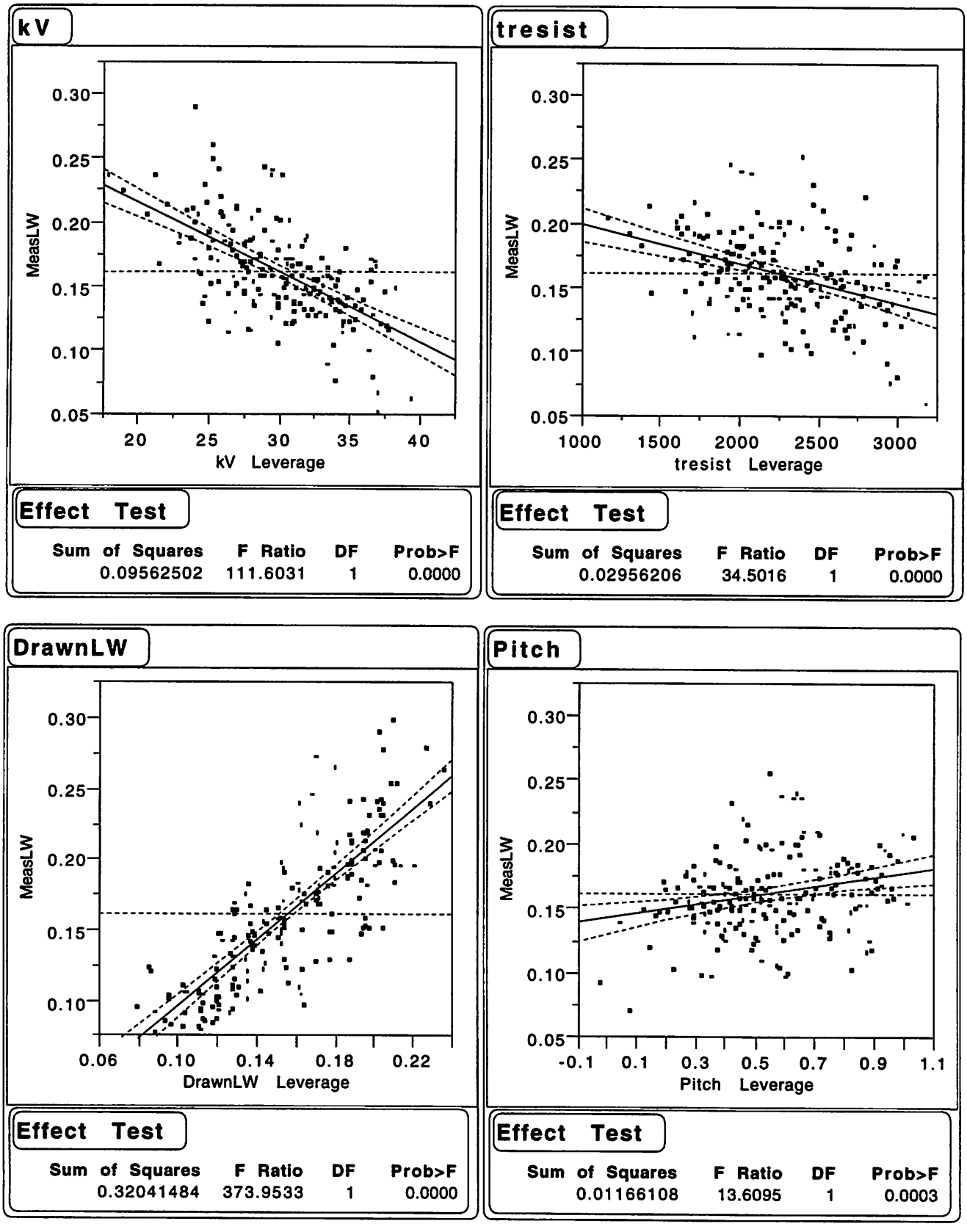


Figure 4.

Linear regression model fit and data vs. experimental variables

leads to overexposure and widens CDs upon development. Figure 5 plots the model fit and data vs. dose The steep slope of the line confirms our earlier analysis that exposure dose is the most significant factor, aside from drawn linewidth.

Table 2 lists the correlation coefficients for second order interactions between the main effects. The highest correlation is seen between dose and the other variables; confirming the conclusion that dose is the most significant effect besides drawn LW. Investigation into 3rd, 4th, and 5th order effects revealed no significant linear cross-interactions. Further analysis using models involving non-linear models is recommened.

At the outset of this experiment it was suspected that both exposure dose and electron energy played a key role in determining developed feature sizes. However, it was not obvious which played the larger role. Now it is clear that exposure dose carries more weight in the final outcome. Prior to this experiment, it was also incorrectly thought that the thickness of the PMMA layer was critical to achieve ultimate resolution. It is now apparent that varying the thickness of the photoresist layer would not have as much affect on the proximity effects and the developed linewidth as

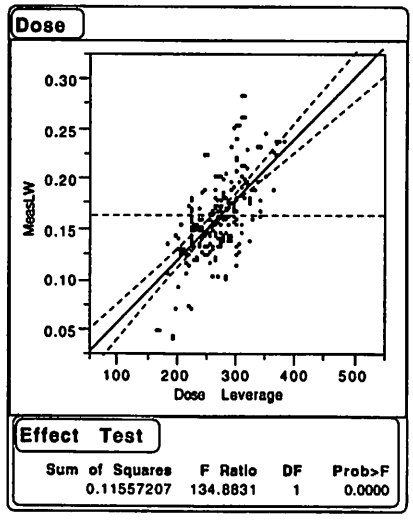


Figure 5.
Model fit vs. dose

varying the energy of the electron source. It can be seen from the data that CD distortion resulting from proximity effects can be reduced by the use of a higher kV source.

5.0 Conclusion

The main factors affecting the resulting linewidths in electron-beam lithography were analyzed using linear regression. Regression was used to fit a model to measured linewidth data from a 3^5 factorial experiment. This model contained the five variables of exposure dose, electron energy, line pitch, resist thickness and drawn linewidth. The results of the analysis point to exposure dose as the most important factor, after drawn linewidth, in determining the developed CD.

Correlations

Variable	kV	tresist	DrawnLW	Pitch	Dose
kV	1.0000	-0.0179	-0.0226	0.0277	0.6968
tresist	-0.0179	1.0000	0.0653	0.0185	0.3431
DrawnLW	-0.0226	0.0653	1.0000	0.1814	-0.2746
Pitch	0.0277	0.0185	0.1814	1.0000	-0.3881
Dose	0.6968	0.3431	-0.2746	-0.3881	1.0000

Table 2.
Second-order Cross-term interaction correlation coefficients

From the analysis we obtained a measure of both the inter and intra-proximity effects. Electron beam writing energy was found be nearly as significant as dose and played the major role in reducing proximity effects by reducing the forward-scattered distribution width. As mentioned earlier, the only 50nm and 75nm lines that did not wash out, for the case of equal L/S pitch, were the ones written at the highest electron energy. This provides us with another measure of the inter-proximity effects.

Although line pitch plays an important role in determining the magnitude of the inter-proximity effects, it was found that it is possible to overcome this effect by the proper choice of exposure dose. In fact current research is under way to find efficient methods for the calculation of dose corrections to compensate for proximity effects [9,10,11,12]. The thickness of the photoresist layer (PMMA) had comparatively little effect with respect to proximity effects and therefore the resulting developed linewidths. The latter two are interesting results and were not anticipated.

References

[1] Broers, A. N., "Resolution limits for electron-beam lithography", IBM Journal of Resist Development, vol. 32, no 4, July 1988, pp.502-513.

[2] Wei C. and Ahmed H., "Fabrication of high aspect ratio silicon pillars of <10nm diameter", Appl. Phys. Lett.62 (12), 22 March 1993, pp. 1116-1120.

[3] Fischer P. B. and Chou S. Y., "Sub-50 nm high aspect-ratio silicon pillars, ridges, and trenches fabricated using ultrahigh resolution electron beam lithography and reactive ion etching", Appl. Phys. Lett.62 (12), 22 March 1993, pp. 1414-1416.

[4] Gentilli M., Grella L., Di Fabrizio E., Luciani L., Baciocchi M., Figliomeni M., Figliomeni M., Maggiora R., Cerrina F., and Mastrogiacomo L., "Development of an elctron-beam process for the fabrication of x-ray nanomasks", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2938-2942

[5] Rosenfield M. G., Thomson R., Coane P. J., Kwietniak K. T., Keller J., Klaus D. P., Volant R. P., Blair C. R., Tremaine K. S., Newman T. H., and Hohn F. J., "Electron-beam lithography for advanced device prototyping: Process tool metrology", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2615-2620.

[6] Nakayama Y., Okazaki S., Saitou N., and Wakabayashi H., "Electron-beam cell projection lithography: A new high-throughput electron-beam direct-writing technology using a specially tailored Si aperture", J. Vac. Sci. Technol. B 8(6), Nov/Dec 1990, pp1836-1840.

[7] Fischer P. B., Dai K., Chen E., and Chou S. Y., "10nm Si pillars fabricated using electron-beam lithography, reacitve ion etching, and HF etching", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2524-2527.

[8] Chen W., and Ahmed H., "Fabrication of sub-10 nm structures by lift-off and by etching after electron-beam exposure of poly(methylmethacrylate) resist on solid substrates", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2519-2523.

[9] Eisenmann H., Waas T., and Hartman H., "PROXECCO-Proximity effect correction by convolution", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2741-2745.

[10] Dobisz E. A., Marrian C. R. K., Salvino R.E., Ancona M.A.,F.K. Perkins, and Turner N.H., "Reduction and elimination of proximity effects", J. Vac. Sci. Technol. B 11(6), Nov/Dec 1993, pp2733-2740.

[11] Bojko R. J., Hughes B.J., "Quantitative lithographic performance of proximity correction for electron-beam lithography", J. Vac. Sci. Technol. B 8(6), Nov/Dec 1990, pp1909-1913.

[12] Owen G., "Methods for proximity effect correction in electron lithography", J. Vac. Sci. Technol. B 8(6), Nov/Dec 1990, pp1889-1892.

A Model for The Formation of Thin Film Anodized Aluminum

Amy Wang

Models have been developed to predict the thickness of electrochemically-formed aluminum oxide films and the effect of anodic films on a solid-state acoustic sensor given film anodization conditions. A method of defining a precise anodization area using photolithography was developed to allow anodization on the chip level, simplifying data collection. Anodization variables investigated include anodization voltage, electrolyte concentration, post-anodization soak time, and limiting current of the power supply.

1.0 Introduction

Porous anodized aluminum coatings are commonly used in a number of commercial applications as decorative, corrosion-resistant, or insulating layers. We would like to use these porous films as a method of increasing the surface area on the active region of a mass sensor. Many studies have reported the results of specific anodization conditions on films; however, it is difficult to find comprehensive studies that explore the effects of a number of different anodization conditions on film characteristics.

Ultimately we would like to measure porosity or surface area of the anodic films and develop a model to maximize surface area through anodization conditions, but first we must investigate the feasibility of coating a solid-state device with an anodic film on the micro-level and characterize how addition of the film affects sensor response.

The device used in these experiments is the Flexural Plate-wave (FPW) sensor, which is a solid-state, acoustic sensor. The key sensing element of the device is a thin plate through which acoustic waves propagate. The velocity of these waves depends on the material characteristics of the film. A thin film deposited on the plate alters the material properties of the plate resulting in a change in acoustic wave velocity from that of a bare device. This velocity change can be detected through a shift in resonant frequency of the device. The frequency shift measured can also be used to extract qualitative information about the effects of anodization conditions on the material characteristics of the film.

2.0 Methodology

Two experimental designs were implemented. The first was an L9 orthogonal array with four anodization variables (Table 1). Each input had three levels that were varied over the experimental space. Response variables measured were film thickness and the shift in resonant frequency of the sensor. Replicate runs were included to measure the experimental error. The second design measured film thickness alone as a function of two anodization parameters, voltage and electrolyte concentration. A full factorial experiment was performed with replicate runs. Input levels

were centered around typically cited values in the literature, and the order of experimental runs and location of film formation on the wafer were randomized.

voltage, V (V)	conc, C (%)	soak time, S (min)	current, I (mA)	thickness, t (Eq. 1) (A)	freq. shift, (kHz) (Δf)
10	15	0	50	4517	-140
20	15	15	150	4769	-150
10	5	30	150	3714	-170
10	5	30	150	3714	-150
15	5	15	50	3840	-50
15	10	0	150	4241	-140
15	15	30	100	4643	-190
20	10	30	50	4367	-60
10	10	15	100	4115	-160
10	10	15	100	4115	-150
20	5	0	100	3966	-20
20	5	0	100	3966	-50

TABLE 1. L9 Orthogonal array for anodization variables to predict sensor frequency shift.

Anodizations were performed under constant voltage conditions. This method was deemed to be more likely to give reproducible film thicknesses since the film simply grows until the voltage drop across the film thickness is not sufficient to further the oxidation reaction. Thus, the first input variable chosen was anodization voltage. The second variable was electrolyte concentration, which is known to affect film thickness and morphology. A sulfuric acid electrolyte was used because it is the most commonly studied electrolyte that yields porous films. The last two variables were not expected to affect film thicknesses, but to control aspects of the film morphology. A post-anodization soak in 10% H_2SO_4 solution was included to determine if further contact with the electrolyte changed the density of the film (increased porosity). A current limit was also set on the power supply to limit the large initial current flow that occurs during constant voltage anodization. This could have possible effects on the morphology of the film during initial growth. Experiments in which only voltage and electrolyte concentration were varied used a 100 mA current limit and no post-anodization soak.

3.0 Implementation

Anodization Technique. A method of spot anodizations on individual dies was developed to form thin film anodized aluminum in lithographically specified areas on a wafer. For the first experiment, aluminum was sputtered on a wafer with fabricated FPW devices. Photoresist was spun on and patterned, exposing the active area of the sensor. The exposed aluminum was then anodized under different conditions. In the second experiment, aluminum was anodized on a bare silicon wafer to enable a more accurate measurement of film thickness. A contact area was opened in the resist for electrode contact to the wafer (Fig.1). Since the areas were defined lithographically, the anodized area and electrode geometry were repeatable for each experiment.

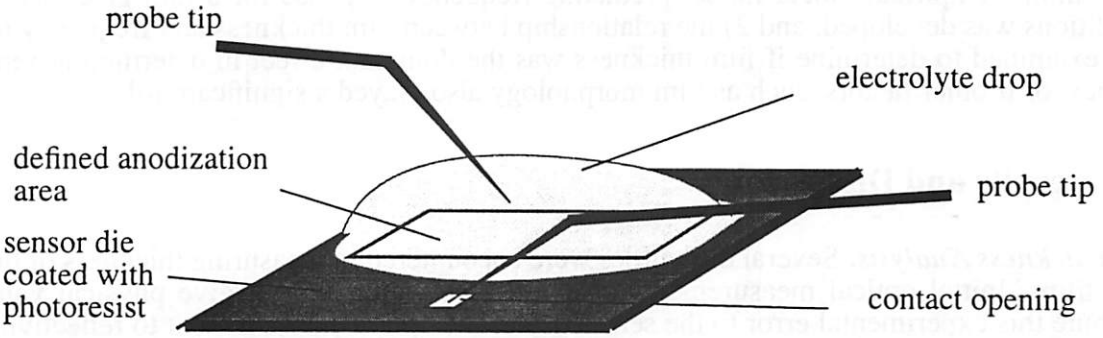
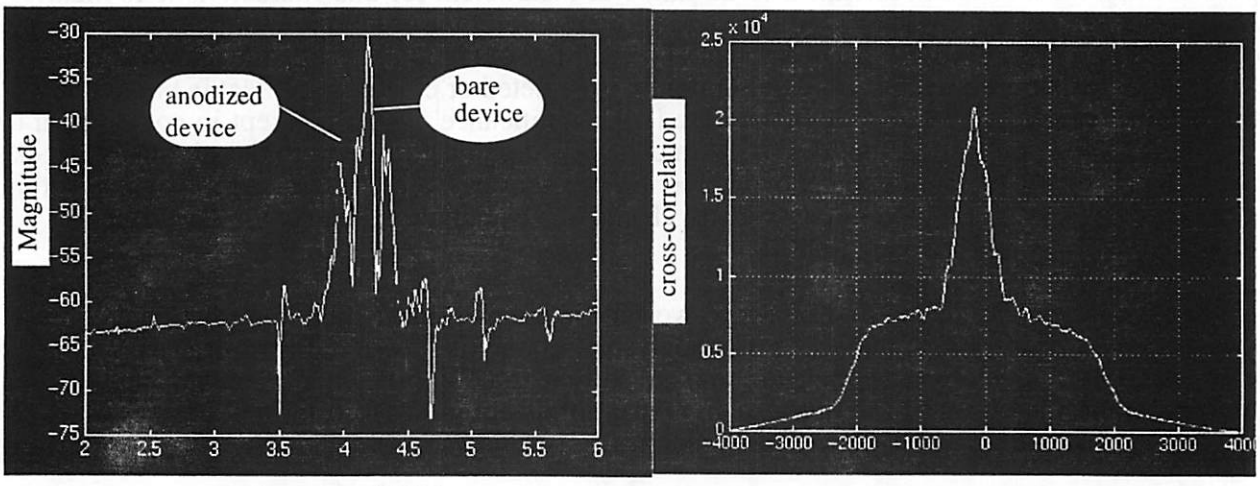


FIGURE 1. Experimental set-up for anodization on a single die.

Anodizations were performed at a probe station. A drop of electrolyte of desired concentration was dropped over the exposed aluminum area. Surface tension and the physical barrier of the photoresist was sufficient to keep the drop confined to a single die. One probe contacted the defined contact area on the die and the other was lowered into the electrolyte drop, completing the current path. A *Tektronix* PS 5010 programmable power supply was used as a constant voltage source. The current limit on the supply was also varied as an input parameter. Current was monitored during the anodization process using a multimeter. Zero current indicated completion of the anodization. With this method, a matrix of anodized films with different anodization conditions was formed on a single wafer for evaluation of film thickness and sensor response.

Film Measurements. Film thickness was measured optically using the NanoSpec film thickness monitor. An index of refraction was assumed from literature values [1]. The frequency spectrum of each device before and after anodization was measured using an HP 4195 Network Analyzer and was captured using *LabView* [2]. The shift in resonant frequency between the two spectra (Fig. 2) was determined by calculating the cross correlation between the two spectra on *Matlab* [3]. The frequency lag yielding the peak correlation was taken to be the frequency shift.

FIGURE 2. Shifted frequency spectrum and cross-correlation between the two spectra.



The JMP [4] software was used to analyze the data and generate predicting models. A linear model was expected and confirmed for film thickness as a function of anodization parameters, and significant effects were determined. Two aspects of the frequency measurement were examined: 1) a simple, empirical, linear model predicting frequency response for a film given anodization conditions was developed, and 2) the relationship between film thickness and frequency response was examined to determine if film thickness was the dominant effect in determining sensor frequency, or if other factors, such as film morphology also played a significant role.

4.0 Results and Discussion

Thickness Analysis. Several difficulties were encountered in measuring thickness of the anodized films. Initial optical measurements on the device wafer did not give physical values. We attribute this experimental error to the sensitivity of the optical measurement to reflectivity of the underlying surface, which was unknown. Analysis of the frequency data showed anodization voltage and electrolyte concentration to be the most significant effects on sensor response, as was expected. Consequently, a second experiment was performed varying only voltage and concentration conditions. These anodizations were formed on a bare silicon wafer so that the underlying layer would have a known reflectivity and accurate thickness measurements could be taken.

A linear model was sufficient to predict film thickness. Physically, film thickness increases with the amount of current that is passed during the anodization process [5]. Thus, increased anodization voltage and electrolyte concentration both result in thicker films. Final results gave the following model,

$$\begin{aligned} \text{Thickness} &= 3128 + 25.2\,V + 70.1\,C. \\ &\quad (\pm 260)\ (\pm 14.3)\ (\pm 12.7) \end{aligned} \tag{1}$$

JMP analysis is shown in Figure 3. The results show insignificant lack of fit. Likely reasons for the non-reproducibility of film thicknesses are measurement error in assuming a constant index of refraction for all optical measurements. A better method would be to use a technique that allows measurement of index of refraction as well as thickness. There is also film non-uniformity over the area anodized causing variance in thickness measurements. This is a disadvantage of the spot anodization technique, which does not allow mixing of the electrolyte. Residual analysis shows the error to be random.

It is unexpected that the anodization voltage does not play a more dominant role in determining the thickness of the film (F Ratio = 3 vs. F Ratio = 30 for the concentration parameter). It is possible that the aluminum thickness (0.25 μm) was insufficient for this characteristic to manifest itself to its full extent. In other words, in an infinitely thick film of aluminum, the oxidation process would proceed until the barrier potential for further oxidation exceeds the anodization voltage, but in this situation, the supply of aluminum is depleted or electrical contact is broken before this point is reached. This would explain the large significance of the intercept in comparison to the other effects, indicating that the oxide thickness is dominated by the initial aluminum thickness. However for the applications of interest for the sensor, we are interested only in very thin alumina films so that they do not drastically change the characteristic device response and so we subsequently only studied very thin films.

Sensor Response. In the first analysis, we form a simple, linear model predicting frequency shift for given anodization conditions. The analysis summary is shown in Fig. 4. It is interesting to note that although voltage and concentration are still the most significant effects, post-anodization soak time and the current limit are also significant, indicating that film morphology also affects the observed frequency shift. The model predicted is as follows,

$$\Delta f = -63.6 + 6.83\,V - 7.82\,C - 0.66\,I - 0.99\,S \quad (2)$$
$$\langle \pm 21.3 \rangle \; (\pm 0.89) \; (\pm 0.93) \; (\pm 0.10) \; (\pm 0.31)$$

The lack of fit in this model was good and the intercept term is less significant than in the thickness measurements. Potential experimental error lies in the lack of resolution of capturing the frequency spectrum on the network analyzer. Only a shift of 10,000 Hz could be resolved. A better method of measurement would be to oscillate the device and measure frequency on a counter with better resolution; this, however, would make data acquisition decidedly more difficult.

In the second analysis we relate film thickness predicted from Eq. 1 with frequency shift measurements to determine if the observed response was primarily a function of film thickness. This would indicate that the dominant cause for frequency change was due to the change in flexural rigidity of the plate (a function of Young's modulus of the film).

The fact that the frequency shifts are all negative suggests a reduction in Young's modulus of the film, which is likely since pure aluminum is much stiffer than aluminum oxide [1]. In other words, as the aluminum film is anodized and converted to aluminum oxide, the composite Young's modulus of the sensor plate is reduced. An upward frequency shift would indicate a reduction in density of the film. If one effect were dominant of the other, the shift in frequency would increase with film thickness. However, we observe the reverse effect, which leads us to believe that as the film grows, the density effect becomes more significant and results in an upward frequency shift.

Fig. 5 shows plots of frequency shift versus voltage for constant electrolyte concentrations. A very simple physical model relating frequency shift to film thickness has a parabolic relationship. Although there are not enough data points to accurately fit a model for the relationship, it is interesting to note that the concavity of the curves change as concentration increases. This shows that there is a critical point where density effects become more prominent over the reduction in composite plate rigidity. Indeed, studies have shown that increased electrolyte concentration reduces density of the film and increased thickness also results in reduced density [6]. The density is inversely proportional to the sensor frequency; thus, the reduced density with increasing thickness agrees with the upward frequency trend shown in plots (a) and (b). However plot (c) shows a different behavior that exhibits an inflection point. A possible interpretation of this data is that this point reflects where density effects become significant, resulting in an upward frequency shift. However there are really not enough data points to make this assertion.

5.0 Conclusions

Models were developed for predicting thickness of an anodized aluminum film and the effect of the film on resonant frequency of a sensor for given anodization conditions. A method of performing anodizations on individual sensors was developed and implemented. The models agree with physical behavior observed in anodic films studied in literature references. These models enable us to characterize sensor response and choose desired anodization conditions when the sensors are used for mass detection.

Response: Thickness

Summary of Fit	
Rsquare	0.808512
Root Mean Square Error	174.9795
Mean of Response	4271.636
Observations (or Sum Wgts)	11

Lack of Fit				
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	6	104870.06	17478.3	0.2496
Pure Error	2	140072.67	70036.3	Prob>F
Total Error	8	244942.73		0.9215

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3128.5714	260.378	12.02	0.0000
voltage	25.2	14.287	1.76	0.1158
concentration	70.130952	12.6641	5.54	0.0005

Effect Test					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
voltage	1	1	95256.00	3.1111	0.1158
concentration	1	1	938957.82	30.6670	0.0005

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1034213.8	517107	16.8891
Error	8	244942.7	30618	Prob>F
C Total	10	1279156.5		0.0013

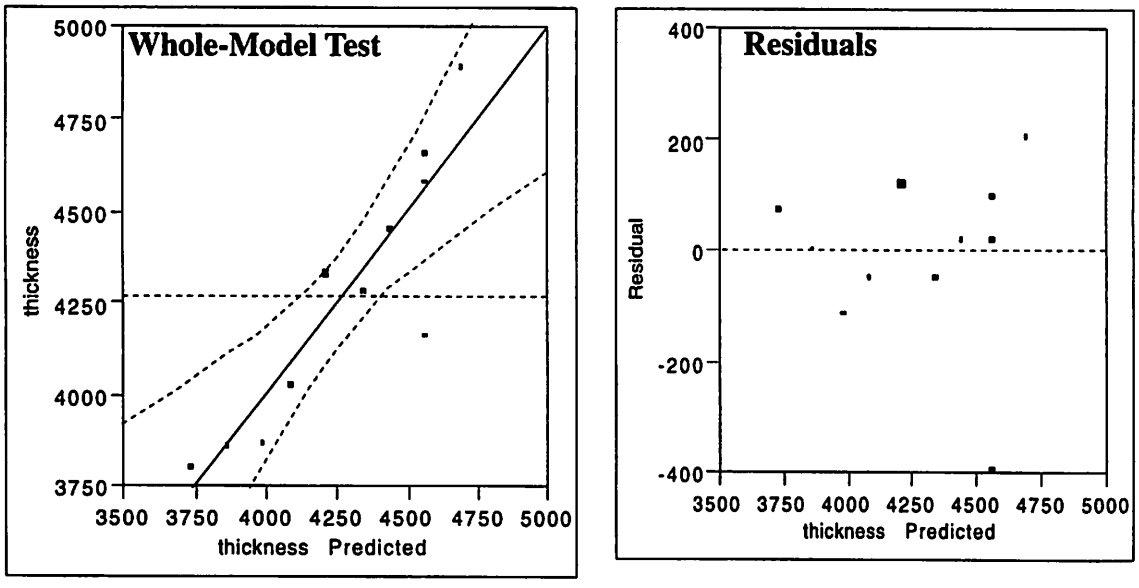


FIGURE 3. JMP analysis for thickness model.

Response: Frequency

Lack of Fit				
Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	4	454.3307	113.583	0.4868
Pure Error	3	700.0000	233.333	Prob>F
Total Error	7	1154.3307		0.7510

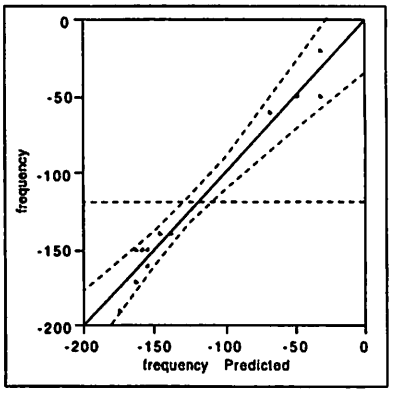
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-63.70079	21.2918	-2.99	0.0202
voltage	6.8293963	0.88917	7.68	0.0001
conc	-7.826772	0.93427	-8.38	0.0001
current	-0.657218	0.09963	-6.60	0.0003
soak time	-0.990376	0.31347	-3.16	0.0159

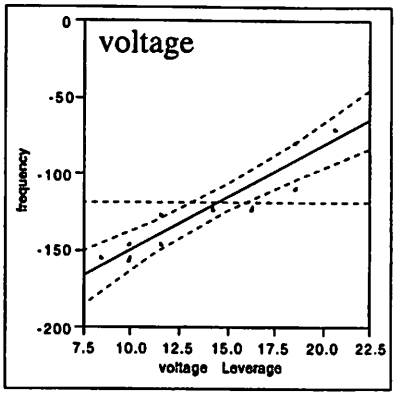
Effect Test					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
voltage	1	1	9728.151	58.9927	0.0001
conc	1	1	11573.273	70.1817	0.0001
current	1	1	7175.902	43.5155	0.0003
soak time	1	1	1645.996	9.9815	0.0159

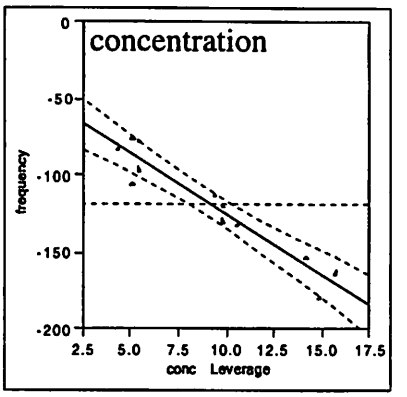
Summary of Fit	
Rsquare	0.967836
Room Mean Square Error	2.84151
Mean of Response	119.167
Observations (or Sum Wgts)	12

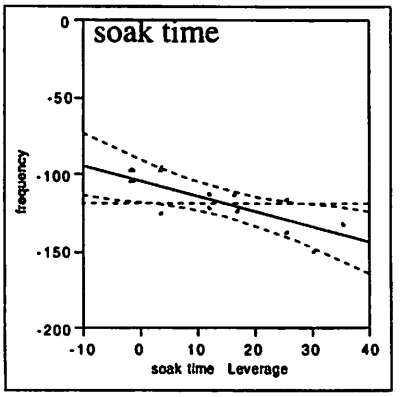
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	34737.336	8684.33	52.6628
Error	7	1154.331	164.90	Prob>F
C Total	11	35891.667		0.0000

Whole-Model Test









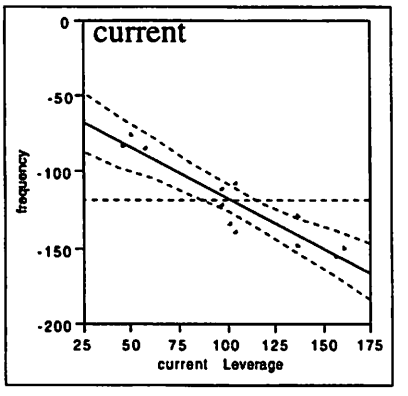
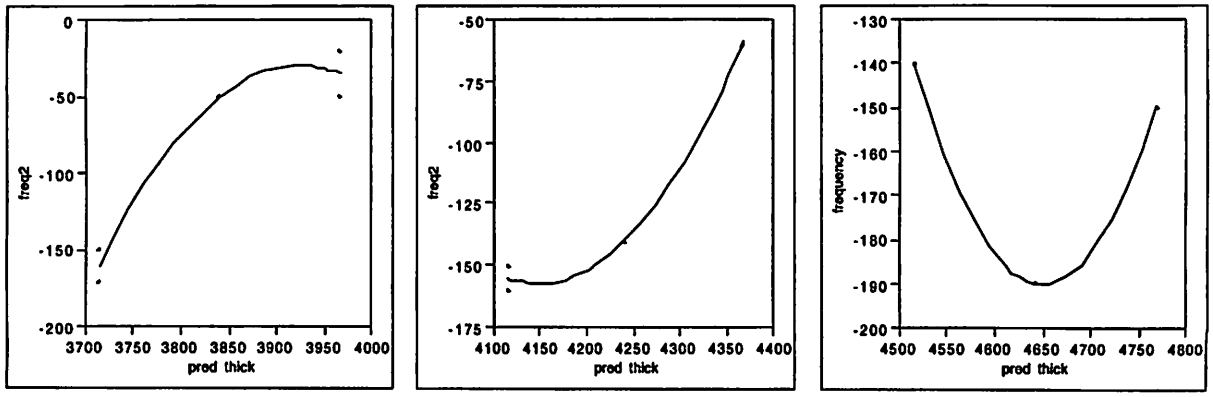


FIGURE 4. JMP analysis for sensor frequency shift.



(a) 5% H_2SO_4 (b) 10% H_2SO_4 (c) 15% H_2SO_4

FIGURE 5. Sensor frequency shift vs. thickness for fixed concentration levels.

Acknowledgments: All devices were fabricated in the Berkeley Microfabrication Laboratory. I would like to thank Chuck Bradley for his assistance with Matlab and the correlation analysis, Jim Bustillo for his help with developing the spot anodization method, Jack Judy for his file transferring programs for the Macintosh, and Tony Miranda for his help with JMP.

6.0 References

[1] K, Safranek, *The properties of electrodeposited metals and Alloys, A Handbook,* 2nd ed., Ch 3. pp. 29-32.

[2] *LabView*, version 3, National Instruments, 1993.

[3] *Matlab*, The MathWorks, Inc., 1991.

[4] *JMP*, SAS Institute, 1989.

[5] L. Young, *Anodic Oxide Films*, New York: Academic Press, pp. 193-210, 1961.

[6] Y. Fukuda, T. Fukushima, M. NagaYama, "The composition, structure and properties of anodic oxide films formed on aluminum in a 13M sulfuric acid solutions," Transactions of National Research Institute for Metals, v. 32, n. 2, pp. 29-36, 1990.

Ion Implant Damage Study Using a Factorial Design

Donggun Park

Factorial design of experiment is applied for the study of the charging damage during the high current ion implantation process. The effects of dose, pattern size, and capping oxide are selected as the experimental factors to be studied in two levels. The MOS capacitor structures which have LOCOS isolation are used for the measurement of the interface trap generation. Implantation dose is the most significant factor to the damage. The Dose of $1x10^{16}$ cm^{-2} with capping oxide 85nm, as a suggested process integration condition, dose not show any damage.

1.0 Introduction

Ion Implantation is one of the most important processes for VLSI integrated circuit fabrication because of its ability to control the precise amount of impurities. However, the ion implantation process is also very expensive. Especially the high current ion implantation process for the transistors source and drain formation takes long time and is a bottleneck process in the IC fabrication line. This process is intentionaly slow in order to prevent the devices from the charging damage during the ion implantation. A decade ago, when the charging damage was shown by studies of the oxide breakdown, the ion beam current was higher than 10mA to reduce the process time. Since then ion beam current of high current implant process has been reduced down to 1mA, and electron shower was introduced.[1] And the ion implantation is known as not only a time-consuming process but also a most expensive process. Recently some academic studies on the ion implantation charging damage were reported showing that there is a charging damage typically on the very large size of antenna ratio (ratio of gate poly pattern to the active oxide area) from the oxide breakdown data.[2,3] In this study the charging damage will be studied in terms of the interface trap generation which is more sensitive in detecting the charging damage than the oxide breakdown comparison.[4] The important factors of the ion implantation for the process integration of ICs will be discussed including the interaction effects between the factors. In addition, we will draw conclusions from the quantitative study of the effects and the possible process integration conditions, which could improve the manufacturing efficiency without any charging damage.

2.0 Methodology

A two level three factor factorial design of experiment was implemented for the study of the ion implantation charging damage study. Three test chips from the wafers were measured and averaged as the data of the experiment. The factors are ion implantation dose, antenna ratio, and capping oxide on top of the gate polysilicon. The significance was estimated by the standard error calculated from the insignificant higher order interaction effects, and also the normal probability plots were used for the estimation of the significance of the effects and residuals. The analysis of variance(ANOVA) was performed for the quantitative estimation of the significance of the effects

compared to the error. [5,6] The three factors and their levels used in the two stages are listed in Table 1.

Level	Implant Dose	Antenna ratio	Capping oxide
+	$3x10^{16}$ cm^{-2}	14:1	0 nm
-	$1x10^{16}$ cm^{-2}	1.6:1	85 nm

TABLE 1. Factors and levels for the factorial experimental design

3.0 Implementation

The p-type silicon wafers were processed with the LOCOS process for the isolation between the capacitors in the Berkeley microfabrication laboratory. The field oxide and the gate oxide thickness were 400nm and 12nm, respectively. Threshold voltage control ion implantation was applied using $1.9x10^{12}$, 60KeV, and $^{49}BF_2^+$. The deposition of in-situ n^+-doped polysilicon was followed by the deposition of 85nm oxide using CVD LTO process. Using photolithography and etching the gate polysilicon were patterned and also the capping oxides were selectively removed by HF etch. Finally, the wafers were implanted with the conditions of $^{75}As^+$, 60KeV, electron flood gun current 40mA, and ion beam current of 8mA using NOVA 10-80 high current ion implanter.

The samples were measured to get the data of the initial interface trap densities before they are sent for the ion implantation. The quasi-Capacitance Voltage measurement technique was applied to get the Low-frequency CV data. The interface trap generation of each sample was calculated from the CV curves using the relation of

$$\Delta N_{it} = \frac{1}{e}\left(\left(\frac{1}{C_f}-\frac{1}{C_{ox}}\right)^{-1}-\left(\frac{1}{C_i}-\frac{1}{C_{ox}}\right)^{-1}\right) \qquad (1)$$

where C_f is capacitance after ion implantation, C_i is the initial capacitance, C_{ox} is the capacitance of accumulation status, and N_{it} is interface trap density as a function of the surface potential.[7] The surface potential, Φ_s, is calculated using

$$\Phi_s = \int_{V_{fb}}^{V_g}\left(1-\frac{C}{C_{ox}}\right)dV \qquad (2)$$

where V_g is the gate voltage, and V_{fb} is the flat band voltage.[8]

Fig. 1 shows the test patterns which were used for this experiment. The active area used here is $80x80\mu m^2$,while the gate poly patterns are $100x100\mu m^2$ (Antenna Ratio 1.6:1) and $300x300\mu m^2$ (Antenna Ratio 14:1).

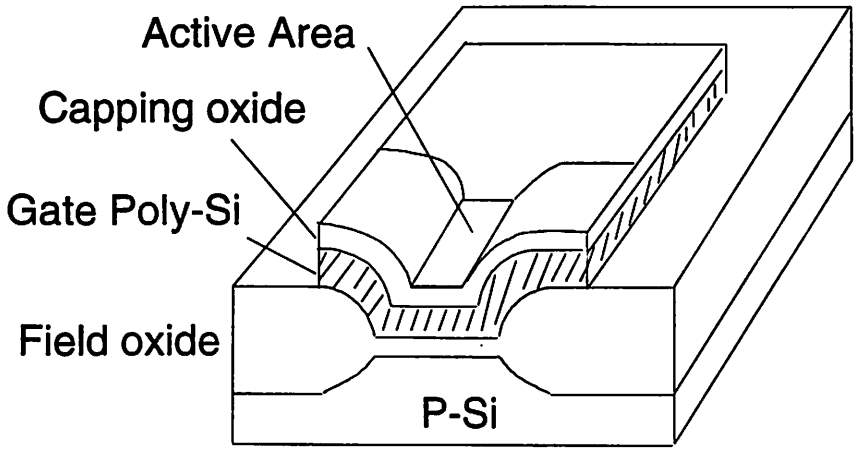
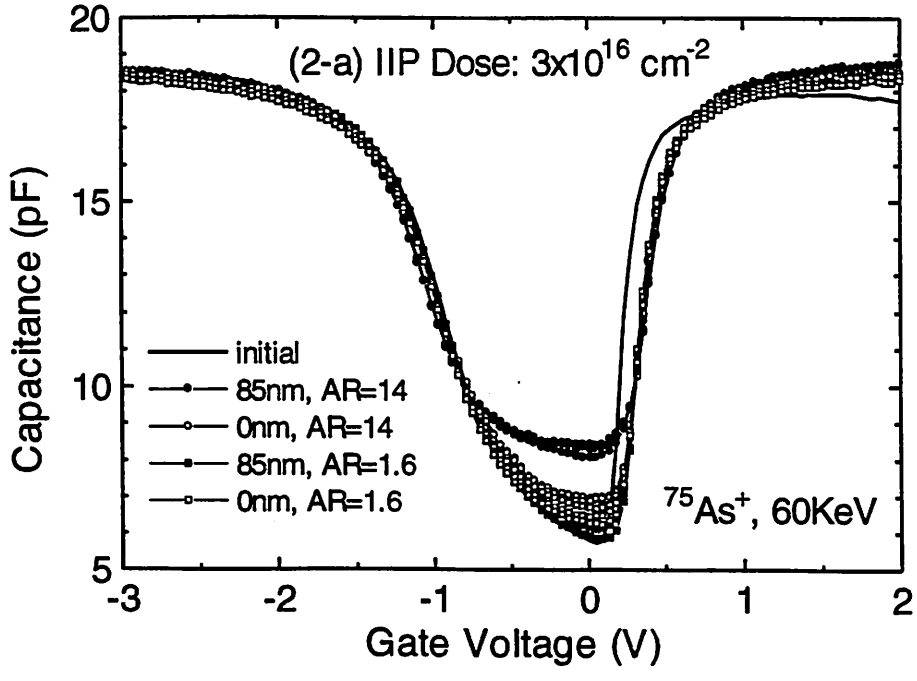


FIGURE 1. Vertical structure of the test devices. Capping oxide on top of the polysilicon gate is one of the factors to be evaluated in this experiment.

4.0 Results and Discussion

The CV curves in Fig. 2 show that the minimum capacitance increases due to the addition of the capacitance which is due to the generated interface traps. The interface traps are generated by the tunneling current passing through the thin gate oxide during the high current ion implantation charging._



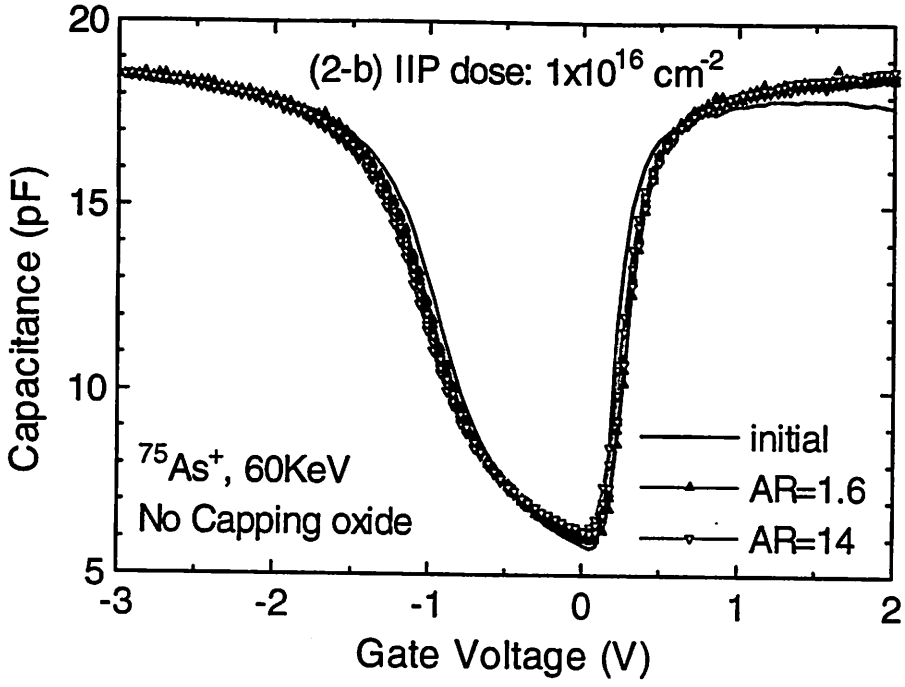


FIGURE 2. CV curves show the changes in minimum capacitance and shape due to the charging damage. Large charging damage is shown at $3x10^{16}$ cm^{-2} dose.

The data which were calculated from the equations (1) and (2) are summarized in Table 2. The effects of each factors and the interaction terms are extracted using Yate's algorithm. The estimated variance for each test conditions within wafer is 0.047, which is good enough so that we can choose the average value for the estimation of the effects. The generated interface trap density data were transformed by taking natural log to the densities because the difference between the low values and high values was very large.In addition, this transformation helped linearizing the effects. The analysis was performed using the raw data and several forms of the transformation to find that the transformation in the logarithmic scale is the most sensitive and easier to interpret. [6]

Fig. 3 shows the normal probability plots of the effects and the residuals. Only the effects of Dose, Capping oxide, and the interaction of Dose and Capping oxide are significant, as they fall far from the line passing through the other points. Therefore the model can be expressed as the equation (3),

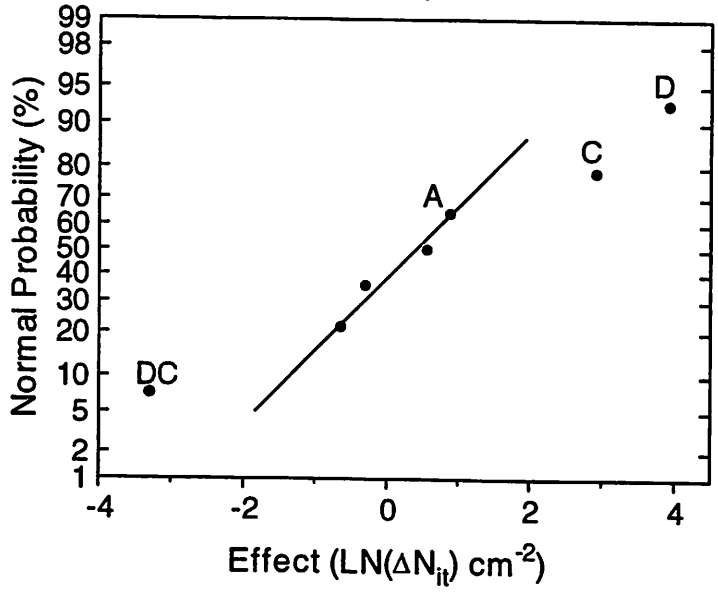
$$\hat{Y} = 23.5 + 1.97D + 1.47C - 1.65DC \qquad (3)$$

On the other hand, the normal probability plot of the residuals lie approximately along a straight line, we do not suspect any severe non-normality in the data. There are no indications of severe outliers. Also the standard error calculated with the insignificant factors is 0.636. Again, we can conclude with approximate 95% confidence intervals, that the effects of Dose, Capping oxide, and DC interaction term are important while the rest of the effects are not.

Implant Dose	Antenna Ratio	Capping oxide	log(N_{it}) (log(cm^{-2}))	(1)	(2)	(3)	Divide	Estimate	Identification
-1	-1	-1	18.42	42.87	88.13	187.97	8	23.5	AVG
1	-1	-1	24.45	45.26	99.84	15.75	4	3.94	**D**
-1	1	-1	18.42	49.31	14.45	3.62	4	0.90	A
1	1	-1	26.84	50.53	1.31	2.26	4	0.57	DA
-1	-1	1	24.29	6.03	2.39	11.71	4	2.93	**C**
1	-1	1	25.01	8.42	1.23	-13.14	4	-3.29	**DC**
-1	1	1	24.97	0.72	2.39	-1.16	4	-0.29	AC
1	1	1	25.56	0.59	-0.13	-2.51	4	-0.63	DAC

TABLE 2. Calculation of the effects using Yate's Algorithm. The estimated within-wafer variance (s^2) is 0.047, which is calculated from the data of 3 dies on a wafer.

Fig.3-a) Normal Probability Plot for the Effects



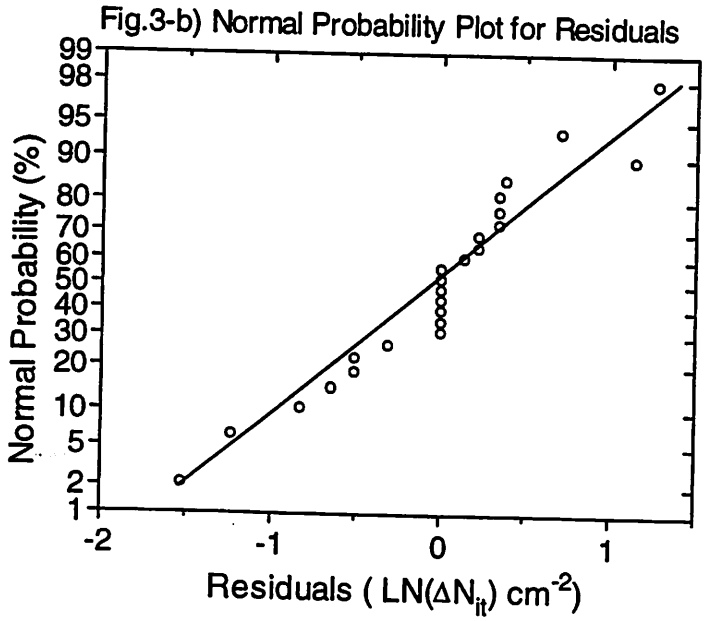


FIGURE 3. Normal Probability plots for the effects and the residuals. It is obvious that only the factors of D, C, and DC are the significant factors from the plots.

Fig. 4 shows that the large positive effect of interface trap generation occurs primarily when capping oxide is at the low level (85nm). If we use the capping oxide on top of the polysilicon with low level of dose, the charging damage can be prevented. However, if we use the higher level of dose, then either capping oxide thickness levels will provide higher trap generation levels.

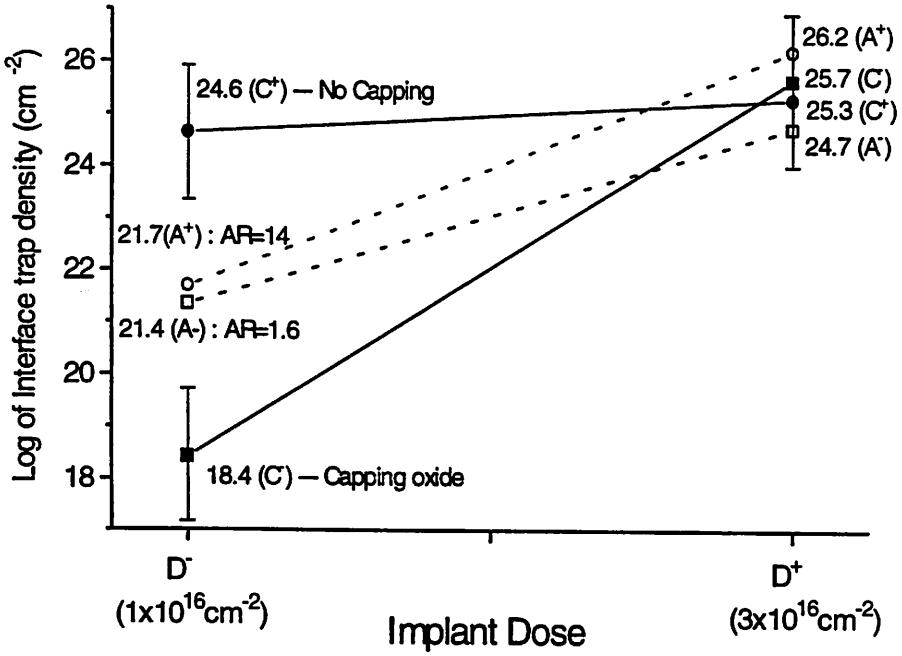


FIGURE 4. Interaction plots of the effects showing that the dose is a primary effect on the charging damage. The 95% confidence intervals of the effects are shown.

The analysis of variance summarized in Table 3 confirms the significant factors which were founded from the normal probability plot. The factors, D, C, and DC, are important even at the 1% level. ($F_{0.01,1,4} = 21.2$)

TABLE 3. Analysis of Variance for the factors of charging damage due to ion implantation

Source of Variance	Sum of Squares	Degrees of Freedom	Mean Square	F_o
D	31.02	1	31.02	38.4
C	17.13	1	17.13	21.2
DC	21.59	1	21.59	26.7
Error	3.23	4	0.81	

$F_{0.01,1,4} = 21.2$, $F_{0.025,1,4} = 12.22$, $F_{0.05,1,4} = 7.71$, $F_{0.1,1,4} = 4.54$

5.0 Conclusions

A factorial experiment was used for the study of ion implantation charging damage in terms of the interface trap density. The factorial experimental design shows obvious results as well as the confidence level. The normal probability plot of effects and residuals is a very useful tool in discovering the significant factors. The dose is primary factor to the interface trap generation. And also the capping oxide on the gate electrode will protect the gate oxide from the implantation charging damage, when the dose level is low enough. For example, with the dose of $2x10^{16}$ cm^{-2} and capping oxide 85nm, the interface trap density is increased by $4x10^{9}$ cm^{-2}, which leads the threshold voltage shift of 2.4mV. Finally, the antenna ratio levels chosen in this experiment might be too close to have significant effects on the output.

References

[1] M. E. Mack, *Nuclear instruments and Methods in Physics Research*, B37, 472, 1989

[2] V. K. Basra and C. M. Mckenna, *Nuclear instruments and Methods in Physics Research*, B21, 360, 1987

[3] M. G. Stinson and C. M. Osburn, *Journal of Applied Physics*, 67, 4190, 1990

[4] K. Nakanishi, H. Muto, H. Fujii, S. Sasaki, H. Yamamoto, S. Matsuda, S. Kato, *Journal of Electronic Materials*, 19, 739, 1990

[5] D. C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley, 2nd ed., 1991

[6] G. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experiments*, Wiley Interscience, 1978

[7] E. Nicollian and J. Brews, *MOS Physics and Technology*, John Wiley, 1982

[8] C. N. Berglund, *IEEE Trans. on Electron Devices*, 13, 701, 1966

Modeling Plasma-Immersion-Ion-Implantation By A Response Surface

Jiang Tao

In this project, a Response-Surface Method has been used to study and model the dependence of ionization efficiency of O+ and O_2+ ions in oxygen plasma on experimental variables (microwave power, oxygen flow rate and magnet current). Through this analysis, an optimized plasma condition has been found which can be used in the SIMOX formation process and to solve the multiple ion implantation problem associated with the Plasma-Immersion-Ion-Implantation (PIII) system.

1.0 Introduction

Recently, extensive research efforts have been focused on SOI (silicon-on-Insulator) technology mainly because SOI CMOS technology is much simpler than bulk CMOS technology and the absence of latch-up, the reduced parasitic source and drain capacitance. In addition, the ease of making shallow junctions will extend the CMOS performance limits beyond scaling limits in bulk CMOS [1]. SIMOX (Separation by Implanted Oxygen) was the first and leading method to provide ultra-thin silicon monocrystalline film on top of SiO_2. It has been generally believed that there is a tremendous promise for SOI technologies as CMOS scaling advances beyond 0.25μm.

The availability of low-cost, low-defect density SOI substrates with Si thin film thickness of about 1000Å is the key bottleneck in the widespread use of SOI in the IC industry. Within the past few years, there has been significant progress in the material quality. It has been demonstrated that the dislocation density in SIMOX is a steep function of oxygen implant dose [2]. Further, it is known that when the implant dose is below "critical" dose, less than 1000defects/cm^2 can be achieved. This critical dose is a function of the implant energy and decreases below $1x10^{18}/cm^2$ as the implant energy is reduced [2]. Therefore, by optimizing ion implantation conditions, the dislocation densities in SIMOX can be significantly reduced. Now, the main challenge is the supply of high quality SIMOX substrates at low cost. Nowadays the commercial SIMOX substrate costs about $200/per wafer compared to about $30/per wafer for normal bulk Si. This is mainly because of the high cost of the conventional ion-implanter and their limitations in ion current and beam area.

A novel ion implantation technique is Plasma-Immersion-Ion-Implantation (PIII), which has the advantages of simple machine design, low cost and high throughput. This method has been proposed recently, and is expected to have a wide application in SIMOX technology. In this technology, oxygen plasma is generated by an ECR source, and oxygen ions diffuse into the process chamber where they are extracted directly from plasma in which the wafer holder is located. Implantation is achieved as the positively charged ions in the plasma sheath are accelerated toward the wafer as shown in Fig.1. The space charge region between the plasma and a negatively biased target can sustain a potential difference up to 100kV, with an implantation flux as high as 1016/cm2-sec. There is no mass selection or ion optics in this technology, and the complexity of the implant machine is greatly reduced. The implantation area can be quite large, and a large

workpiece can be accommodated without scanning the beam. The simple machine design, high throughput, and small reactor size make PIII a good candidate for cluster tool implantation.

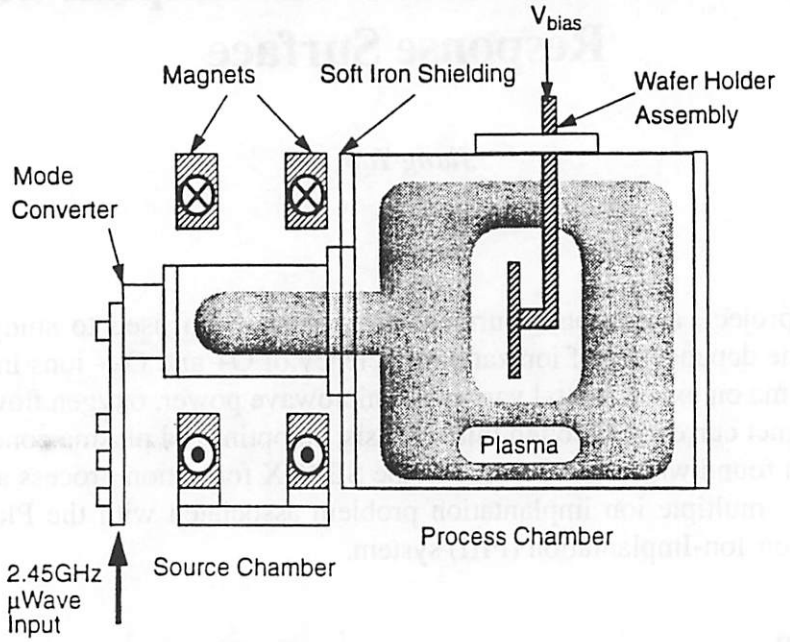


FIGURE 1. Schematic plot of the PIII system

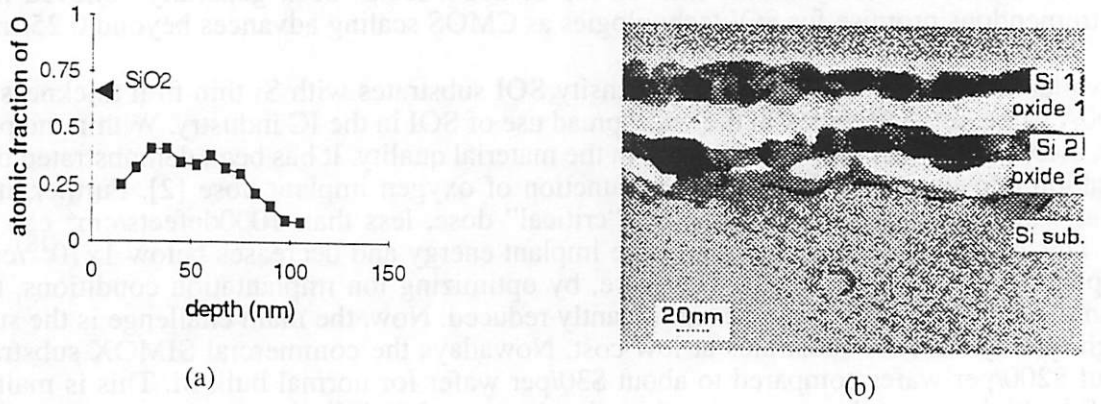


FIGURE 2. (a) Oxygen concentration profile calculated from RBS measurement on the as implanted wafer and (b) XTEM micrograph of the double buried oxide layer formed after annealing.

However, oxygen plasma consists of both O_2+ and O+ ions. Because the mass of O_2+ is twice that of O+, the projected range Rp will be roughly two times different for these two kinds of ions. Therefore, it will form two discrete peaks after the implantation as shown in Fig.2, which shows

two separate SiO_2 layers were formed after thermal annealing. This has been a major problem since low implantation energy was required in order to form very thin top single crystalline Si.

In this project, the response-surface method was used to study the effects of process condition on plasma composition, and establish a mathematical relationship between process variables and the ratio of O+ and O_2+ ion density and find the process conditions that can maximize this ratio. In other words, we seek to optimize this process so that only one kind of ion will dominate in the plasma.

2.0 Experimental Design

In PIII technology, plasma is generated by an electron cyclotron resonance (ECR) source powered by a 2.45GHz microwave power supply (x1) and a matching network (magnet field, which is controlled by the magnet current x3). The ion generation efficiency will be determined by x1, x3 and oxygen flow rate x2 (these are also the only adjustable variables in experiment). A 20-run experiment was designed in order to perform full resolution surface response analysis, and it is shown in Table I.

Table I. Experimental Design

Run	Block	x1	x2	x3	comment
1	1	-1	-1	-1	FF
2	1	-1	-1	1	FF
3	1	-1	1	-1	FF
4	1	-1	1	1	FF
5	1	1	-1	-1	FF
6	1	1	-1	1	FF
7	1	1	1	-1	FF
8	1	1	1	1	FF
9	1	-1.682	0	0	Axial
10	1	1.682	0	0	Axial
11	1	0	-1.682	0	Axial
12	1	0	1.682	0	Axial
13	1	0	0	-1.682	Axial
14	1	0	0	1.682	Axial
15	1	0	0	0	Center
16	1	0	0	0	Center
17	1	0	0	0	Center
18	1	0	0	0	Center
19	1	0	0	0	Center
20	1	0	0	0	Center

The levels of the variables in coded units were:

code	x1 (microwave power (W))	x2 (gas flow rate (sccm))	x3 (magnetic current(mA))
1	500	40	220
0	400	30	210
-1	300	20	200

The ion density was measured by using Mass Spectrometry (SXP 500). By setting the mass at 16 or 32 amu (atomic-mass-unit), we can measure the ion density of O+ or O_2+. Therefore, for each run, I measured O+ and O_2+ ion densities separately by changing the mass selection.

3.0 Results and Discussion

To carry out the experiment, the real experiment run sequence has been randomized as shown in Table II. The measured ion density ratio of O+ and O_2+ is also shown in Table II.

Table II. Experimental Results

Run	x1	x2	x3	ratio(O+ / O2+)
1	0	0	0	1.0875
2	1	1	1	1.5964
3	-1	-1	-1	1.0426
4	0	0	0	3.1784
5	-1.682	0	0	0.6923
6	0	0	-1.682	0.3288
7	0	0	0	1.6544
8	-1	1	-1	0.6821
9	0	0	0	0.7641
10	0	-1.682	0	5.2035
11	1	1	-1	0.4582
12	-1	-1	1	2.8622
13	0	1.682	0	0.8542
14	1.682	0	0	1.4168
15	-1	1	1	0.9221
16	0	0	1.682	0.9316
17	0	0	0	0.4285
18	1	-1	-1	1.0115
19	1	-1	1	2.7091
20	0	0	0	1.7928

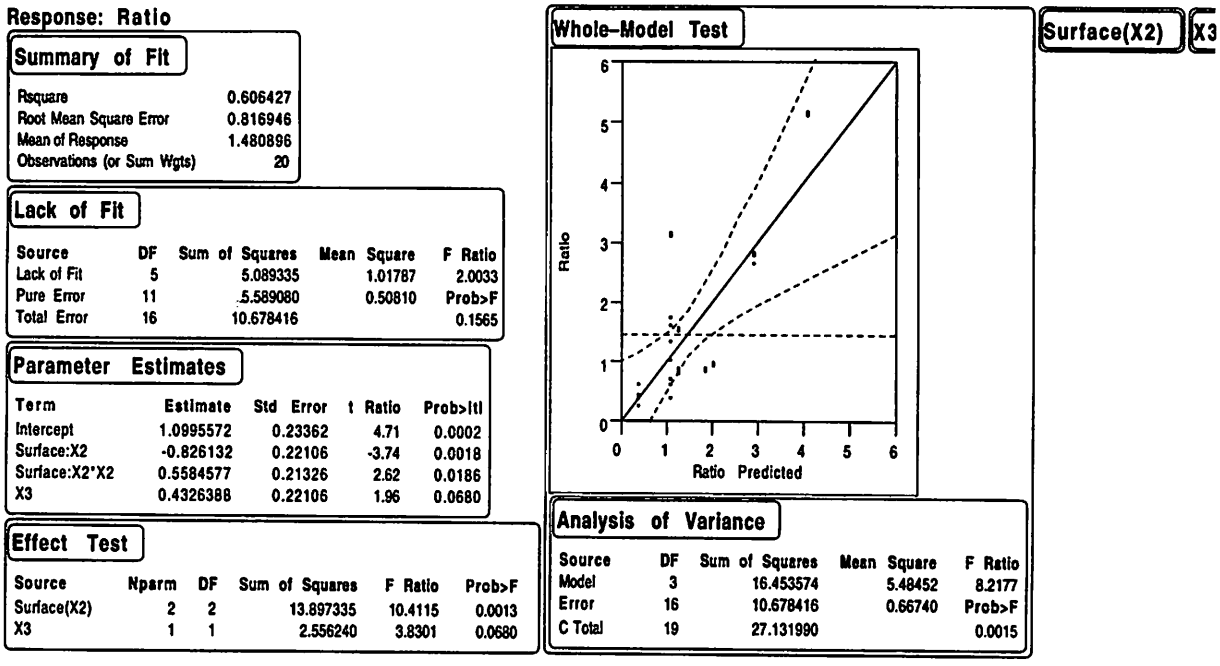
We used the JMP statistical analysis software to analyze the data, and try to fit the data to different models. First, I fit the data to a factorial model, and found that x1, x1*x2, x1*x3, x2*x3, and x1*x2*x3 are not important terms. By ignoring these terms, I tried to build different models in order to get the best fit. Based on some theoretical and fundamental understanding of plasma chemistry, I found the best model that fits the experiment data is:

$$Y(ratio) = \underset{(\pm 0.234)}{1.1} - \underset{(\pm 0.221)}{0.826X_2} + \underset{(\pm 0.213)}{0.5585X_2^2} + \underset{(\pm 0.221)}{0.4326X_3} \qquad (1)$$

The fitting results are shown in Table III. The physical explanations are following:

(1) From this model, we can see the ion density ratio (O+ /O_2+) is independent of microwave power (x1) in our experiment range (300W to 500W). This is because from plasma chemistry

point of view, there exists a critical power for generating O+ and O_2+ ions in oxygen plasma. When the microwave power is larger than this critical power, the plasma chemistry, i.e. generation of O+ and O_2+ ions and electrons, will not be affected by the input microwave power. From this experiment, we found the critical power to be below 300W in our plasma system.



Response: Ratio

Summary of Fit

Rsquare	0.606427
Root Mean Square Error	0.816946
Mean of Response	1.480896
Observations (or Sum Wgts)	20

Lack of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	5	5.089335	1.01787	2.0033
Pure Error	11	5.589080	0.50810	Prob>F
Total Error	16	10.678416		0.1565

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.0995572	0.23362	4.71	0.0002
Surface:X2	-0.826132	0.22106	-3.74	0.0018
Surface:X2*X2	0.5584577	0.21326	2.62	0.0186
X3	0.4326388	0.22106	1.96	0.0680

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Surface(X2)	2	2	13.897335	10.4115	0.0013
X3	1	1	2.556240	3.8301	0.0680

Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	16.453574	5.48452	8.2177
Error	16	10.678416	0.66740	Prob>F
C Total	19	27.131990		0.0015

Surface(X2) X3

(2) From this model, we can see the ion density ratio (O+ /O_2+) is linearly dependent on the magnet current (x3) or the magnet field. This is because by changing the magnet field, the ECR zone location (or plasma location) will be changed. Therefore, the ion loss to the aluminum chamber wall or the ion density will be affected.

(3) From this model, we can see the ion density ratio (O+ /O_2+) is strongly dependent on oxygen gas flow rate or the chamber pressure. This is because plasma chemistry strongly depends on oxygen source concentration. By changing the oxygen flow rate or the available oxygen to generate plasma, the generated ion density can be greatly varied.

Fig.3 shows the 3-D plot of Eq.(1), and we can see that the optimal condition for SIMOX process is set x2 at the lowest level and x3 at the highest level.Fig.3 3-D plot of the ratio of O+/O_2+ as a function of oxygen flow rate x2 and magnet current x3.

4.0 Conclusion

In this project a Response Surface method has been used to study oxygen plasma conditions in a SIMOX formation process. A model has been developed to model the ratio of O+/O_2+ as a function of process variables (microwave power, oxygen flow rate and magnet current). The results show that microwave power has little effect on the O+/O_2+ ratio. The physical meaning of this model has been briefly presented. By using this model, it is found that by setting oxygen flow rate to the lowest level and the magnet current to the highest level in our experiment range, we can achieve the optimal plasma condition for a SIMOX formation process.

Acknowledgment

I would like to thank my colleagues S. Sundar Kumar Iyer for helping me set up the Mass Spectrometry to carry out the experiment, Xinhui Niu for helping me use the JMP software to analyze the experiment data, and professor C. Spanos for guidance and valuable discussions.

References

[1] J.P. Colinge, <<Silicon-on-Insulator Technology: Materials to VLSI>>, Kluwer Academic Publisher, 1991.

[2] S. Nakashima, et al., Proc. 5th International Symposium on SOI Technology and Devices, Electrochem. Soc., vol.92-13, p.358, 1992.

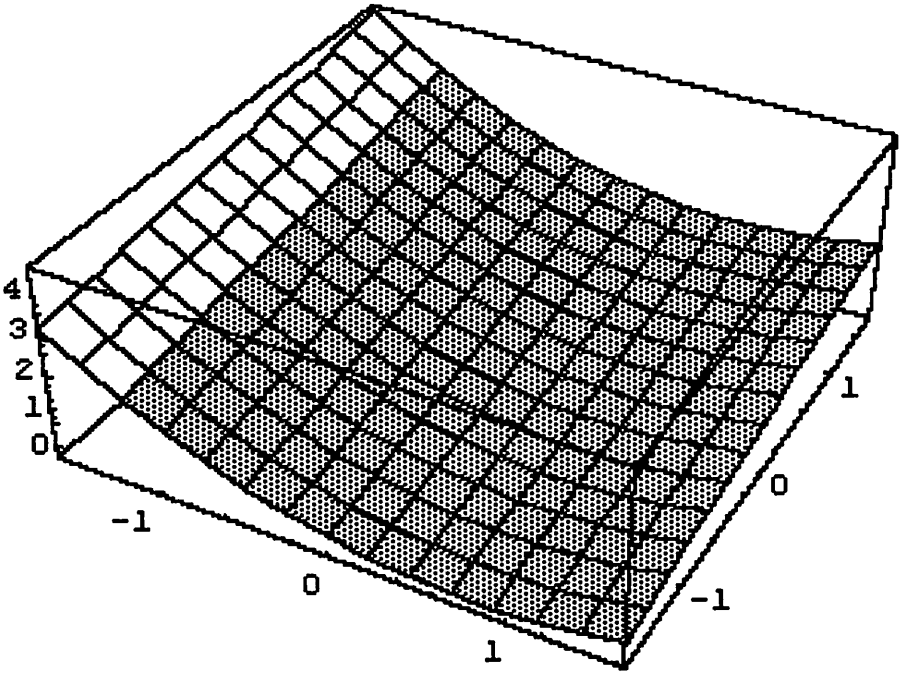


FIGURE 3. Fig.3 3-D plot of the ratio of $O+/O_2+$ as a function of oxygen flow rate x2 and magnet current x3.

Multiple Data Streams in Real-Time Multivariate Statistical Process Control

Herb Huang

A real-time multivariate statistical process control (SPC) methodology using time-series filters and multivariate analysis techniques has been proposed and implemented by previous authors. This paper first describes the methodology, and then attempts to expand its scope to include multiple, asynchronous data streams; in addition, some practical issues are resolved. The work is incorporated into a commercially available software package that employs these techniques. Examples using data from a plasma etcher are given.

1.0 Motivation

The real-time SPC methodology described in the next section has been shown to be useful for monitoring equipment faults. However, it has the following limitations:

1. It is limited to synchronous data, i.e., data in which all measurements are sampled at the same times and at equal time intervals.
2. The models created may be inappropriate.

In addition, the following practical limitations are observed when the methodology is implemented into software:

1. Scaling is needed when data streams vary greatly in absolute value, otherwise numerical round-off errors become substantial.
2. Extremely large data sets must be compressed in order to fit into computer memory.

This work addresses items #1 in both of the above lists. The goal is to allow real-time SPC of all types of data streams. In principle, there should be no limit to the number of data streams that can be analyzed. To this end, a method of synchronizing multiple data streams is presented. Also, scaling is implemented into the software, so that data streams with greatly varying absolute values can be accommodated.

The remaining issues are the subject of ongoing research.

2.0 Introduction

Traditional control charts use measurements that are assumed to be independent from one sample to the next [1]. The measurements are then plotted on a chart and points which fall outside a set of control limits are considered to be "out-of-control." In contrast, the data which this paper considers are real-time data that violate the assumptions of independence. The data are, in fact,

highly auto-correlated (correlated between samples), as well as cross-correlated (correlated between variables). Therefore, traditional control charts can not be applied directly on these data.

The real-time SPC methodology used for this work will now be described. This methodology has been implemented in a commercially available software package, known as RTSPC. For a more complete description, see the original authors' text [2][3][4][5]. The methodology can be separated into two parts: system modeling and fault detection.

2.1 System modeling

Before faults can be detected, the system to be controlled must be characterized with in-control data. These data are assumed to represent the normal operating mode of the system. System modeling consists of the following steps: 1) selection of signals, 2) identification of time-series models, 3) estimation of model parameters, and 4) calculation of the covariance matrix of the model residuals. The selection of signals and the identification of the time-series models are difficult in general. Moreover, they often require subjective decision-making and employ physical or intuitive knowledge. The selection of signals may require a designed experiment or some kind of trial and error [6].

Assuming the desired signals have been selected, the next step is to model them with time-series models. The models used are also known as ARIMA models. Much literature is available on these models[7][8][9][10][11][12][13][14], thus this paper will only briefly introduce them. In short, the models are of the general form:

$$\phi(B)\, w_t = \theta(B)\, a_t \tag{1}$$

$$\phi(B) = 1 - \sum_{k=1}^{p} \phi_k B^k, \quad \theta(B) = 1 - \sum_{k=1}^{q} \theta_k B^k$$

$$w_t = \nabla^d z_t, \quad d \geq 0$$

$$\nabla z_t = z_t - z_{t-1}, \quad B^k z_t = z_{t-k}$$

where z_t are the data, w_t are the differenced data, and a_t are the prediction errors.

The identification of time-series models and estimation of the model parameters have been implemented into an automatic model generation program by [4]. The method used in RTSPC for identifying time-series models is based on the modified Yule-Walker equations, augmented with known heuristics and experience with the selected signals [4][16]. Once the models are identified for each signal, the model parameters are estimated and the covariance matrix of the model residuals are calculated. Together, the models and the covariance matrix characterize the system.

2.2 Fault detection

2.2.1 Overview

Fault detection uses the results from the system modeling to monitor the operation of the system and generate alarms when the system has deviated from the previously modeled state. A simulation tool developed at the University of California, known as Ptolemy [15], will be used as a

framework in which to discuss the fault detection algorithm. This framework was chosen because it allows a high-level signal-processing view of the SPC methodology and, at the same time, illustrates some of the practical issues involved with implementing it into a software utility.

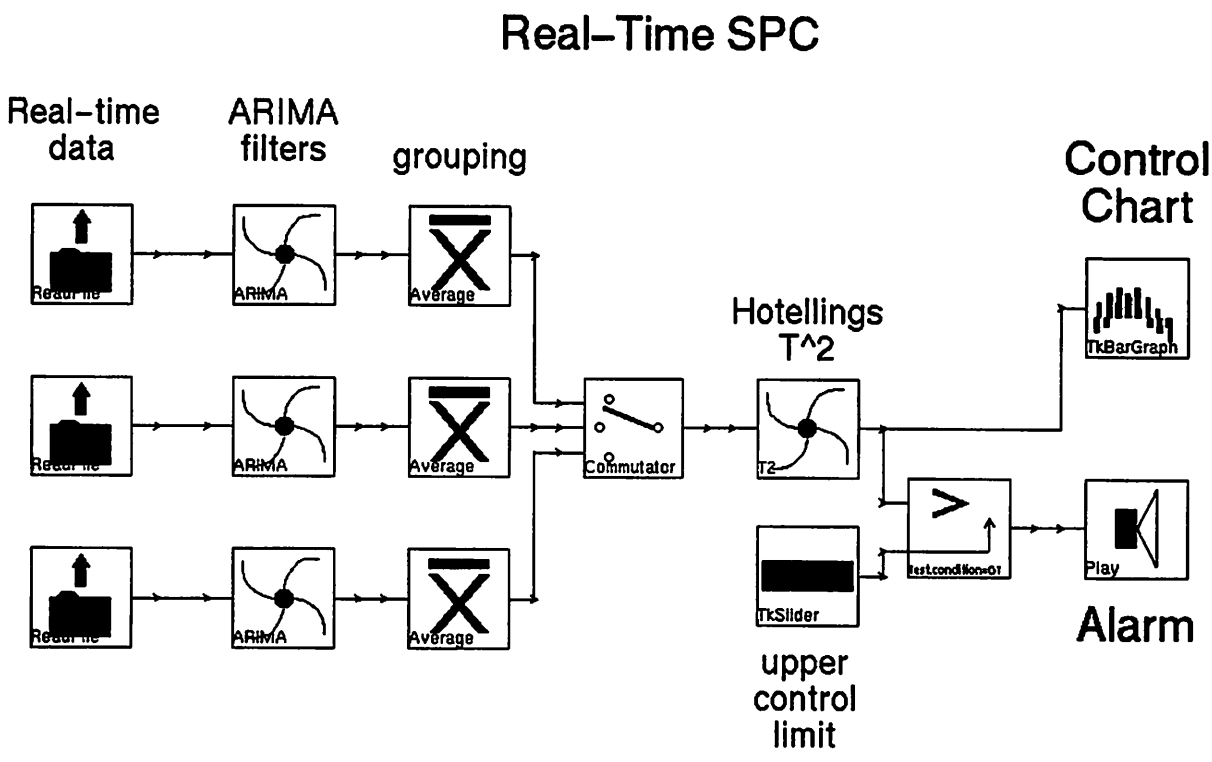


FIGURE 1.

FIGURE 1. shows an overview of the real-time SPC methodology. The Ptolemy diagram consists of blocks (also known as "stars" in Ptolemy) that each process their inputs and then send their outputs along the connecting link to the next block. The left-most blocks in the figure ("ReadFile") read data from a file on disk. The figure shows only three streams of real-time data corresponding to three different sensors, but any number can be supported. The data then feeds into the ARIMA filter blocks, which use the previously calculated time-series models. These models forecast the output for each sample, which is then subtracted from the actual value to result in the forecasting residual. The result of a ARIMA filter is to whiten the data, i.e., to remove autocorrelation from the data.

After ARIMA filtering, the data is grouped to adjust the sensitivity of alarms. The group size is adjustable by the user. A high group size averages the data, thus smoothing it.

In RTSPC, there are actually two ARIMA models for each variable; one is a seasonal model (between wafers), and the other is a non-seasonal model (within wafer). The entire procedure shown in FIGURE 1. is simply repeated for each of the two models, using different group sizes.

After grouping, the output of the three streams feeds into the "Commutator." The purpose of this block is to interleave the three streams into one. The interleaved stream is fed into the "Hotelling's T^2" block, where the T^2 statistic is calculated. The T^2 statistic is then fed to a graphical plotter, as well as a decision block, which checks to see if the statistic has crossed the upper control limit. The upper control limit in FIGURE 1. is given by the "TkSlider" block, which allows the user to adjust the upper control limit as desired. The theoretical upper control limit for the T^2 statistic is related to the F-distribution and depends on the Type I error, α, the number of samples, N, and the number of monitored variables, P [1]:

$$UCL_{\alpha, P, N} = \frac{P(N-1)}{N-P} F_{\alpha, P, N-P}$$

When an alarm occurs, the "Alarm" block is triggered.

2.2.2 ARIMA filter

The ARIMA filter block is shown in FIGURE 2. The filter consists of three "FIR" (finite impulse response) blocks. The diamonds placed on the connections denote delays (i.e., the backward shift operator). The user must fill in the actual model parameters (not shown in the figure) obtained from system modeling. The first FIR performs differencing, the second calculates the auto-regressive (AR) portion of the model, and the third calculates the moving average (MA) portion of the model. The output is the forecasting residual. The overall filter corresponds to a rational transfer function [16][17].

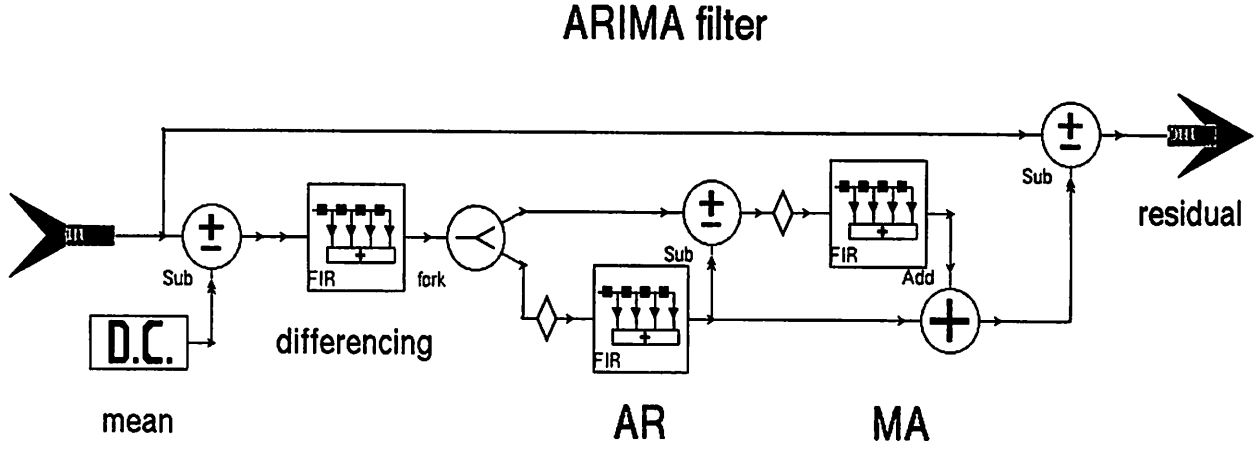


FIGURE 2.

2.2.3 Hotelling's T^2

The calculation of the T^2 statistic is shown in FIGURE 3. The input to this block is an interleaved stream of data, and the outputs are the scalar values of the statistic. The block at the bottom-left labeled "Covariance Matrix" is the covariance matrix calculated from the residuals of the in-control models. If x is the input (column) vector, then:

$$T^2 = nx^T S^{-1} x \qquad (2)$$

where n is the group size (a scalar) and S is the covariance matrix.

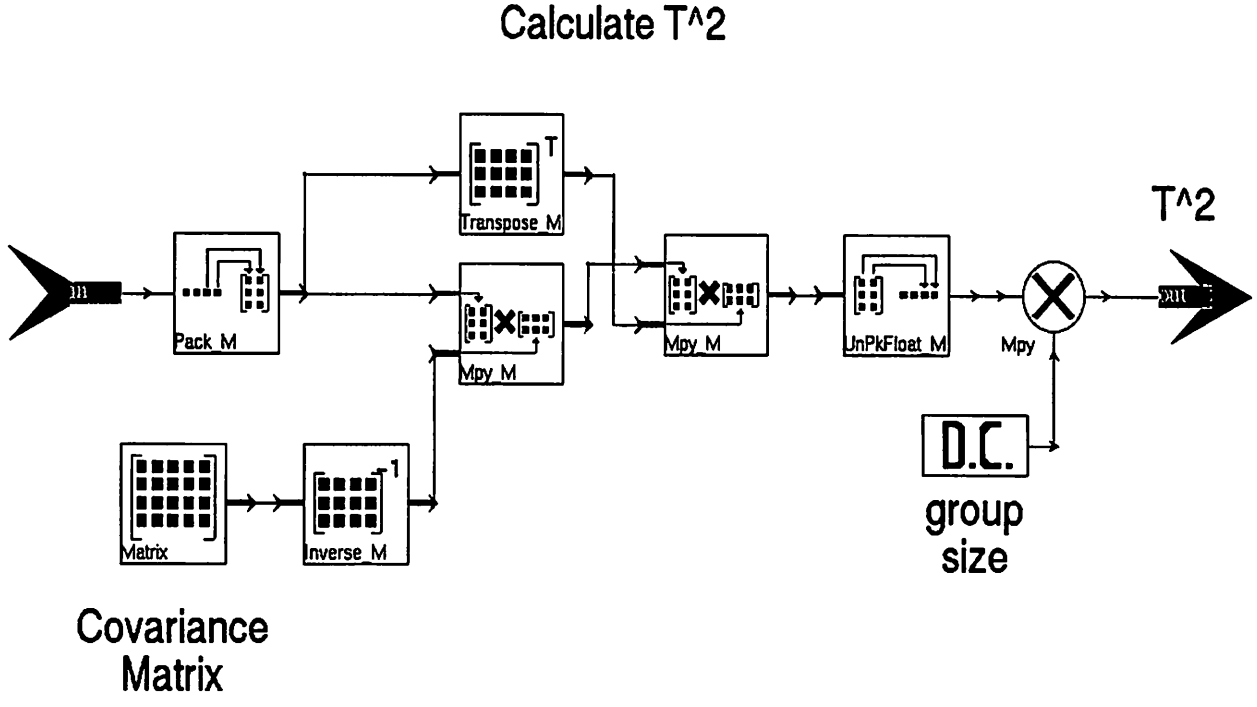


FIGURE 3.

3.0 Implementation

The two important modifications required are the synchronization of data streams and the scaling of data. Synchronization of data streams is required in order to incorporate multiple data streams into a single analysis. Scaling of the data is important to maintain an acceptable numeric precision when using digital arithmetic. Another more practical (and perhaps just as difficult) problem is the management of limited computer memory.

3.1 Synchronizing data streams

An example pair of data streams is drawn in FIGURE 4. to illustrate the problem of asynchronous data streams. In the figure, Data A and Data B represent two streams of data. Each stream has its own sample times and is not necessarily sampled at regular intervals. However, the real-time SPC methodology described above requires that all data streams be sampled at identical and equally spaced times.

The proposed synchronization method is to linearly interpolate between measurements and take samples of all measurements at equally spaced times. The linear interpolation is given by:

$$y_s = y_{t_1} + \frac{s - t_1}{t_2 - t_1}\left(y_{t_2} - y_{t_1}\right) \tag{3}$$

where s is the desired sample time and $t_1 \leq s < t_2$. From the figure, the method is easily seen pictorially. If the desired sample times are 1, 2, 3,..., then the method consists of simply taking the samples as the points that cross the vertical lines extending from each sample time.

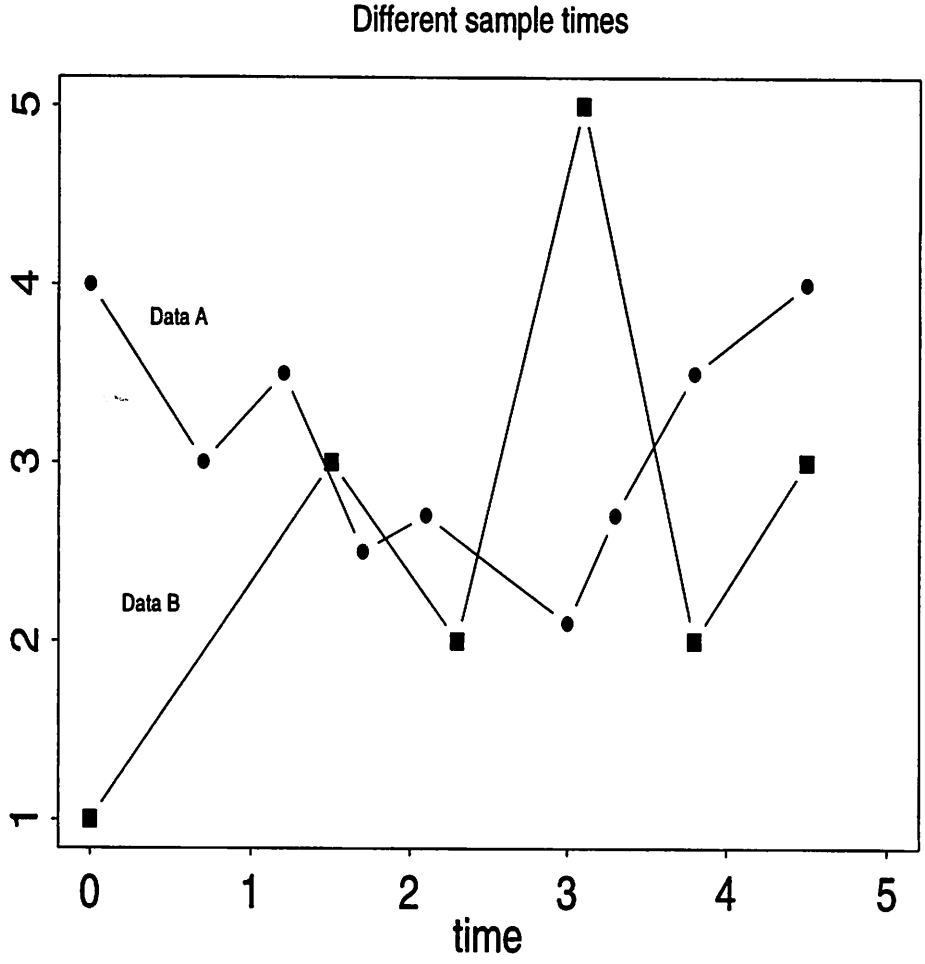


FIGURE 4.

The sampling frequency needs to be carefully chosen. A sampling frequency that is too fast will introduce additional correlation into the data; on the other hand, a sampling frequency that is too slow will throw away useful information. A good compromise is to sample at the rate of the slowest data stream.

3.2 Scaling

The motivation for scaling is to improve numerical precision. All operations on computers are done with finite precision, and if one is not careful, round-off errors can result in miscalculations. An example of an algorithm vulnerable to numerical precision problems is matrix inversion. In RTSPC, matrix inversion is required during model generation and also for the calculation of the T^2 statistic.

The effects of finite precision are especially damaging when the data streams vary greatly in absolute value. Such data streams produce highly ill-conditioned matrices; the result is that round-off errors can cause a matrix to become numerically singular, thus causing matrix inversion

to fail [18]. The usual cure for this malady is to scale the data to unity variance before doing matrix inversion, so that the data values are of the same order of magnitude.

4.0 Results and Discussion

Examples of the real-time SPC methodology will be given using data from the Lam 4400 Rainbow etcher in the Berkeley MicroLab and using the RTSPC software package. Real-time data are available from three sources: the Brookside LamStation software which obtains measurements via SECS-II, the Comdel Real Power Monitor (RPM-1) via RS232, and the Chromex Imaging Spectrograph (OES data). Currently, RTSPC is limited to analysis of data from a particular source (the LamStation).

Data scaling was implemented before the automatic model generation. This allowed successful generation of models for RF power, RF voltage, and RF phase error using data from the RPM-1. Previously, the model generator had failed when attempting to model these data. FIGURE 5. displays original and forecasted data using a generated model for RF voltage.

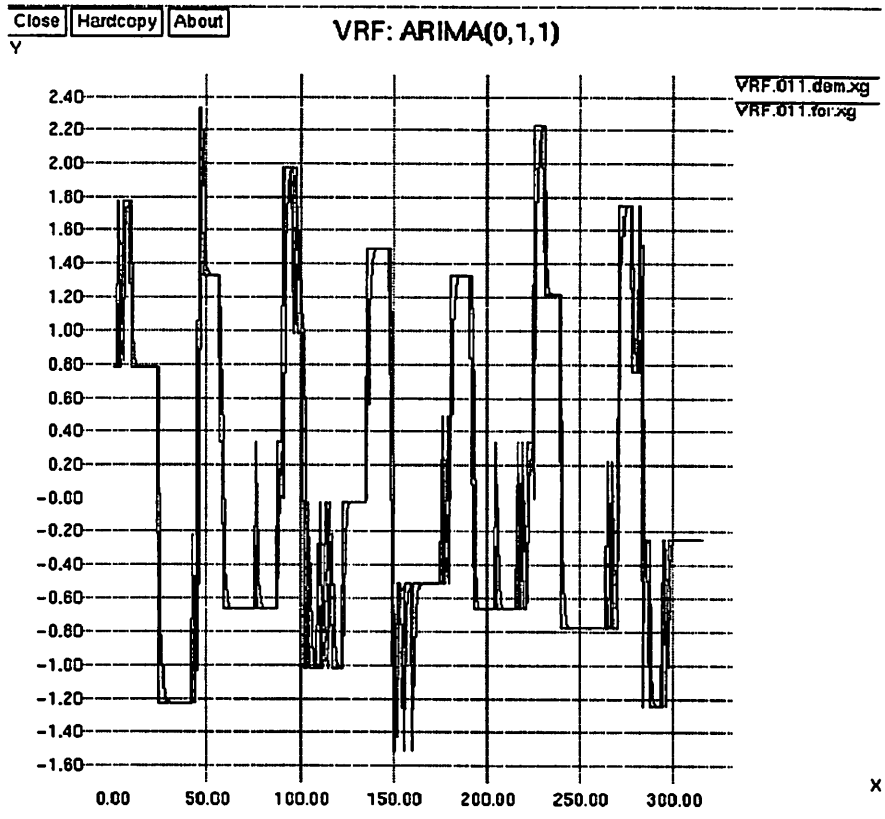


FIGURE 5.

The interpolation scheme discussed in the previous section was implemented into a utility called "interp." This utility takes data streams and samples them at regular intervals, using linear interpolation. The utility was run on both LamStation signals and RPM-1 signals in order to synchronize the streams. Then another utility, "merge", was created to merge the data streams into a single file. Baseline models were created and the results fed into the SPC module of RTSPC. The merging with compressed data from OES has not yet been tried, but there are no difficulties

expected, since the OES data are simply another stream of data to be synchronized, merged, modeled, and monitored.

Data scaling was also implemented before inverting the covariance matrix in the T^2 statistic. This allows RTSPC to calculate control charts for all types of data. No figure is included for this; it suffices to say that the combination of baseline signals from the RPM-1 and LamStation produced a control chart that appeared in-control.

A more detailed summary of software changes to RTSPC is included in the appendix.

5.0 Future Work

The following is a list of ideas for future work.

The methods presented in this paper for supporting multiple data streams could be incorporated into the RTSPC package.

Many more methods are known for identifying and estimating time-series models. The Mayne-Firoozan method and Akaike method are some examples [16].

Currently, RTSPC employs two different models for each variable, a seasonal model and a non-seasonal model. The T^2 statistic is calculated using both types of models. A different multivariate control chart, the CUSUM chart, might be tried instead [19]. A CUSUM chart may prove to be more sensitive than the T^2 chart [20]. However, α is difficult to calculate in general.

A stability check for the ARIMA model generator should be added. The stability criteria for an ARIMA model specifies that the roots of the polynomials, $\phi(B)$ and $\theta(B)$, should lie outside the unit circle in the complex plane [8]. When unstable or nearly unstable models are generated, the user should be warned. Furthermore, this information can be used to improve model identification. Various polynomial root finding algorithms can be found in [18].

The overall memory management of RTSPC desperately needs to be improved. The current mis-management of memory limits the number of variables and amount of data that can be analyzed.

The current matrix inversion algorithm is an LU-decomposition. This could be replaced by a Singular Value Decomposition (SVD). The SVD is perhaps twice as slow, but it is numerically very stable [18]. Moreover, it produces extremely useful diagnostic information. The condition number of the matrix can be calculated to alert the user of potential numerical precision problems. Also, when near-singular matrices are detected, the algorithm can eliminate bad parameters from a model.

Future work might implement machine recipes as another input to RTSPC; this would obviate the need to create a new baseline model every time a machine recipe is changed.

Another project could involve using real-time data for final wafer state prediction, but this would involve a major re-working of the current RTSPC framework.

References

[1] Douglas C. Montgomery, *Introduction to Statistical Quality Control*, 2nd. ed., NY: John Wiley & Sons, 1985.

[2] Costas J. Spanos, Hai-Fang Guo, Alan Miller, and Joanne Levine-Parill, "Real-Time Statistical Process Control Using Tool Data," *IEEE Transactions on Semiconductor Manufacturing*, Nov. 1992.

[3] Sherry F. Lee, Eric D. Boskin, Hao Cheng Liu, Eddie H. Wen, and Costas J. Spanos, "RTSPC: A Software Utility for Real-Time SPC and Tool Data Analysis," *IEEE Transactions on Semiconductor Manufacturing*, Feb. 1995.

[4] Hao-Cheng Liu, "Automatic Time-Series Model Generation for Real-Time Statistical Process Control," Memorandum No. UCB/ERL M93/45, Berkeley: Electronics Research Laboratory, June 1993.

[5] Hai-Fang Guo, "Real Time Statistical Process Control for Plasma Etching," Memorandum No. UCB/ERL M91/61, Berkeley: Electronics Research Laboratory, July 1991.

[6] Sherry Fen-hwei Lee, "Semiconductor Equipment Analysis and Wafer State Prediction System Using Real-Time Data", Memorandum No. UCB/ERL M94/104, Berkeley: Electronics Research Laboratory, Dec. 1994.

[7] George E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed., Englewood Cliffs, N.J.: Prentice Hall, 1994.

[8] Robert H. Shumway, *Applied Statistical Time Series Analysis*, Englewood Cliffs, N.J.: Prentice Hall, 1988.

[9] Alan Pankratz, *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, NY: John Wiley & Sons, 1983.

[10] Walter Vandaele, *Applied Time Series and Box-Jenkins Models*, NY: Academic Press, 1983.

[11] Bovas Abraham and Johannes Ledolter, *Statistical Methods for Forecasting*, NY: John Wiley & Sons, 1983.

[12] Peter J. Brockwell and Richard A. Davis, *Time Series: Theory and Methods*, NY: Springer-Verlag, 1991.

[13] David R. Brillinger, *Time Series: Data Analysis and Theory*, SF: Holden-Day, 1981.

[14] Bruce L. Bowerman and Richard T. O'Connell, *Forecasting and Time Series: An Applied Approach*, 3rd. ed., Belmont, Calif.: Duxbury Press, 1993.

[15] *The Almagest: Vol. 1--Ptolemy 0.5 User's Manual*, University of California, 1994.

[16] Steven M. Kay, *Modern Spectral Estimation*, Englewood Cliffs, N.J.: Prentice Hall, 1988.

[17] Steven M. Kay, S. L. Marple, Jr., "Spectrum Analysis--A Modern Perspective," *Proceedings of the IEEE*, Nov. 1981.

[18] William H. Press, et al., *Numerical Recipes in C*, Cambridge University Press, 1992.

[19] Ronald B. Crosier, "Multivariate Generalizations of Cumulative Sum Quality-Control Schemes," *Technometrics*, Aug. 1988.

[20] William H. Woodall and Matoteng M. Ncube, "Multivariate CUSUM Quality-Control Procedures," *Technometrics*, Aug. 1985.

Appendix for developers: software changes to RTSPC

arima

general

- cleaned up entire code; *code is now much easier to read and to work with*
- fixed calculations of auto-correlation and covariance when data is not zero-mean
- patched up dynamic memory leaks
- temporary file name prefix changed to: arima$PID, where $PID is the process ID number

data file

- data is read one line at a time and *only the necessary amount of memory is used*
- fixed a bug that used to cause the parser to skip the first line after each header (except the first header)
- headers are identified by the first word of the first header (can be any string); these headers mark the separation between wafers
- wafers with insufficient data are ignored; a warning is printed
- blank lines are ignored, as are lines beginning with the comment character '#'

configuration file

- added MAX_DIFF and MAX_SEA_DIFF as options that allow the user to specify the maximum differencing order
- allowed use of '#' as a comment character

spcwish

- cleaned up entire code; *code is now much easier to read and to work with*
- residuals are scaled to unity variance before calculating inverse of covariance matrix
- fixed some dynamic memory leaks, but there is still a big problem with dynamically allocated memory
- data file parser no longer looks for "SAS"
- prints warnings when seasonal upper control limit is bogus
- temporary file name prefix changed to spc$PID, where $PID is the process ID number

rtspc

- added a "Postscript" button in the plot window that creates a Postscript file named "/tmp/canvas-[exec whoami].ps", where [exec whoami] is replaced with the user's login name.
- made changes in the procedures *fetch_newmodelfile* and *create_spc_config* to allow data files that trigger on a keyword other than "step".

Real-Time SPC for Plasma Etching Using Optical Emission Spectroscopy

Roawen Chen

A new approach of real-time statistical process control (RTSPC) for plasma etching is proposed in this term project. Instead of using the real-time signals from other conventional sensors such as an RF monitor and SECSII-based signals, we monitored the signals from optical emission spectroscopy (OES) from 250nm-750nm. Principal component analysis (PCA) was used to filter real-time OES signals before further analysis. Time-series models, as well as multivariate T^2 analysis, were used to construct an SPC chart.

1.0 Introduction

Plasma etching has been widely used in the manufacturing of submicron IC devices. Traditionally, test wafers are run and measured after machine recipe changes, and then used to examine the monitored parameters (e.g., etching rate). When the monitored parameters exceed the control limits, an alarm will be generated to indicate a possible machine malfunction. However, with the advent of smaller feature sizes and larger wafer dimensions, the requirements for better real-time machine monitoring are increasing. Because of this, recent efforts have focused on several sensing techniques that are needed to provide real-time plasma process information. In our plasma etcher (i.e. Lam4400), RF monitor, LamStation, and end-point detector (i.e. monochromatic spectrography) were installed to collect the real-time signals. In this project, instead of monitoring these common real-time signals, we attempted to monitor signals from the optical emission spectrum. More specifically, 1024 wavelengths from 250nm to 750nm were monitored. In addition, because the spectra were collected from three different positions inside the plasma, this technique takes into account etching uniformity. However, applying standard control charts to the real-time OES signals is not a viable method because these signals are highly correlated, and the mean value of a given wavelength may also drift with time. Also, the analysis of 1024 variables for real-time SPC is unnecessarily time consuming. Therefore, Principal Component Analysis (PCA) was used to compress the OES data set, and then baseline ARIMA time-series filters and multivariate statistics were used to analyze these variables.

2.0 Methodology

This section starts with the collection of real-time OES data, followed by then a brief introduction of our data compression techniques. After that, the reduced data set is processed by RTSPC, a software utility which was developed in Berkeley Computer-Aided Manufacturing group. This software provides the capability to automatically generate time series models and also combine the residuals into a multivariate T^2 chart. The concepts of time-series modeling and

Hotelling's T^2 statistic will be briefly discussed in the last two parts. An overview of the real-time SPC flow chart is shown below.

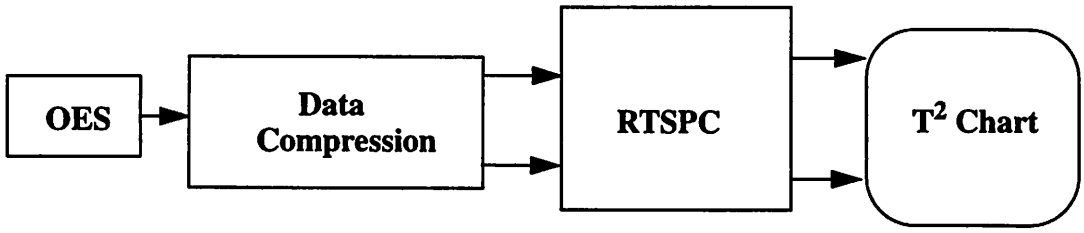


Figure 1. Flow Chart of this real-time SPC study

2.1 Real-Time signals

The real-time signals are collected using a spatially resolved OES system which has been installed in the Berkeley Microfabrication Laboratory. The Chromex spectrograph (Chromex 250IS), with the Princeton Instruments 256x1024 CCD camera is mounted in the instrument rack. Three optical fibers are connected from the spectrograph to the Lam 4400 reactor viewport. The spectrograph is set up to view the plasma across three distinct spots arranged laterally above the etching wafer. We have chosen a 150 groove/mm grating with spectral resolution of 0.4nm, in order to acquire spectral range from 250 to 750nm in the reasonable acquisition time of about one second/sample.

Since baseline wafers must first be processed to build the real-time series models, nine wafers were etched on the Lam Rainbow 4400 using the given baseline recipe (i.e. recipe #400). In this example, poly-Si was etched to endpoint and the total etching time was approximately 40-50 seconds. To simplify the time-series modeling, the same number of data points, or step length, is used for each wafer. Fifteen continuing spectra were selected from the same time frame for each wafer with the acquisition rate of 1/sec, excluding a few starting, unstabilized spectra and the spectra detected from the overetching period. In this term project, only the data collected from the middle lens are used. A typical spectrum collected during poly-Si etching is shown in Figure 2.

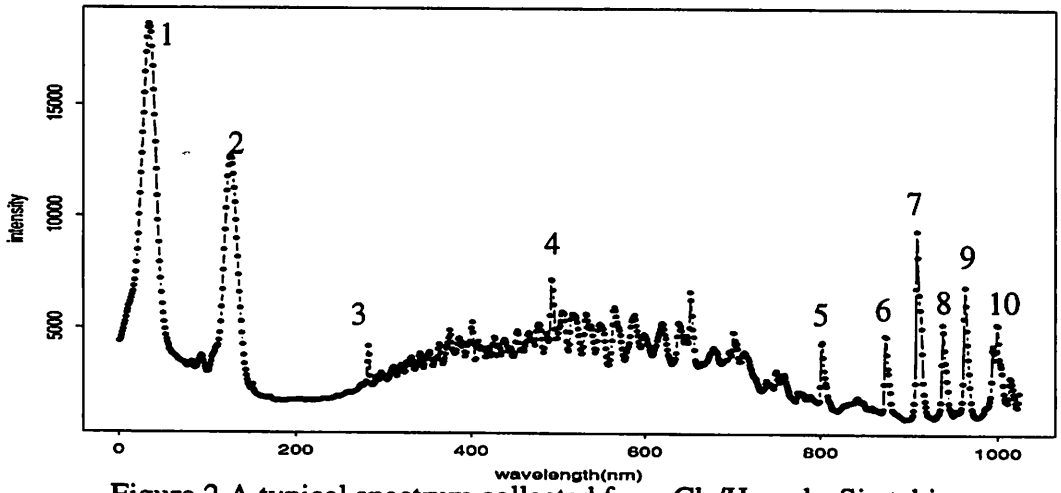


Figure 2 A typical spectrum collected from Cl_2/He poly-Si etching

However, the profile of intensity vs. time for wafer #4 shows some unusual discontinuities, as shown in Figure 3. These might be due to either some problems with the wafer, which were

unknown prior to the run, or the malfunction of machine. Thus, only eight wafers were used to create the baseline time-series models, excluding the wafer #4. Once the time series models are created, wafer #4, along with other eight wafers is analyzed by RTSPC in order to demonstrate RTSPC's alarm generation capability.

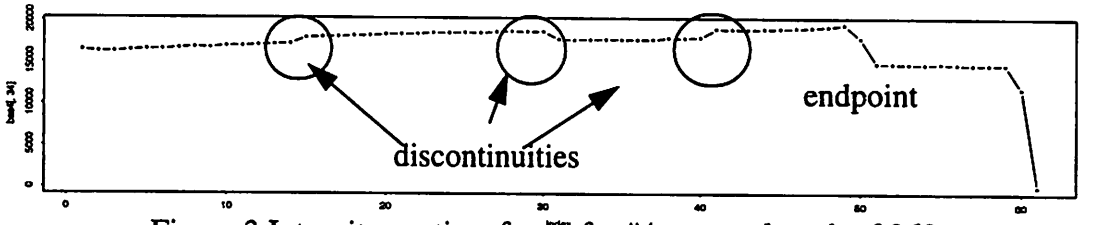


Figure 3 Intensity vs. time for wafer #4 at wavelength of 260nm

2.2 Data Reduction

Since typical OES data contains 1024 correlated variables in each spectrum, it is not practical to use them for real-time SPC. Two data compression techniques are performed to reduce the variable numbers in this study.

First, Principal Component Analysis (PCA) is used to transform the input variables to a set of orthogonal variables. The transformed variables, known as principle components (PCs), are the linear combinations of the original variables. Since the covariance matrix of the original input matrix X is symmetric, it can be decomposed to $X^TX=V\Omega V^T$, where the diagonal elements of the Ω are the eigenvalues and the columns of V are the eigenvectors of X^TX. The coefficients of the original variable are the eigenvectors V. The general equation of PCA is thus

$$PC = (X-\overline{X}^T)V \qquad (1)$$

where $\overline{X}^T$ is the vector of average values of each variable in X. The higher value of PC (i.e. larger score) denotes the larger variation along this transformed coordinate axes. In short, the purpose of PCA is to explain most of the variance in the original data set by only a few PCs without losing too much information. A better way to understand PCA is to illustrate it geometrically, as shown in Figure 4.

Another method is to simply select those distinct atomic and molecular spectral peaks from the spectra. That is, instead of using the entire spectrum, only those wavelengths contributed by the important chemical species are selected for real-time SPC. As such, the monitoring variables are reduced to only about 10, which are presumed to explain most of variance in the etching process. Those distinct wavelengths are listed in Table.1.

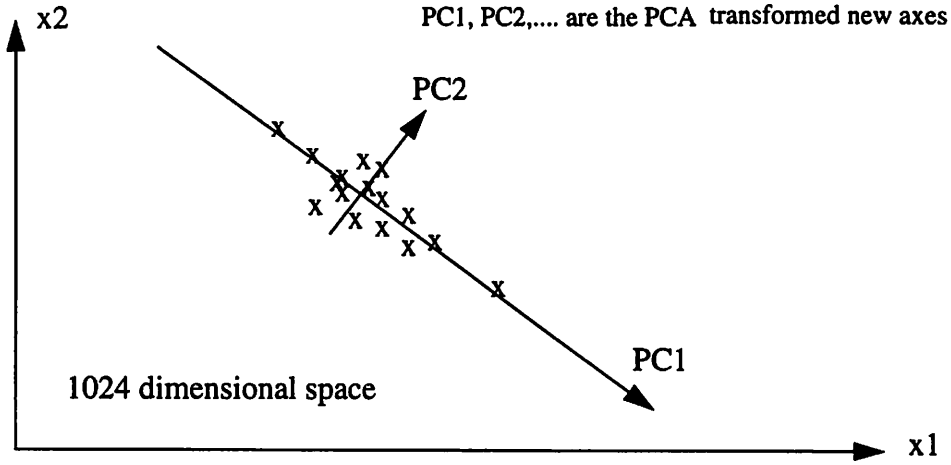


Figure 4 Principal component transformation for one wafer. 15 data points (i.e. 15 spectra for each wafer) in 1024 dimensional space

Table 1: Distinct wavelengths in our OES data

	wavelength(nm)	possible element
1	260	Cl_2
2	308	Cl_2
3	388	He
4	499	?
5	664	He
6	703	He
7	722	Cl
8	738	Cl
9	752	Cl
10	771	Cl

3.0 Results and Discussion

3.1 Reduced OES Data

Method 1: PCA. As described in the previous section, the original correlated OES data can be transformed to a set of orthogonal variables using PCA. Usually, only few principal components are enough to capture most of the variance of the original variables. In this case, the eigenvalue of the first PC explains 99% of the total variation of the input matrix, that is, most of process variation exists along one specific eigenvector.

In order to observe wafer-to-wafer variation or a long-term trend, the data points for each wafer need to be compared under the same eigenvector, as illustrated in Figure 5. For each wafer, the within-wafer variation can be illustrated by the geometrical shrinking or expanding of the data domain. The wafer-to-wafer variation can be explained by the difference between the center points of each wafer's data domain under the same coordinate axes. Therefore, the same eigenvector should be used for each baseline wafer to conduct principal component transformation. In our case, the eigenvector is given by

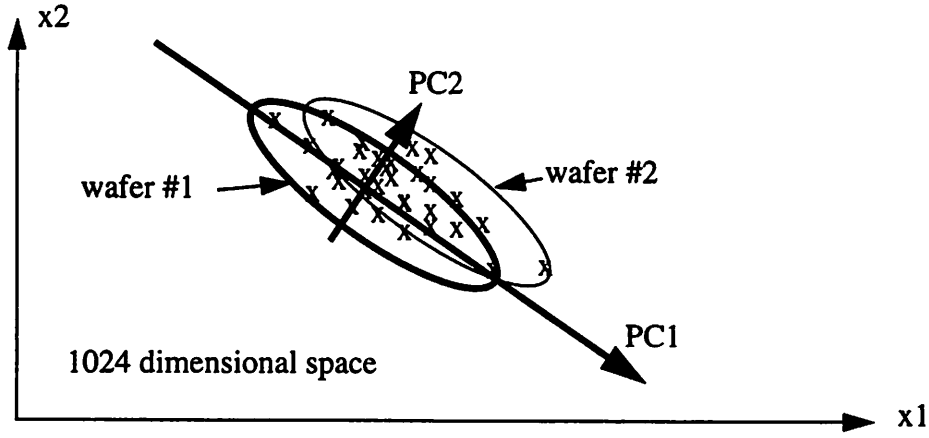


Figure 5. An illustration of the wafer-to-wafer variation in the PCA transformed coordinate axes. Note that the PC coordinate axes are fixed for all wafers

$$V_{baseline} = eigenvector\left(X_{mean}^{T} X_{mean}\right); \qquad X_{mean} = \frac{\sum_{i}^{8} X_i}{8}$$

where X_i is the intensity matrix for wafer i, which contains 1024 variables and 15 observations. The equation for the transformed variables PCs is thus

$$PC = (X - \overline{X}^{T})V_{baseline} \tag{2}$$

Figure 6 shows that the values of these PCs change with time. Notice that the score for each principal component roughly follows a seasonal upward trend.

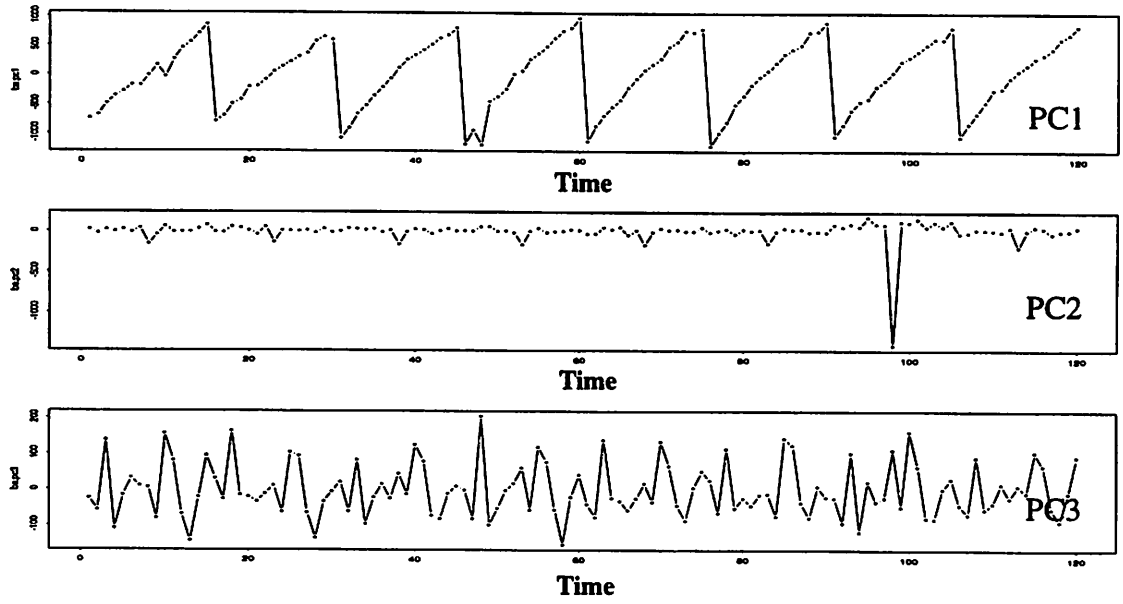


Figure 6 The scores over eight wafers for three principal components

Method 2: Using Ten Distinct Wavelengths. In this case, only 10 wavelengths were selected from the original OES data for time-series modeling. The profiles of intensity vs. time for these 10 wavelengths are shown in Appendix A.

3.2 RTSPC Algorithm

3.2.1 Baseline Model Generation

Up to this point we have two different filtered data sets with only three uncorrelated variables (reduced by PCA) or ten correlated variables (by selecting ten distinct wavelengths from the original OES data). These data were acquired while the machine was operating in control. Next, these data will be decomposed to wafer-to-wafer and within-wafer components, and time-series models will be generated using these selective data sets and our automatic model generation routine, RTSPC.

The results of the baseline model generation are shown in Figure 7, which displays the double T^2 chart for the data used to generate the model. The baseline data along with their corresponding within-wafer and wafer-to-wafer filtered residuals are shown in Appendix B. Note that in the double T^2 chart the long-term signals (the bar plots) seem to be zero. This is because the sample size (i.e. n) is close to the variables number (i.e. p), the control limits of wafer-to-wafer T^2 is much higher than that of within-wafer components[1]. While RTSPC scaled two separate T^2 statistics into one plot, the bar plot for wafer-to-wafer T^2 are scaled down to a small value and, as a result, can-

1. The control limits of the double T^2 chart: In the case of PCA reduced data, p=3, n=8 for long-term components and n=120 for short-term components. Thus T^2 control limits are

$$T^2 = \frac{(3 \cdot 7)}{5} F_{0.01,\,3,\,5} \approx 50\ (long-term)\ ;\ T^2 = 3F_{0.01,\,3,\,120} \approx 11\ (short-term)$$

In the case of data set with 10 wavelengths, p=10, n=8 for long-term components and n=120 for short-term components. Because n < p for long-term case, RTSPC just select a number without statistical meaning. Thus T^2 control limits in our case are

$$T^2 = 25\ (long-term) \qquad ;\ T^2 = 10F_{0.01,\,10,\,120} \approx 25\ (short-term)$$

not to be observed. Note that the value of the double T^2 statistic (shown on the left side of double T^2 chart) only indicates the T^2 statistic of the short-term components.

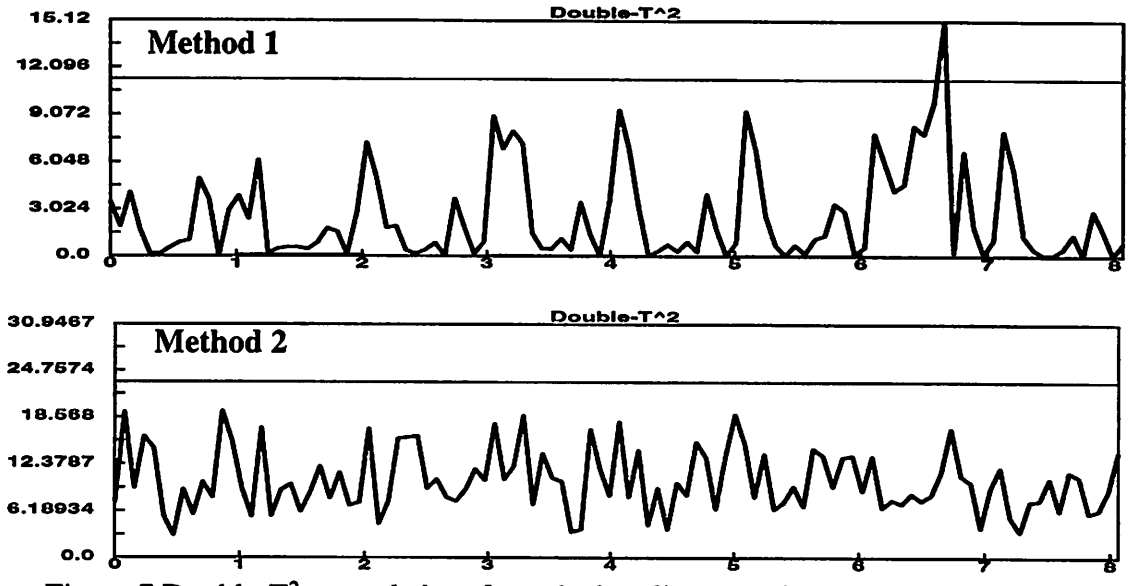


Figure 7 Double-T^2 control chart from the baseline experiment (NOTE: The line plots are the within-wafer T^2 and the bar plot are the wafer-to-wafer T^2)

Both data compression techniques show similar double T^2 charts. However, for the PCA reduced data set, the short-term signals (the line plots in the double-T^2 chart) indicate a slight problem with wafer #7. This is obviously due to the noisy spike in the original data of the second principal component shown in Fig.6.

The within-wafer and wafer-to-wafer ARIMA time series models generated by RTSPC are shown in Appendix C.

3.2.2 Fault Detection Example

Once the time-series models for the baseline process were created, the RTSPC software was able to generate real-time alarms in the case of misprocessed wafers. The peculiar wafer #4 mentioned in Section 2.1, was used along with other baseline wafers to demonstrate the alarm generation capability of RTSPC. Figure 8 gives the results for this additional run. One can see that both data reduction methods result in similar double T^2 statistics.

For method 1, both wafer-to-wafer and within-wafer T^2 statistics clearly identified the fault in wafer #4 even though this wafer only causes limited shift in the original signals, as shown in Fig. 3. For method 2, the within-wafer T^2 statistics were able to signal problems with wafer #4. However, no wafer-to-wafer T^2 alarm was generated. Again, this was obviously due to the insufficient amount of baseline wafers and the large number of variables.

These data and their corresponding wafer-to-wafer residuals and within-wafer residuals are shown in Appendix D.

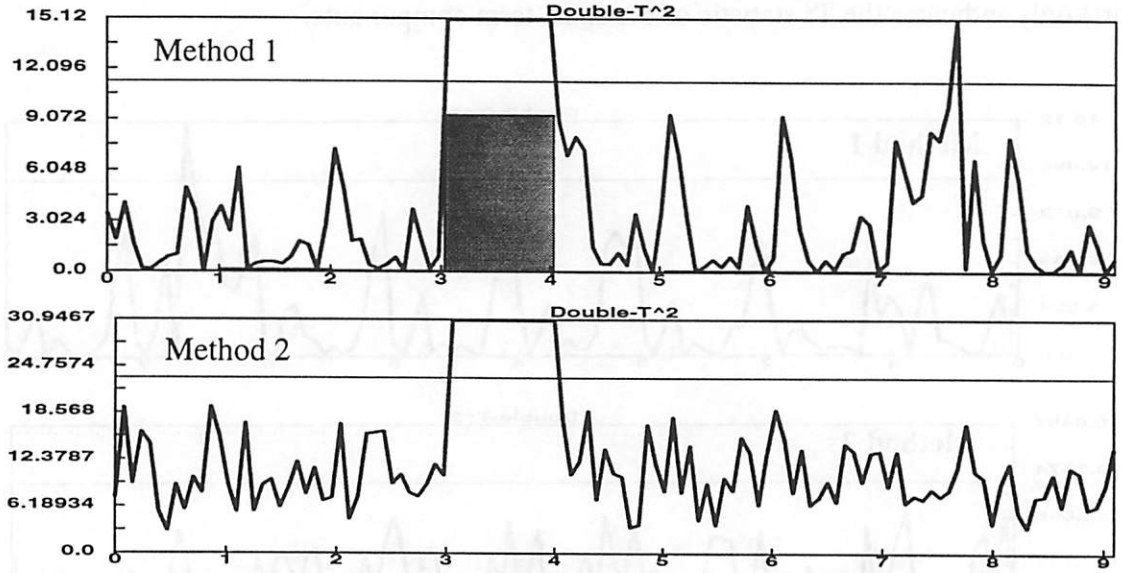


Figure 8 Double-T^2 control chart from the verification testing

4.0 Conclusion and Future Work

In this term project we have demonstrated a new approach for performing real-time SPC using full-range OES data. The entire algorithm including the PCA and the following RTSPC takes only few minutes, which is fast enough in real manufacturing sites. BCAM's RTSPC shows its fault detection capability for within-wafer variation. Two different data compression methods give us similar results, even though PCA only used three variables, compared to 10 or so variables chosen for the second method. In fact, since the first principal component captures most of the variance in the OES data, one might be able to monitor this single variable instead of using multivariate T^2 statistics. In the case of wafer-to-wafer components, our ARIMA analysis and its following T^2 statistics don't mean much because we have too few observations in this study.

Future work includes a new baseline experiment with large number of wafers (more than 20 wafers). It would improve our wafer-to-wafer ARIMA filters and give a better performance of RTSPC. Also, instead of monitoring the entire spectrum, one can collect the OES data from selective spectral range with higher resolution. For example, we can acquire a narrow spectral window that contains the endpoint wavelength (405nm for poly-Si etching). In addition, sensor fusion which integrates OES data with other real-time signals is another promising approach to conduct RTSPC. Furthermore, by including the OES data from three different spatial lens, within-wafer etching uniformity can be monitored in real-time. The results of RTSPC for etching uniformity will be included in Appendix E.

Acknowledgments: My thanks go to Mr. Shenqing Fang for allowing me to collect the data from his Microlab baseline experiment. I also wish to thank Herb Huang and Eric Boskin for their assistance with the RTSPC software.

References

[1] S. F. Lee, "Linear Regression Prediction Models of Wafer Characteristics After Plasma Processing", Master Thesis, Statistical Department, U.C. Berkeley, 1994

[2] A. Pankratz, "Forcasting with Univariate Box-Jenkins Models", New York, Wiley, 1983

[3] C.J. Spanos, H. F. Guo, A. Miller and J. Levine-Parrill, "Real-time Statistical Process Control Using Tool Data", *IEEE Transaction on Semiconductor Manufacturing*, Vol.5, No. 4, Nov. 1992, pp. 308-318

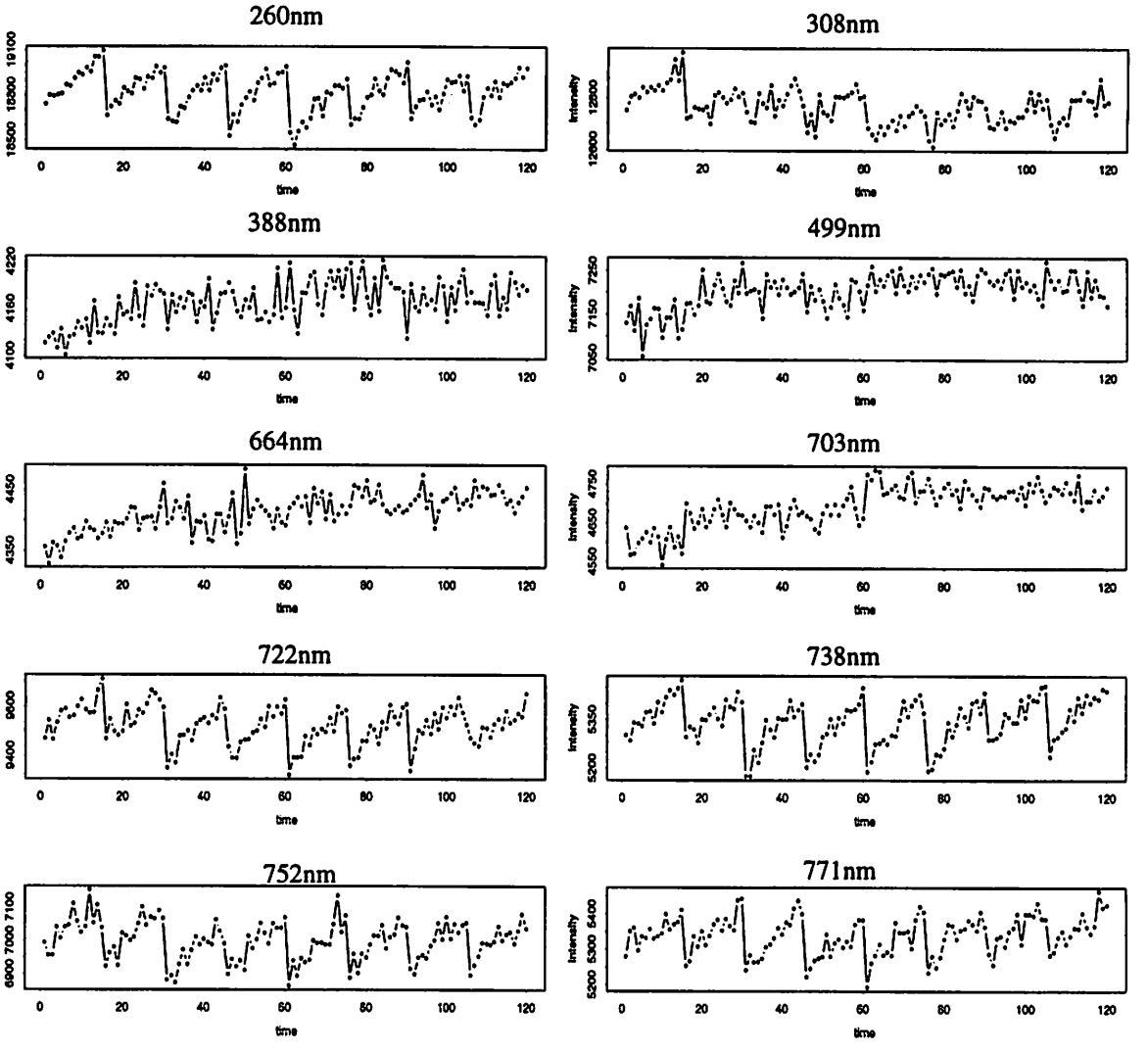
[4] H.C. Liu, "Automatic Time-series Model Generation For Real-Time Statistical Process Control", Master Thesis, EECS Department, U.C. Berkeley, 1993

[5] S.F. Lee, E.D. Boskin, H.C. Liu, E.H. Wen and C.J. Spanos, "RTSPC: A Software Utility for Real-Time SPC and Tool Data Analysis", *IEEE Transaction on Semiconductor Manufacturing*, Vol.8, No. 1, Feb. 1995, pp. 17-26

[6] D. C. Montgomery, "Introduction to Statistical Quality Control", 2nd ed., New York, Wiley, 1991

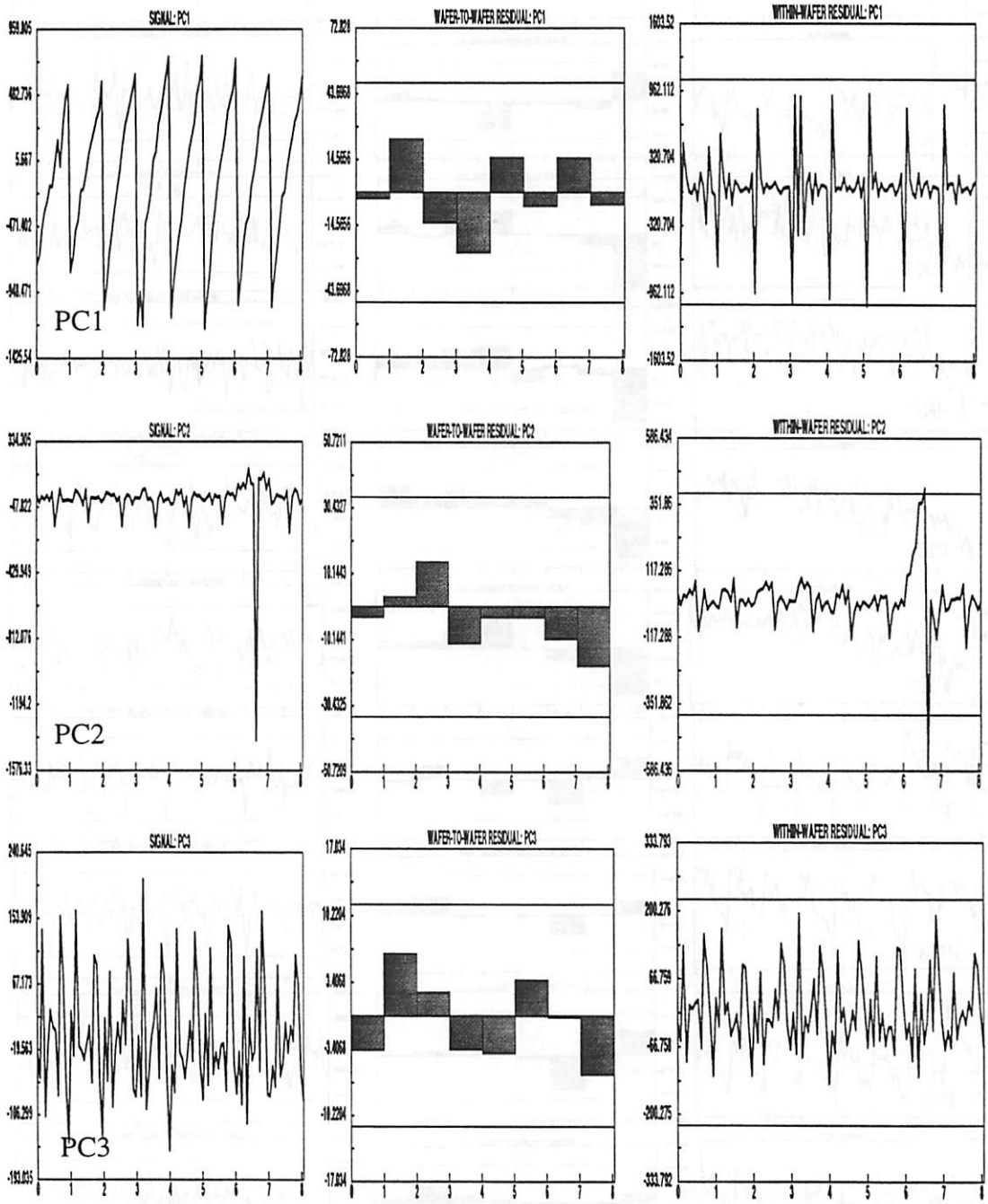
Appendix A.

The original real-time intensity data for the ten selective wavelengths

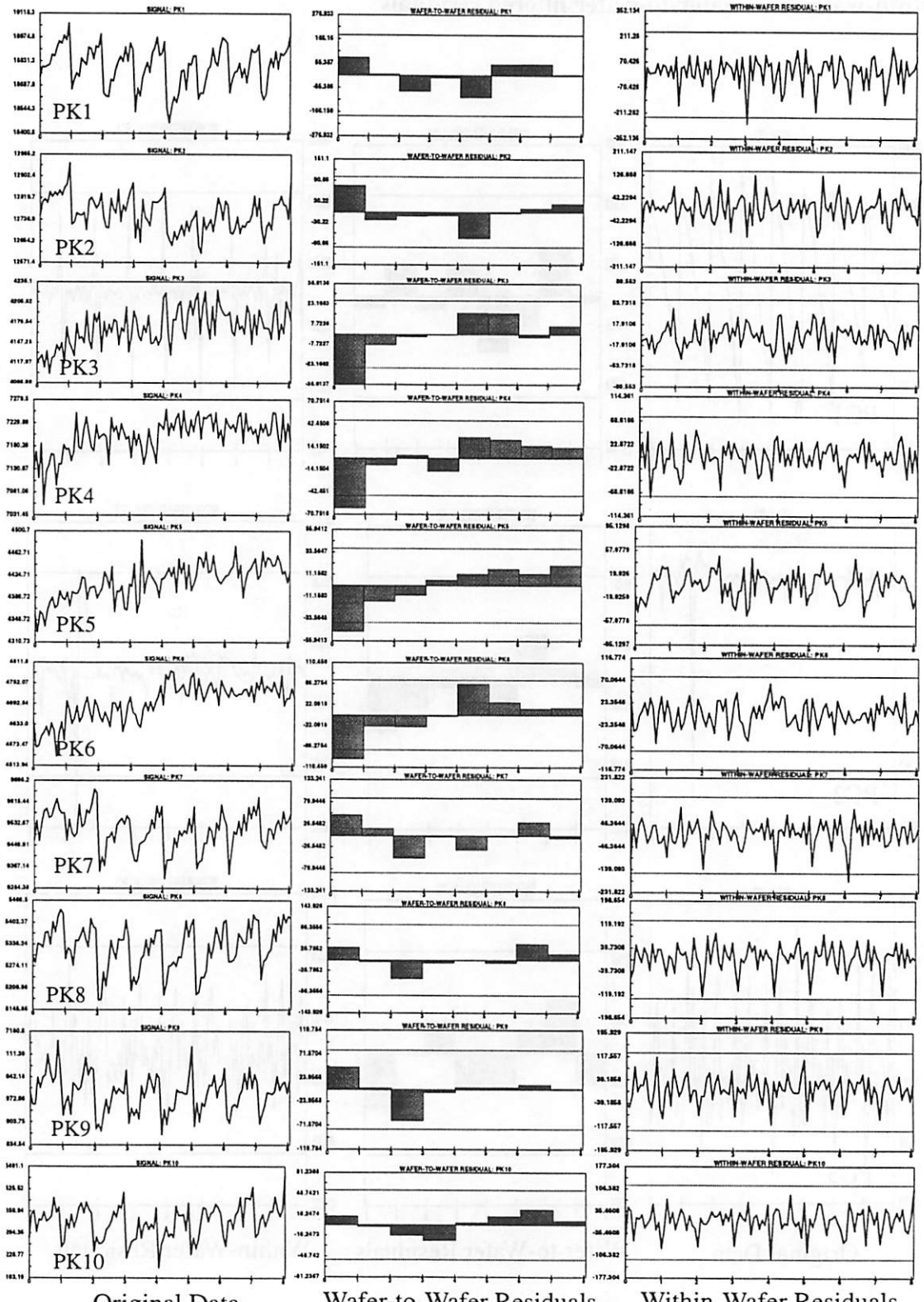


Appendix B

B.1 Baseline data for three principal components along with their corresponding within-wafer and wafer-to-wafer filtered residuals



B.2 Baseline data for ten selective variables along with their corresponding within-wafer and wafer-to-wafer filtered residuals



Appendix C. Within-wafer and wafer-to-wafer ARIMA baseline models

Method 1

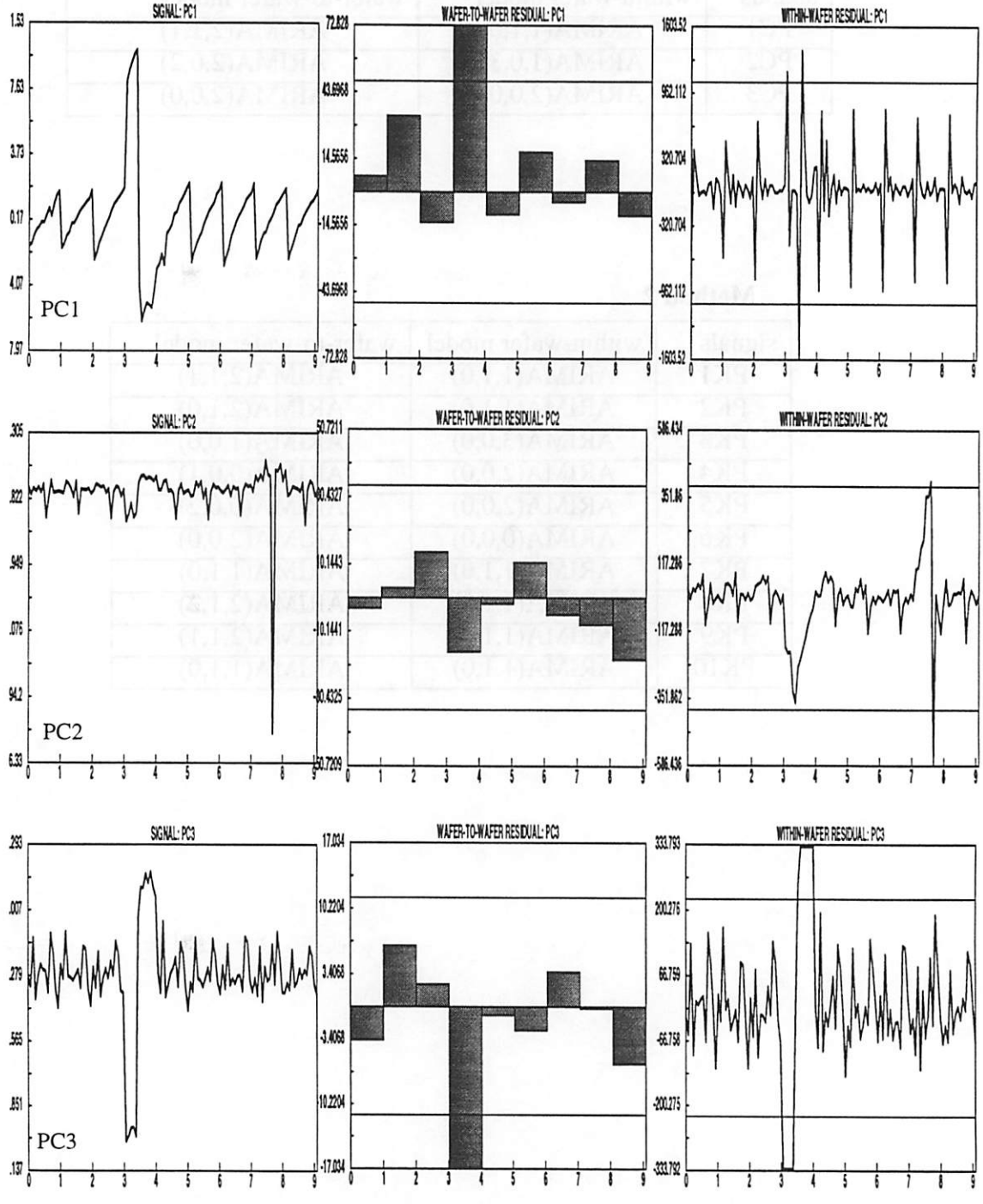
signals	within-wafer model	wafer-to-wafer model
PC1	ARIMA(1,1,0)	ARIMA(2,1,1)
PC2	ARIMA(1,0,3)	ARIMA(2,0,2)
PC3	ARIMA(2,0,0)	ARIMA(2,0,0)

Method 2

signals	within-wafer model	wafer-to-wafer model
PK1	ARIMA(1,1,0)	ARIMA(2,1,1)
PK2	ARIMA(1,1,0)	ARIMA(2,1,0)
PK3	ARIMA(3,0,0)	ARIMA(1,0,0)
PK4	ARIMA(2,0,0)	ARIMA(0,0,1)
PK5	ARIMA(2,0,0)	ARIMA(0,0,2)
PK6	ARIMA(0,0,0)	ARIMA(2,0,0)
PK7	ARIMA(1,1,0)	ARIMA(1,1,0)
PK8	ARIMA(1,1,0)	ARIMA(2,1,2)
PK9	ARIMA(1,1,0)	ARIMA(2,1,1)
PK10	ARIMA(1,1,0)	ARIMA(1,1,0)

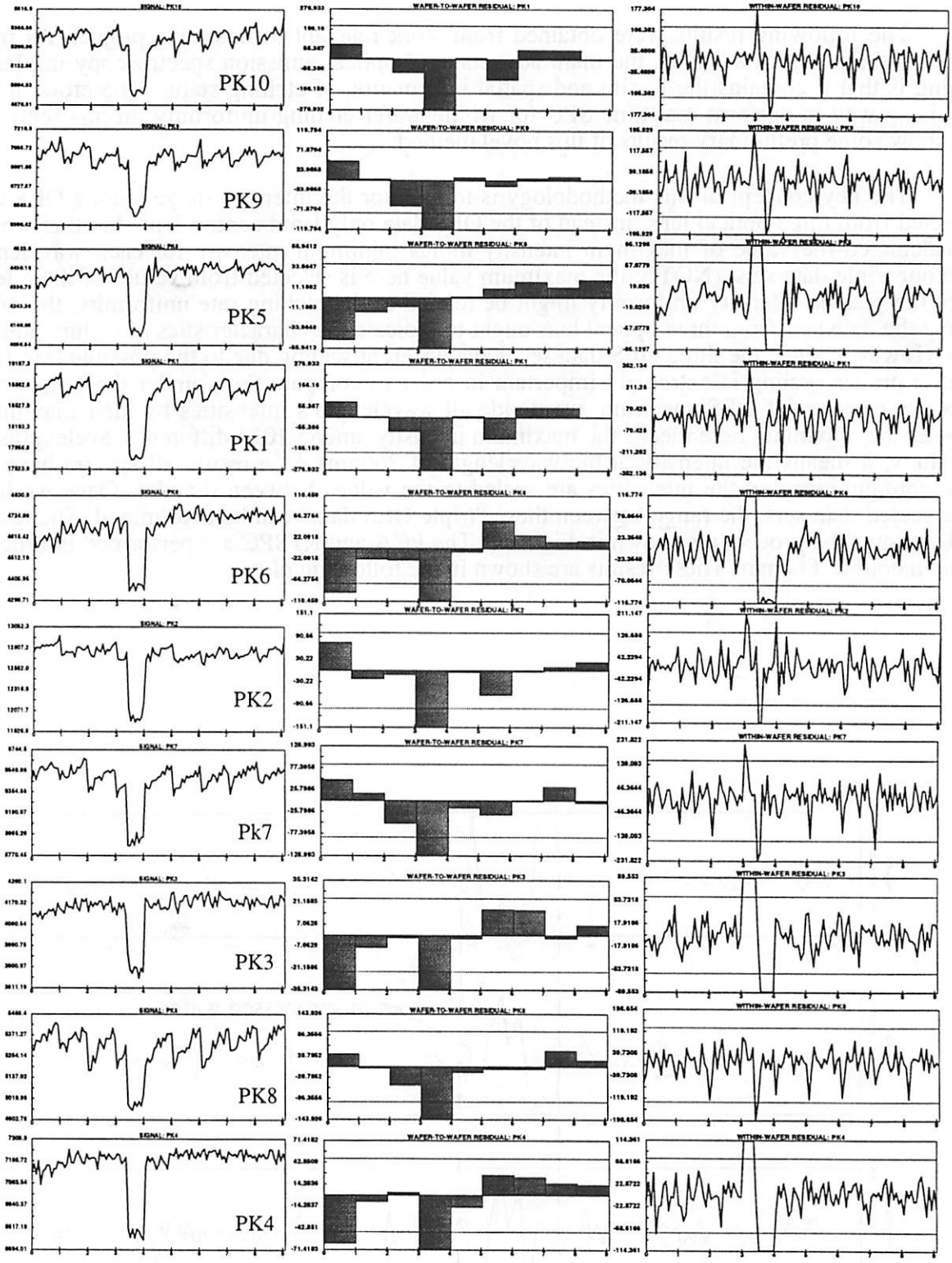
Appendix D

D.1 Original data for three principal components along with their corresponding within-wafer and wafer-to-wafer filtered residuals



Original Data Wafer-to-Wafer Residuals Within-Wafer Residuals

D.2 Original data for ten selective variables along with their corresponding within-wafer and wafer-to-wafer filtered residuals

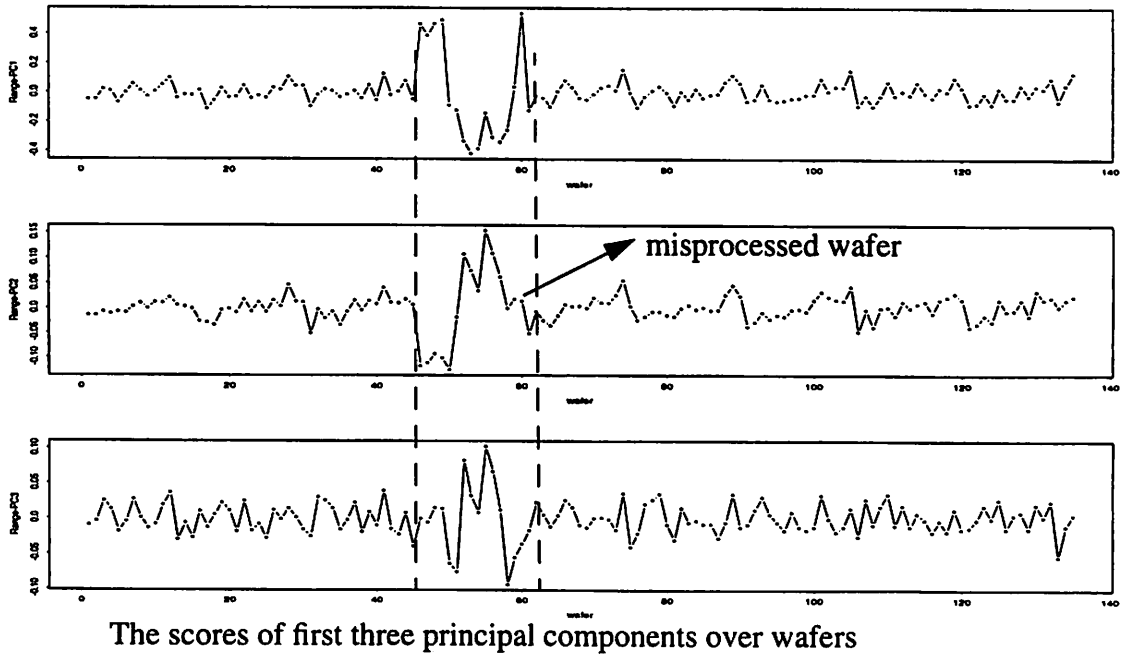


Appendix E

Real-time SPC for wafer etching uniformity using OES data

The following results were obtained from work relevant to this term project. As mentioned in the previous sections, the main advantage of optical emission spectroscopy in plasma etching is that it contains the in-situ and spatial information of etching status. Therefore, it is a promising way to perform real-time SPC for within-wafer etching uniformity. In this section, I will show some preliminary results of this novel method.

The key concept of this methodology is to monitor the intensity range among OES data collected from three optical lens, instead of the OES data only from central lens. In other words, we calculated the value of maximum intensity minus minimum intensity for each wavelength from our triple data sets. (NOTE: the maximum value here is selected from values of three lens) Since the plasma intensity uniformity might be related to the etching rate uniformity, the range among the data sets from three optical lens ought to indicate the characteristics of etching uniformity. However, since the three OES data sets have different scaling due to the possible misalignment, a proper scaling is extremely important in order to compare them under the same scale. Thus, for each set of OES spectrum, we divide all wavelength's intensities by their maximum. (NOTE: the maximum here means the maximum intensity among 1024 different wavelengths. In this study, it means the intensity at the wavelength of 260nm) As a result, all spectra have the same configuration but the intensities are scaled to the values between 0 and 1. Once we have these scaled data sets, the range between these "triple OES data" can be determined. The rest of work follows the procedures shown in Figure 1. The PCA and RTSPC are performed in order to obtain a double T^2 chart. These results are shown in the following plots.



The scores of first three principal components over wafers

Modeling Temporal Variability on a Lot to Lot Basis in Semiconductor Manufacturing Equipment

Anna M. Ison

The purpose of this project is to study the time-dependent behavior of semiconductor manufacturing equipment on a lot-to-lot basis. Real-time signals determined to be most sensitive to changes in the machine state were collected using LamStation software from a Lam TCP 9600 metal plasma etcher in an experiment conducted by Texas Instruments. The methodology developed in this paper is based on simulated data generated by combining a theoretical lot-to-lot model with wafer-to-wafer models identified from real data. Statistical process control (SPC) techniques were then employed to monitor the wafer-to-wafer, and lot-to-lot residuals respectively. In particular, to simulate machine aging, a drift component was added to the wafer-to-wafer models in order to study the sensitivity of the lot-to-lot model to drifting in the wafer behavior.

1.0 Introduction

Driven by the need to remain competitive in a growing marketplace, companies in the semiconductor manufacturing industry are striving to maintain high yield, decrease circuit geometries, and increase throughput in the face of growing costs and tighter product specifications. As the demands of the industry mount, variation in the fabrication process limits production and thus becomes more of a concern. Furthermore, the increasing complexity of microfabrication machines makes it advantageous to apply real time control procedures to reduce the effects of that variation.

A model capturing the variability of the machine is necessary before one can devise a methodology to design a controller. There are two types of variability present in the semiconductor manufacturing process - spatial and temporal. Spatial variability describes variation over some physical dimension. In the case of semiconductor manufacturing, this type of variability occurs primarily across the wafer. In contrast, temporal variability refers to variation over time. In single wafer machines, the wafers are loaded into the machine one at a time during processing. The wafers are processed in groups called lots, where the time in between processing of lots can vary substantially. Furthermore, the input settings of the machine may be changed, or some maintenance may be performed on the machine during the interim period between processing of lots. Thus, we can consider temporal variability to occur over different time scales. Specifically, variation occurs within the processing time of each wafer, from wafer to wafer, and from lot to lot.

The monitored signals used in this work were those suspected to be most sensitive to changes in the chamber state of the etcher [1]. This is important because the machine state is directly related to the state of the wafer, and of course, we are ultimately concerned with the wafer output characteristics. It has been found in the past that these observations are highly autocorrelated. One would expect time series patterns to be most evident within the processing time of a wafer

(referred to as real-time), but they have also been found on the wafer-to-wafer level [1]. It is possible that there may also be some time dependency from lot to lot. However, due to limitations imposed by the amount of actual data available at the time of this study, this behavior could not be investigated. Instead, the lot means were assumed to follow an Exponentially Weighted Moving Average (EWMA) trajectory, and the methodology developed was based on this assumption. Modeling of the time-dependent behavior on a lot to lot basis can be conducted in the future as more data becomes available. This can easily be incorporated into the framework developed in this paper.

Traditional SPC techniques such as the Shewhart quality control chart methods assume that the underlying process is stationary, i.e. that the mean and variance do not vary with time, and that the observations are identically, independently, and normally distributed (IIND) [2]. Applying these techniques directly to data which contain time series patterns will result in increased false alarm and missed alarm rates [1]. In order to deal with this situation, time series modeling can be used to filter out the time dependencies, and then traditional or multivariate SPC methods can be applied to the residuals. This is the approach taken in this work.

2.0 Methodology

2.1 Selection of Real Time Signals for the TCP Etching System

The ability of a model to describe the state of a machine is highly dependent on the data used to create the model. Much of the past work in modeling plasma etch equipment has used response surface methodology (RSM) models to map the input settings of the machine directly to output states of the wafer (etch rate, uniformity, selectivity and anisotropy). However, it has been shown in [3] that for plasma processes, electrical and mechanical signals, such as the load impedance and coil positions, provide a better description of the chamber state than do the input settings. These signals are referred to as real-time equipment signals [1].

	LamStation	Description
Bottom RF Coil	RF Tune Vane Position	Equivalent position of the tune vane position in matching network of the lower coil
	RF Load Coil Position	Equivalent position of the load coil position in matching network of the lower coil
	Line Impedance	Apparent input impedance of the lower matching network
	RF Phase Error	Phase error between the current and voltage at the bottom coil
	DC Bias	Measures the charge on the electrodes due from electrons

	LamStation	Description
Top TCP Coil	TCP Tune Vane Capacitor Position	Position of the tune vane capacitor of the matching network for the top coil
	TCP Phase Error	Phase error between the current and voltage at the top coil
	TCP Load Capacitor Position	Position of the load capacitor of the matching network for the top coil
	Line Impedance	Apparent input impedance of the upper matching network
	RF Bias	DC bias when both sources are powered
	Endpoint	Reads the intensity of the plasma in the chamber at a particular wavelength

TABLE 1. Real Time State Signals Collected for the Lam TCP 9600

The plasma source of a TCP etching system consists of two planar coils located at the top and bottom of the chamber. An induced RF field causes the gas near the coil to ionize, creating the plasma. The real-time signals found to be most sensitive to the chamber state of the Lam TCP 9600 metal etcher are related to both the top and bottom coils. The signals used in this study are summarized in Table 1 [1].

2.2 ARIMA Time Series Modeling and SPC

Univariate time series models, and in particular, ARIMA (Auto-Regressive Integrated Moving Average) models are used to account for dependencies among sequential measurements of the same signal. The assumption is that a major part of the signal's behavior can be explained by using past observations of the signal. This explanation is formalized through an ARIMA(p,d,q) model, where p is the auto-regressive order, d is the integration order, and q is the moving average order. Thus, with ϕ representing the autoregressive and θ the moving average parameters respectively, a non-stationary time series x_t can be described by an equation of the following form:

$$w_t = -\sum_{k=1}^{p} \phi_k w_{t-k} + \sum_{k=0}^{q} \theta_k a_{t-k} \tag{1}$$

where θ_0 is set to 1, $|\phi_1| < 1$, the error term $a_t \sim N(0, \sigma^2)$, and w_t are the differenced data

$$w_t = \nabla^d x_t \tag{2}$$

where ∇^d is the dth order differencing operator,

and $\nabla^1 x_t \equiv x_t - x_{t-1}$, $\nabla^2 x_t \equiv \nabla^1 x_t - \nabla^1 x_{t-1}$.

For a more thorough explanation of time series models, the reader is referred to [4] and [5].

Real-time signals are monitored in order to detect changes in machine state which are not representative of the normal operating behavior of the process. These changes indicate that a fault or malfunction may have occurred. Using SPC methods, this type of change would presumably be detected if the process were out of statistical control. However, to make such a determination requires a point of reference, and this is provided by the baseline behavior of the process. In other words, ARIMA models are derived from data taken while the process is running under normal operating conditions; this establishes the baseline behavior, and the limits under which the process is considered to be in statistical process control. The models are then used with current readings to forecast each new value of the signal. The differences between the forecasted values of the model, and the actual values of the signal are the residuals. If the process is operating under statistical control, these residuals are IIND variables by definition, and can be plotted in standard SPC charts.

Thus, after applying an iterative ARIMA modeling procedure to the collected data, the "goodness" of the model (as measured by its ability to capture the time series patterns in the signal) can be determined by examining the residuals to check that they do in fact resemble IIND variables. This was the method employed in this study.

2.3 Generation of Simulated Data Sets

In general, time series modeling requires several observations, preferably at least 50 data points. Unfortunately, the available data spanned a total of 41 wafers, but comprised only two lots. Thus, time series methods could be used to model wafer-to-wafer, but not lot-to-lot behavior. To overcome this problem, simulated data were used for the purposes of this study.

2.3.1 Expressing the EWMA as an ARIMA Model

The exponentially weighted moving average, or EWMA, is a univariate method which expresses the forecast ($\hat{z}_t$) for time t as a weighted mean of the most recent observation (z_{t-1}) and the last forecast ($\hat{z}_{t-1}$). A common form of the EWMA is given by the following equation:

$$\hat{z}_t = (1-\theta_1) z_{t-1} + \theta_1 \hat{z}_{t-1} \tag{3}$$

where $0 < \theta_1 < 1$.

Now consider an ARIMA(*0,1,1*) model given by the following equation:

$$z_t = z_{t-1} - \theta_1 a_{t-1} + a_t \tag{4}$$

and suppose that θ_1 is known. The forecast of z_t formed at time (t-1) based on (4) is

$$\hat{z}_t = z_{t-1} - \theta_1 a_{t-1} \tag{5}$$

since at time (t-1), a_t is not known and is assigned its expected value of zero. By subtracting (5)

from (4) we obtain

$$z_t - \hat{z}_t = a_t \tag{6}$$

Thus, the difference between the observed value z_t and the forecasted value $\hat{z}_t$ is the random shock a_t for time *t*. Returning to the EWMA in (3), the terms may be rearranged to get

$$\hat{z}_t = z_{t-1} - \theta_1 (z_{t-1} - \hat{z}_{t-1}) \tag{7}$$

Using (6) to substitute a_{t-1} into (7) for $z_{t-1} - \hat{z}_{t-1}$, we see that a forecast from the EWMA in equation (7) is identical to a forecast from the ARIMA(*0,1,1*) in equation (5). Thus, the EWMA may be interpreted as an ARIMA(*0,1,1*); the reader is referred to [4] for a more detailed discussion of this topic.

2.3.2 Simulating Data Sets

The equivalence between EWMA and ARIMA(*0,1,1*) models was used to simulate data based on the assumption that the lot means followed an EWMA trajectory. An identification procedure was conducted on this simulated data to estimate a lot-to-lot model. Wafer-to-wafer ARIMA time series models based on wafer averages were identified using the available baseline data, and these were combined with the lot-to-lot model to generate data representing several lots. To simulate the effect of machine aging, linear drift components were added to the wafer models in a random fashion, with a different drift occurring in each lot. A change in recipe was simulated by adding an abrupt shift in the lot means for a few consecutive lots. The generation of these data sets and the underlying assumptions are discussed in more detail in section 3.4.

3.0 Implementation

All of the identification procedures, data set simulations, and control charts were implemented using *S-PLUS* software in an *S-PLUS* environment [6].

3.1 Real-Time

The readings collected for each signal track the behavior of the process in real-time, or in other words, within the processing time of each wafer. Although time series patterns would be expected to be most prevalent at this level, the models developed in this study do not capture real-time behavior. Instead, an average value for the signal was calculated for each wafer, and these wafer averages were used to determine a wafer-to-wafer model. For results involving time-series modeling and filtering on a real-time level, the reader is referred to [1].

3.2 Wafer-to-Wafer

On the wafer-to-wafer level, where each point in the series was an average value for the signal over a wafer, analysis of the data demonstrated that although no significant autocorrelations were visible in the majority of the monitored signals, two signals were found that exhibited time dependent behavior. Standard control charts could be applied directly to many of the signals with no correlated behavior. Since this is not such an interesting case, this work focuses on one of the two

signals demonstrating time series patterns, namely the *RF Tune Vane Position*. The methodology for analysis is fully developed for this signal, and can be generalized to other signals exhibiting time series behavior, such as the *RF Load Coil Position*. However, specific results for the latter signal are not included in this work.

Using baseline data taken from the machine under normal operating conditions, an ARIMA time series model was identified for the *RF Tune Vane Position* signal, and the residuals were used to construct an xbar chart. The data set consisted of a total of 38 points corresponding to 38 wafer averages. The model was estimated using the first 30 points so that the last 8 points could be compared against a forecast of the model.

3.3 Lot-to-Lot

On the lot-to-lot level, each point in the series would represent the average value for the signal over a lot. The available data comprised only two lots, and hence only two lot averages - or two points in the series. Since the analysis could not be continued based on only two data points, the lot averages were assumed to follow a known trajectory given by an EWMA. The series of points representing the lot averages was generated by computer simulation using an ARIMA*(0,1,1)* model with a moving average parameter $\theta_1 = 0.6$. The lot-to-lot variance was assumed to be a multiple of the wafer-to-wafer variance. More specifically, the wafer-to-wafer variance for the *RF Tune Vane Position* signal was calculated from the data to be $\sigma^2_{wafer} = 232$. The lot-to-lot variance for this signal was chosen to be about seven times this value, namely $\sigma^2_{lot} = 1600$. This generated series of points representing 50 lot averages was then used as a data set to estimate an ARIMA model.

3.4 Generating Data for Different Cases

To simulate data collected from the machine under other conditions representing real scenarios that might be encountered, two cases were considered. First, drifting behavior as a result of machine aging was assumed to occur at a wafer-to-wafer level; a drift at the lot-to-lot level would then arise from the cumulative effect of this wafer-to-wafer drift. Secondly, an abrupt shift in lot averages might result from maintenance performed, or from changing the input settings of the machine (changing the recipe) between lots.

3.4.1 Adding a Random Drift Component

A linear drift was added at the wafer-to-wafer level by the following mechanism:

$$x_{drift}(i) = x(i) + mi \tag{8}$$

where x is a series of wafer averages, i denotes the index corresponding to the wafer number (i.e. for 30 wafers, $i = \{1, 2, ..., 30\}$) and m is the slope of the linear drift. Thus, because we considered a sample size of 50 lots with 30 wafers per lot, 50 values of m had to be chosen corresponding to the wafer-to-wafer drifts in each lot. The values of m were determined by a random number generator operating on a normal distribution. A small drift component was determined by choosing m from ~N(0.5,0.25). The non-zero mean ensures that a cumulative drifting effect will occur over the lot-to-lot behavior. In a similar manner, a large drift was simulated by choosing m from ~N(0,4). The mean was chosen to be zero in this case because the drifts from wafer-to-wafer were already quite large, so that adding any cumulative effect over lots would just amplify the

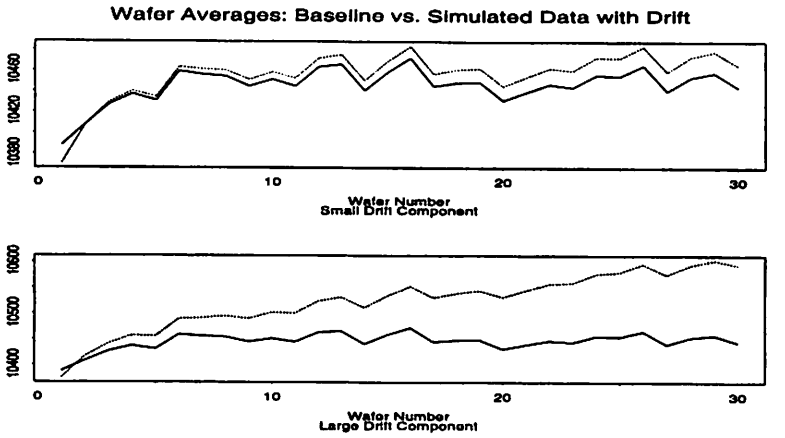


FIGURE 1. Wafer Averages: Baseline (solid) vs. Simulated Data with Small/Large Drifts

magnitude of these drifts. Examples of typical values for m corresponding to a small drift (m = 0.7), and a large drift (m = 5) at the wafer-to-wafer level are displayed in Figure 1.

The final version of the simulated data combining wafer-to-wafer and lot-to-lot models is given by

$$y_{drift}[i, k] = y(k) + m(k) \times i \tag{9}$$

where y is a series of lot averages, k denotes the index corresponding to the lot number (i.e. for 50 lots, k = {1, 2,..., 50}), $m(k)$ is the slope of the wafer-to-wafer drift for lot k, and i is the index denoting the wafer number (i.e. for 30 wafers, i = {1, 2,..., 30}). The simulated data, $y_{drift}[i, k]$, is in matrix form, with each row representing a different wafer, and each column corresponding to a different lot. To plot the data in the proper time sequence, one would concatenate the columns together, so that one set of 30 points (wafer averages) would be followed by another set of 30 points, over a total of 50 sets (each representing a lot). The result of this concatenation for the small and large drift components added at the wafer-to-wafer level are shown in Figure 2, which plots the resulting lot averages for these two cases.

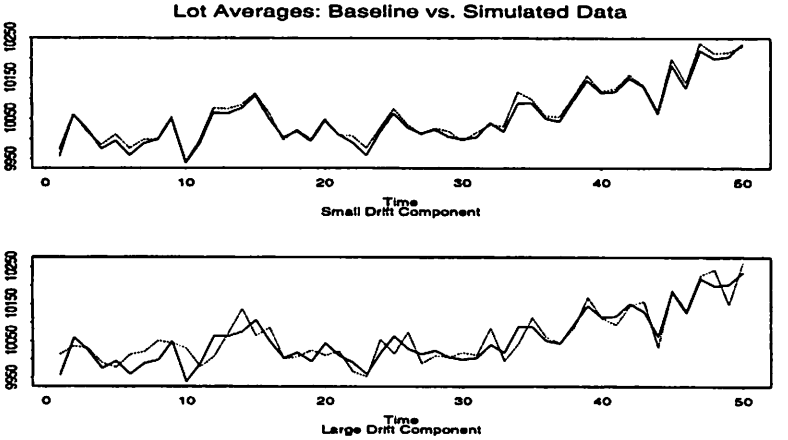


FIGURE 2. Lot Averages: Baseline (solid) vs. Simulated Data with Wafer-to-Wafer Drifts

3.4.2 Adding an Abrupt Shift

An abrupt shift in this context refers to a sudden jump in value in the trajectory of the signal. This kind of effect might be seen from one lot to another if the input settings for the machine were changed (recipe change), or if maintenance were performed on the machine in between processing of lots. Data was generated to mimic the effect of a recipe change or maintenance event by arbitrarily adding a constant term to consecutive lot averages for a selected number of lots. Figure 3 shows the result of adding an offset of +50 to the average values in lots 8 through 15, an offset of -250 to lots 40 to 43, and an offset of +375 to the average value of lot number 26. This data set was analyzed against the standards set by the baseline data to determine whether these shifts would be detected using SPC.

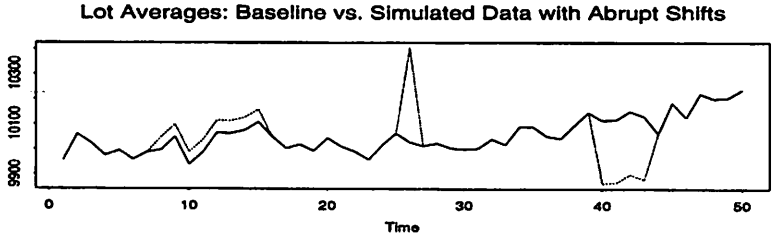


FIGURE 3. Lot Averages: Baseline (solid) vs. Simulated Data with Abrupt Shifts

4.0 Results and Discussion

4.1 Wafer to Wafer Results

4.1.1 Time Series Modeling

An ARIMA time series model was identified for the *RF Tune Vane Position*, which exhibited autocorrelated behavior in its wafer averages. The model was found to be an ARIMA(*0,0,1*), with a moving average parameter $\theta = -0.292$. The model was identified using the first 30 wafer averages, so that the remaining 8 points could be compared against a forecast using the model. This is depicted in Figure 4, which plots the forecast of the model with the upper and lower two standard deviation limits. From the plot we can conclude that the forecasting capability of the model is acceptable based on a comparison with the actual data. The residuals of this model are displayed on a quantile-quantile plot which is a graphical display to test the distribution of the data. In this case, the data are tested against the normal distribution. If the plot is approximately a straight line, then the residuals follow a normal distribution. The residuals for this model are shown to satisfy the IIND criteria required for SPC as demonstrated by the quantile-quantile plot in Figure 4.

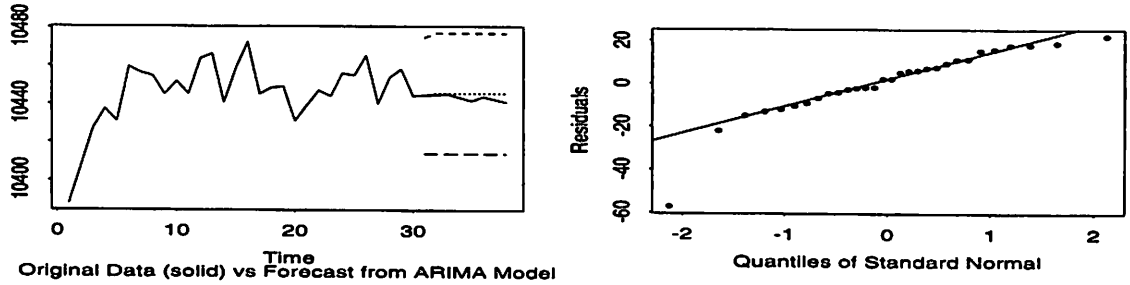


FIGURE 4. (a) Original Data (solid) vs. Forecast from ARIMA Model and (b) Residuals

4.1.2 Fault Detection

SPC techniques were used on the residuals of the wafer-to-wafer model described in the previous section to establish limits (based on baseline data) for an xbar chart at the 95% confidence level. This chart is plotted in Figure 5. Data sets corresponding to small and large drift components added at the wafer-to-wafer level were filtered by this model, and the resulting residuals plotted on the xbar chart with limits established by the baseline data. Figure 6 shows the residuals corresponding to the data plotted in Figure 1 for typical values corresponding to small and large drift components at the wafer-to-wafer level. In both cases, it is evident that the drift component is detected successfully by the out-of-control points visible on the xbar charts in Figure 6.

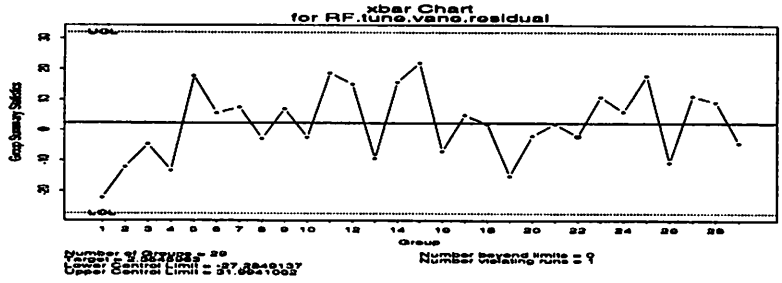


FIGURE 5. Baseline Xbar Chart for RF Tune Vane Position Wafer-to-Wafer Residuals

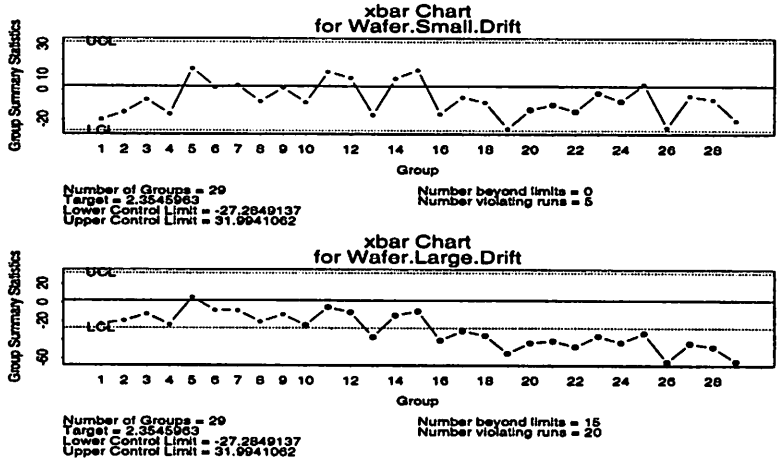


FIGURE 6. Xbar Charts for Wafer Residuals - Small (top) and Large (bottom) Drifts

4.2 Lot to Lot Results

4.2.1 Time Series Modeling

Because of an inadequate amount of actual data available for lot-to-lot modeling, simulated data representing 50 lot averages generated by an EWMA model with a theoretical moving average parameter $\theta_1 = 0.6$, and lot-to-lot variance of $\sigma^2_{lot} = 1600$, were used to estimate an ARIMA model.

The identified lot-to-lot model was found to be an ARIMA(*0,1,1*), with a moving average parameter $\theta_1 = 0.54$. Thus, the modeling procedure was found to perform adequately, with the estimated model closely approximating the original theoretical model used to generate the data. The series of points generated by these two (original and identified) models is plotted in Figure 7.

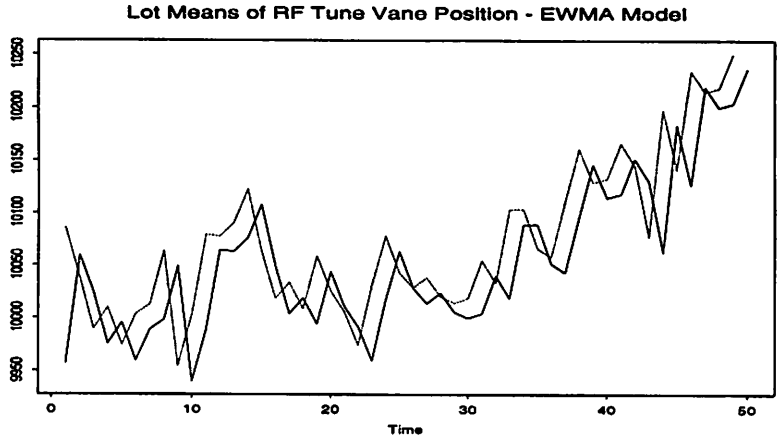


FIGURE 7. Original Lot-to-Lot Data (solid) vs. Simulated Data from Identified Model

4.2.2 Fault Detection

SPC techniques were used on the residuals of the lot-to-lot model described in the previous section to determine the upper and lower control limits for an xbar chart at the 95% confidence level for the baseline data. This chart is plotted in Figure 8. Data sets corresponding to small and large drifts added at the wafer-to-wafer level as described in section 3.4.1 and plotted in Figure 2 were filtered through the lot-to-lot model. The residuals were then plotted against the xbar chart established by the baseline data; these are depicted in Figure 9. Note that the small drift component, which was easily detected at the wafer-to-wafer level, is more difficult to detect at the lot-to-lot level. In contrast, for the larger drift components, the effect on the lot averages was so strong that deviations were clearly visible on the xbar chart even on a lot-to-lot level.

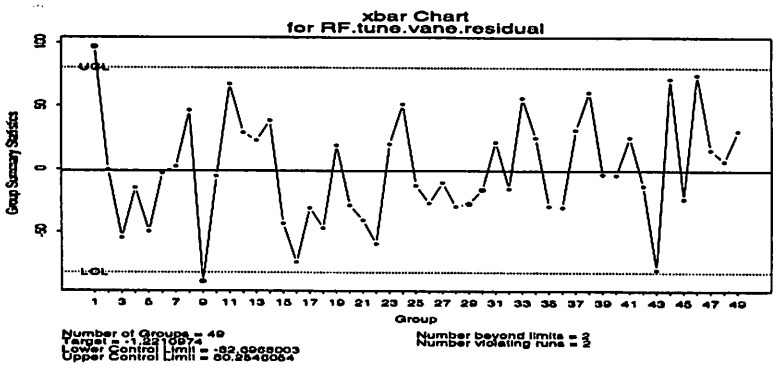


FIGURE 8. Xbar Chart for Lot-to-Lot Residuals - Baseline Data

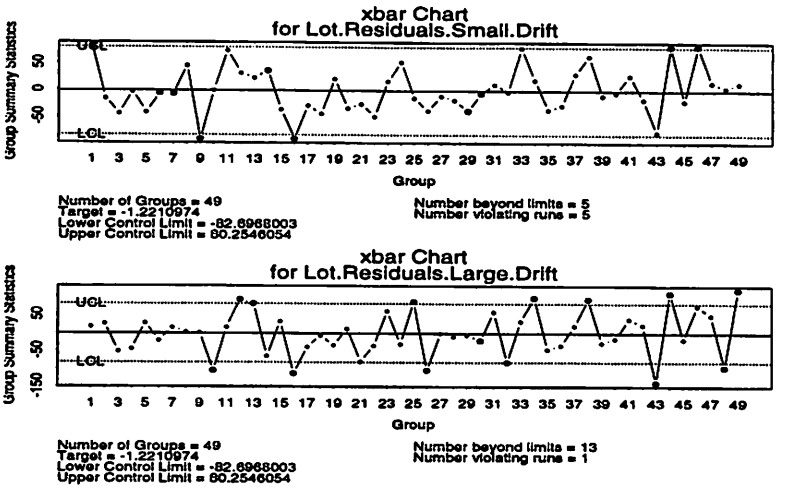


FIGURE 9. Xbar Chart for Lot-to-Lot Residuals - Small and Large Wafer-to-Wafer Drifts

Finally, the data injected with abrupt shifts at the lot-to-lot level (described in section 3.4.2 and plotted in Figure 3) were filtered by the lot-to-lot model, and the resulting residuals plotted in an xbar chart shown in Figure 10. This demonstrates that on a lot-to-lot level, abrupt shifts in lot averages arising from a change in recipe or maintenance would be very effectively detected using this methodology.

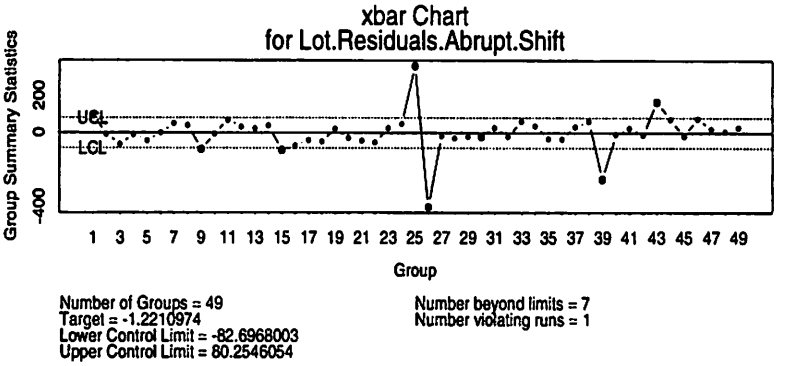


FIGURE 10. Xbar Chart for Lot-to-Lot Residuals with Abrupt Shifts in Lot Averages

5.0 Conclusions and Future Work

By monitoring real-time signals believed to best reflect the chamber state, the time-dependent behavior of a metal plasma etcher was studied on both wafer-to-wafer and lot-to-lot levels. Due to a lack of data on the lot-to-lot level, simulated data were generated by combining an assumed lot-to-lot model with wafer-to-wafer models identified from real data. ARIMA time-series models were used to filter out the time-dependent patterns in the signal. The resulting residuals were then used to establish xbar charts based on the baseline data. Data sets were generated to simulate common phenomena such as machine aging, maintenance, or changes in recipe, in order to study the effectiveness of the xbar charts in detecting these events on wafer-to-wafer and lot-to-lot levels. In general, the small drift component was more visible at the wafer-to-wafer level. Since these drifts were added at the wafer-to-wafer model, this result is not surprising. On the lot-to-lot level, large drifts in wafer behavior were easily detected, as were abrupt shifts in the lot averages.

Several assumptions were made in order to conduct this analysis. First, the wafer-to-wafer model was assumed to remain the same from lot to lot, changing only by an offset given by the lot averages. Secondly, it was assumed that the lot averages followed an EWMA trajectory. This assumption could be relaxed in the future. However, we must still assume that the lots are processed closely enough in time that a time-dependency will exist; otherwise there will be no time-dependent behavior to model. It was also assumed that a linear drift would occur on the wafer-to-wafer level, and that drifting on the lot-to-lot level would be the result of a cumulative effect of this wafer-to-wafer drift. Finally, it was assumed that machine maintenance and recipe changes would take place in between processing of lots, and that this would cause an abrupt jump in value in the lot averages.

Future work will focus on applying this methodology to actual data obtained over several lots. However, due to potentially large time intervals between processing of lots, in addition to maintenance events and recipe changes which may take place during those intervals, we would not expect the same kind of time-dependency to occur from lot to lot as was found for wafer to wafer, and within the processing time of a wafer. The assumptions made above will have to be checked against results obtained from the actual data. If the lot averages are found to follow some kind of time-series behavior, then this methodology should be able to identify significant changes which shift the lot averages. Using actual data, we can check the sensitivity of the SPC techniques at a lot-to-lot level to a real change in the input settings of the machine, or a real maintenance event. If successful, a next step would be to employ multivariate techniques such as Hotelling's T-squared statistic to capture correlations among the different signals at a lot-to-lot level. However, depending on the time series behavior of the actual lot averages, a different methodology may be required to deal with variation on a lot-to-lot basis.

Acknowledgments

I wish to acknowledge Sematech and Texas Instruments for providing the experimental data used in this study. I would also like to thank the members of the Berkeley Computer Aided Manufacturing group, especially Xinhui Niu and Sean Cunningham for help in using the *S-PLUS* software. Finally, I would like to thank Professor Costas Spanos for his guidance throughout this project, and his reviews of the report drafts.

References

[1] S. F. Lee, *Semiconductor Equipment Analysis and Wafer State Prediction System Using Real-Time Signals*, Ph.D. Thesis, University of California, Berkeley, Memorandum No. UCB/ERL M94/104, Dec. 1994.

[2] D. C. *Montgomery, Introduction to Statistical Quality Control*, 2nd ed., John Wiley & Sons, 1991.

[3] C. J. Spanos, H. Guo, A. Miller, J. Levine-Parrill, "Real-Time Statistical Process Control Using Tool Data," *IEEE Trans. Semiconductor Manufacturing*, vol. 5, no. 4, Nov. 1992.

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

[5] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, 1994.

[6] *S-PLUS User's Manual*, version 3.0, Statistical Sciences, Inc., Seattle, WA, Sept. 1991.

In-Situ Poly Gate Photoresist Metrology and Control

Xinhui Niu

Photoresist film thickness and photo-active compound concentration are two important quantities in photolithographic process. In this work, metrology is developed for measuring them *in-situ* during gate polysilicon patterning. Experimental design is used to model the spin-coat equipment for run-to-run process control. Finally, an EWMA procedure is applied for baseline monitoring and control.

1.0 Introduction

To go beyond classical Statistical Process Control (SPC) in semiconductor manufacturing, we must develop effective methods that are capable of quickly detecting and responding to the changes in the performance of the photolithographic process. By combining statistical process control and feedback/feed-forward control, a run-to-run control application can provide more accuracy and flexibility. In this work we focus on the run-to-run control application in a photoresist coating process.

This work consists of three parts. First we develop a method for extracting the film information *in-situ*, using adaptive simulated annealing as the optimization computation engine. Fast and accurate real-time or quasi real-time metrology is the foundation of run-to-run control. Then we build an accurate regression model for the coater, considering the machine aging. This model will be used for run-to-run coat recipe design. The coating process exhibits autocorrelated observations, so an Exponentially Weighted Moving Average (EWMA) control chart is designed for baseline statistical process control.

2.0 Methodology

2.1 *In-situ* automated metrology for photolithography

In-situ processing information is very important to real-time control. Most film thickness measurement systems require that the wafer be brought to the instrument. A metrology which can measure the photoresist thickness and photo-active compound concentration (PAC) has been developed by Sovarong Leang, et al. [7]. However, like other metrologies, the operation of this method needs specific prior information about all the films. This requirement will be impossible when we have many variations in previous process steps, especially when we have more than two films. So the first task of this project is to develop metrology to extract all the interesting information of any material films. This metrology belongs to the reflectometry category.

Reflectometry is a fast, accurate and convenient way to measure film thickness, reflective indices and other characteristics on the wafer. It is convenient for monitoring IC processes for

several reasons: high good spatial resolution, high throughput, high precision, accuracy and can be easily automated. For example, in this work we use a modified SC Inspector [17]. Features on the order of 2 mm can be measured. Because of the high data acquirement rate (0.25 seconds), the measurement rate is limited by the extraction computation. Most importantly, once the equipment is set up, the measurement is fully automated, and communication links between the computer and the instrument eliminate the need for human intervention.

Interference of light waves reflected from each interface of a multi-layer film structure determines the structure's reflectance [1]. A typical film structure is shown in FIGURE 1.

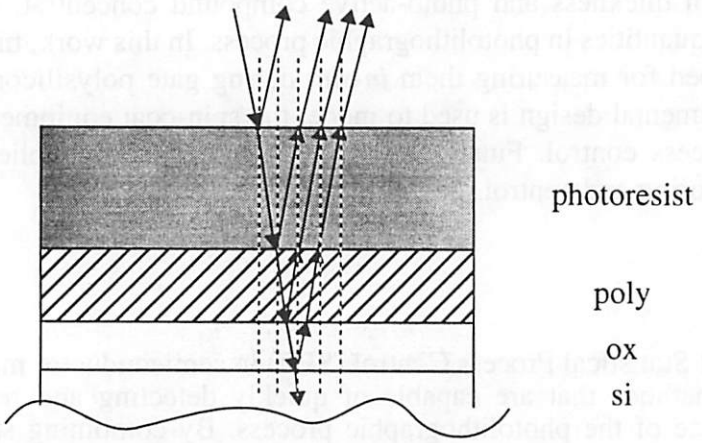


FIGURE 1 Multi-layer film example in the reflectometry.

Each film has its own index of refraction n and absorption coefficient k. In this work, we use Cauchy coefficients Na and Nb to calculate n,

$$n = Na + \frac{Nb}{\lambda^2} \quad . \tag{1}$$

For photoresist, k is determined by Dill's A and B parameters,

$$k = \lambda \frac{(A(\lambda) PAC + B(\lambda))}{4\pi} \tag{2}$$

where $A(\lambda)$ is the bleachable absorption, and $B(\lambda)$ is the nonbleachable absorption [12].

The optical properties of a layer of film are described by its characteristic matrix M. Assuming that, after careful alignment, the incident angle is approximately equal to 0, then

$$M = \begin{bmatrix} \cos(k_0 nl) & \frac{1}{i \cdot n} \sin(k_0 nl) \\ \frac{n}{i} \sin(k_0 nl) & \cos(k_0 nl) \end{bmatrix} \tag{3}$$

where n is the index of refraction, l is the film thickness, $k_0 = \frac{2\pi}{\lambda}$. The characteristic matrix of a stack of N films is then

$$M = \prod_{j=1}^{N} M_j \ . \tag{4}$$

Assume that the two end films are semi-infinite, in other words, the thickness values of the air and silicon substrate are ∞, the reflectivity of the entire stack is

$$R = \frac{(M_{11} + M_{12} n_{si})\, n_{air} - (M_{21} + M_{22} n_{si})}{(M_{11} + M_{12} n_{si})\, n_{air} + (M_{21} + M_{22} n_{si})} \ . \tag{5}$$

The relative value to the reflectivity of bare silicon (including a typical 2.5~4.5 nm of native oxide) will be the theoretical reflectivity for each wavelength. Reflectance spectrograph can be measured by the Inspector from 320 nm to 820 nm wavelength range. Measured and theoretical curves can be matched by optimizing the film thickness, the PAC of the resist, etc. The problem is formulated as

$$\min \left\{ \text{cost} = \sum_{\lambda=\lambda_1}^{\lambda_2} (R_{\text{measured}} - R_{\text{theoretical}})^2 \right\} . \tag{6}$$

In the typical problem of FIGURE 1, we will have following parameters to be optimized:

Table 1

NO.	Parameter	Description	Range
1	PAC	PAC of photoresist	[0, 1]
2	PhTh	thickness of photoresist	[1100, 1300] nm
3	PhNa	*Na* of photoresist	[1.57, 1.62]
4	PhNb	*Nb* of photoresist	[11500, 13500]
5	PolyTh	thickness of poly	[480, 660] nm
6	OxTh	thickness of oxide	[90, 120] nm

The selection of these parameters will be discussed in section 3.2.

Because there exist local minima in the optimization, traditional optimization algorithms are not appropriate here. FIGURE 2 shows an example of the local minimum phenomena in the optimization process. In this example, we have 2 parameters to be optimized, the thickness of photoresist and poly. It is known that for this process, the range of photoresist thickness is between 1100 nm and 1300 nm, and the range of the poly thickness is between 480 nm and 660 nm. In this example we can clearly see several local minimums. To find the global minimum, in this work we use stochastic methods, such as simulated annealing.

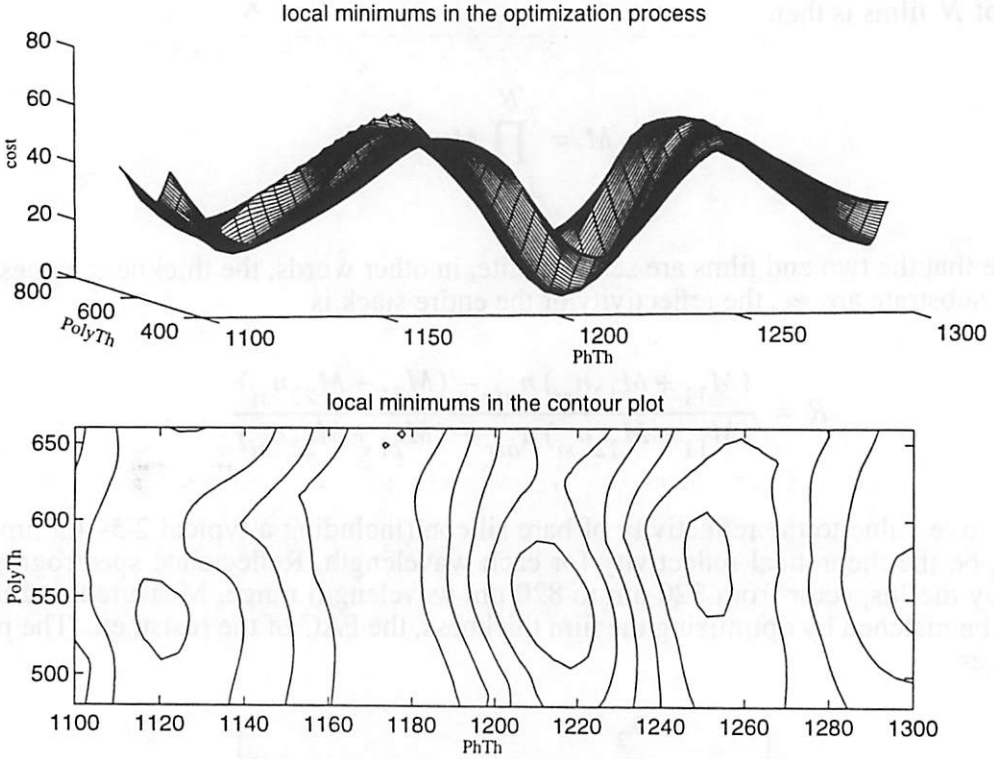


FIGURE 2 Local minimums in the optimization.

2.2 Simulated annealing

Simulated Annealing (SA) exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure and the search for a minimum in a more general system.

SA's major advantage over other methods is its ability to avoid becoming trapped at local minimums. The algorithm employs a random search, which not only accepts changes that decrease the objective function, but also some changes that increase it. In general, simulated annealing consists of four functional relationships:

1. $g(x)$: the probability density function of the state-space of D parameters

   $x = \{x_i, i = 1 \ldots D\}$

2. $h(x)$: the probability density function of accepting a new value given the last examined value.
3. $T(k)$: the schedule of "annealing" the "temperature" T in annealing-time steps k.
4. $f(x)$: the objective function.

Generally SA optimization methods choose new points at various distances from their current point x. Each new point is generated probabilistically according to a given generating distribution $g(x)$. Then the algorithm calculates the objective function $f(x)$ and probabilistically decides whether to accept it. The new point may be accepted even if it has a larger function value than the

current point. The criteria for acceptance are determined by an acceptance function $h(x)$, the temperature parameter T and the difference in the function values of the two points. As the optimization progresses, T is reduced, thus lowering the probability that the acceptance function will accept a new point if its functional value is greater than that of the current point.

Based on above, SA can:

1. deal with arbitrary degrees of nonlinearities, discontinuities, and stochasticities;
2. deal with arbitrary boundary conditions and constraints associated with the cost functions;
3. statistically guarantee finding an optimal solution.

It is important to note that first, the cost function (either a sum of squares of the difference between the model and data, or a maximum likelihood function) can be stochastic simply because one of the parameters is defined to be generated by a random number generator. If the cost function is stochastic, then derivatives often will not exist, and so those algorithms that rely on derivative information cannot be used; second, a statistical guarantee for simulated annealing means that the total space will be truly sampled, although the time could be infinite and the requirement of machine precision could be unrealistic.

The primary criticism is that SA could be impractical slow if not used appropriately. In this work, we use a very powerful SA public software, which is called "Adaptive Simulated Annealing (ASA)", developed by Lester Ingber [6]. ASA is found to be at least exponentially faster than other simulated annealing algorithms [5]. Several optimization parameters are tuned to enhance SA performance in this work, as will be discussed later in this report.

2.3 Equipment modeling

The characterization of IC processes through modeling has become a necessity in semiconductor manufacturing. The models may be physical, empirical or a combination of both. In this work, an equipment model is derived using a statistically designed experiment and linear regression analysis [2][15]. The quality of the model is tested by its R^2 value and residual plot.

2.4 EWMA for autocorrelated measurement

Traditional statistical process control techniques, like the Shewhart family, have a fundamental assumption that the observations are uncorrelated [9]. In other words, a reasonable model for the observations from the process is

$$X_t = \mu + \varepsilon_t \tag{7}$$

where μ is the process mean and ε_t is a sequence of independently and identically distributed random variables. In a photo-lithographic process, wafers are usually processed in a pipeline fashion[1]. For example, the throughput of a coater is typically around 30~60 wafers/hour. The high sampling frequency makes the observations autocorrelated.

Time series modeling is one of several approaches dealing with autocorrelated data. In this approach, one first fits an appropriate time series model to the observations and then applies con-

1. Our metrology needs less than half of the longest cycle time in the original pipeline, so introducing the *in-situ* metrology does not bring any pipeline delay.

trol charts to the stream of residuals from this model. The typical time series model employed is the autoregressive integrated moving average (ARIMA) model. However, in practical control, using this approach is frequently awkward, especially when many control variables are in question. Developing an explicit ARIMA model for each variable of interest is potentially time-consuming. One alternative approach is exponentially weighted moving average (EWMA).

The EWMA can be used as the basis of a control chart. The EWMA is defined as

$$Z_t = \lambda X_t + (1-\lambda) Z_{t-1} \tag{8}$$

where $0 < \lambda \leq 1$. The EWMA can be used in certain situations where the data are autocorrelated. Actually it can be shown that ARIMA(0,1,1) can be interpreted as an EWMA. If $\hat{X}_{t+1}(t)$ is the forecast for the observation in period $t+1$ made at the end of period t, then

$$\hat{X}_{t+1}(t) = Z_t \ , \tag{9}$$

and the sequence of one-step-ahead prediction errors

$$e_t = X_t - \hat{X}_t(t-1) \tag{10}$$

are independently and identically distributed with zero mean and standard deviation σ_p, assuming that the underlying process is really IMA(1,1). Therefore, control charts could be applied to these one-step-ahead prediction errors. In addition, even If the process is not exactly IMA(1,1), and if the observations from the process are positively autocorrelated and the process mean does not drift too quickly, the EWMA with an appropriate value for λ will provide an excellent one-step-ahead predictor [10].

One procedure for constructing an EWMA control chart is described by Montgomery and Mastrangelo as following[10]:

1. Choice of λ: select the value of λ that minimizes the sum of the squares of the one-step-ahead prediction errors.
2. Estimation of σ_p: if λ is chosen as suggested above over n observations, then calculate the variance by

$$\sigma_p^2 = \frac{\sum e_t^2}{n} \ . \tag{11}$$

3. First control chart: plot one-step-ahead EWMA prediction errors.
4. Second control chart: plot original observations on which the EWMA forecast is superimposed.

Assume that the one-step-ahead prediction errors $e_t \sim \text{IIND}(0, \sigma_p)$, then for the first control chart, the control limits are

$$\begin{aligned} UCL_{t+1} &= w\sigma_p \\ LCL_{t+1} &= w\sigma_p \end{aligned} \quad ; \tag{12}$$

for the second chart, the control limits are

$$\begin{aligned} UCL_{t+1} &= Z_t + w\sigma_p \\ LCL_{t+1} &= Z_t - w\sigma_p \end{aligned} \quad , \tag{13}$$

where w is the "distance" of the control limits from the center line, expressed in standard deviation units and determined by the type I error of the control chart.

3.0 Results and Discussion

In order to demonstrate the application of run-to-run process control, several experiments have been designed and carried out.

3.1 Experimental setup

1. Wafers: p-type, 4-inch silicon.
2. Oxidation, about 100 nm of oxide[2].
3. LPCVD, about 600 nm of polysilicon.
4. Equipment: SVG8626/36 photoresist coater track, SC Inspector.
5. Photoresist: OCG820 G-line positive photoresist.

3.2 *In-situ* film thickness extraction

Two of the top goals of designing an *in-situ* sensor are accuracy and speed. Even through simulated annealing can guarantee finding an acceptable solution, the computational cost for practical accuracies might be prohibitive. Several techniques are applied to shorten the optimization time.

Accurate material coefficients are required for the optimization cost function. Unfortunately we only know the standard (or textbook) values for these coefficients. The actual values have some deviations from the standard values. For example, the A-B-C values of the photoresist change over the time, and different texts give different Cauchy coefficients *Na* and *Nb* for Silicon, etc. To use the ASA algorithm efficiently, it is important to narrow the number and range of free variables.

Even though the material coefficients differ from the standard values, the form of the relative relationship over wavelength is assumed known. Under this assumption, the absolute intensity waveform might change, but the relative shape of the waveform will not. In the strategy of the optimization, when the cost function drops down below a certain value, we put more weight on the similarity between the theoretical and observed waveform, especially at those peak and valley positions (wavelength). This is very important in order to avoid re-annealing, which is extremely slow.

2. The thickness extraction of oxide will not affect the extraction of other layers.

While tuning the optimization algorithm, we found that at higher wavelength range (520nm ~ 820nm), the variation of photoresist *Na* and *Nb* has little effect on the cost function, the oxide thickness variation also has little effect on the cost function. So at higher wavelength, we extract the photoresist thickness by optimizing 2 variables, photoresist thickness and polysilicon thickness, providing the oxide thickness. At the lower wavelength range (320nm ~ 470nm), the variation of photoresist *Na* and *Nb* has large impact on the cost function. So at this range, we extract the PAC by optimizing 3 variables, photoresist *Na* and *Nb* value and its PAC, providing the thickness values of photoresist, polysilicon and oxide.

FIGURE 3 shows the optimization performance versus both the "accepted run number" and "generated run number". In this example we extract the photoresist thickness by optimizing 3 variables, the thickness of photoresist, poly and oxide. From the figure we can see that to get a good optimization result, we need about 400 runs to be generated and 70 runs to be accepted. These are used as the stopping criterion in this specific problem.

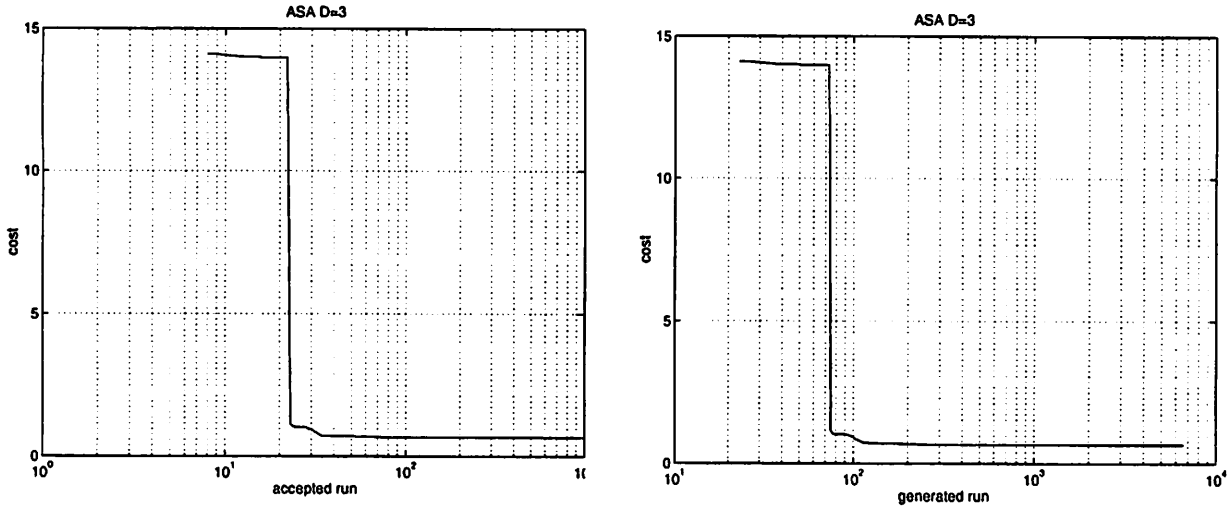


FIGURE 3 The cost of the ASA algorithm in the photoresist thickness extraction.

FIGURE 4 shows the comparison between the observed and theoretical relative intensity after PACbp and THK extraction. The two waveforms are not exactly matched because that the material coefficient are different from the standard values. Since we concentrate on optimizing the similarities of two waveforms, they look similar, especially in the location of peaks and valleys.

As stated before, our goal of the *in-situ* measurement is its accuracy and speed. It takes about 0.25 seconds for data collection and about 25 to 35 seconds for thickness and PAC extraction. This method is very powerful because that it allows us to extract the information of those films under the photoresist simultaneously. This is one of the foundation of the idea of "Custom Photolithographic Process Design", which is to design and control the photoresist thickness depending on the film to be etched. FIGURE 5 is the poly thickness extraction for 2 lots. The 47 wafers were processed in 4 different occasions under the same process recipe. We can see both large variation between lot and within lot in LPCVD process. FIGURE 6 shows the photoresist in two different locations (related to the wafer center) in a lot.

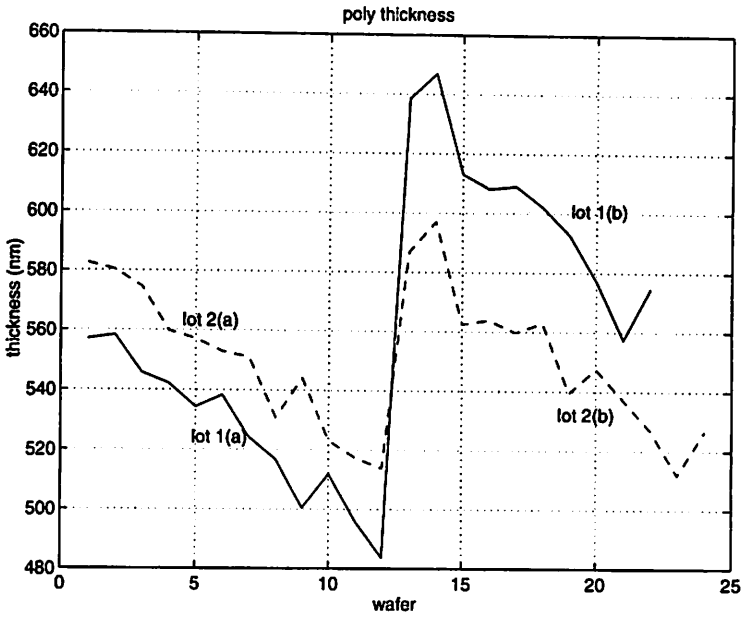


FIGURE 4 Poly thickness extraction from two lots.

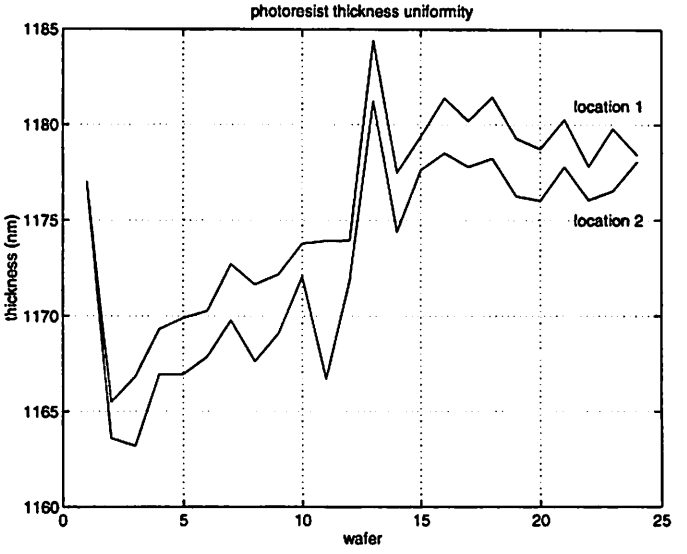


FIGURE 5 Photoresist thickness variation within wafer.

3.3 Experimental Design and Modeling

In order to build effective equipment models for process control, it is desirable to determine a minimal set of input/output parameters to observe and control. For the photoresist coater, the input parameters include spin speed (SPS), bake temperature (BTE) and bake time (BTI). The output parameters include photoresist thickness (PhTh) and photo-active compound concentration (PAC). PAC is further divided into the PAC before exposure (PACbp) and the PAC after exposure (PACxp).

It has been shown that PACbp is highly correlated with PhTh and mainly determined by the BTE [8]. However, changing the bake temperature is time consuming. For example, SVG8626 will take about 5 minutes if to decrease the temperature by 5 °C. From the viewpoint of control, it would be better to avoid changing the bake temperature.

Table 1 is the designed experiment which is conducted in two different days according to different block. The experiment consists of two 2^3 experiment design, each with 3 center points. We measured 3 random locations close to the center of the wafer. The averages of each run are used to build the model. Since on each wafer, the locations of three measurement are close, we do not expect to see large variation among them. From the extracted data we can see that the metrology is self-consistent. Equation (14) is the linear regression model with an R^2 of 0.9977. FIGURE 7 is the model fitting and residual plot. From the residual plot we can see the residuals are normally distributed. The high value of R^2 and normally distributed residuals indicate that the model is adequate and accurate.

Table 2 Experimental design for coater

Run	Block	SPS (rpm)	BTE (°C)	BTI (sec)	PhTh (nm)	PhTh (nm)	PhTh (nm)
1	1	4600	60	90	1167.196	1164.986	1168.637
2	1	3600	90	85	1305.012	1312.360	1300.286
3	1	5600	30	85	1057.972	1059.117	1056.263
4	1	3600	30	95	1336.918	1332.783	1329.071
5	1	3600	90	95	1315.420	1313.697	1311.961
6	1	5600	30	95	1061.549	1055.810	1060.625
7	1	3600	30	85	1333.289	1328.629	1332.508
8	1	5600	90	95	1051.291	1050.889	1049.330
9	1	4600	60	90	1161.115	1166.964	1160.433
10	1	5600	90	85	1053.542	1053.647	1053.147
11	1	4600	60	90	1181.660	1177.723	1177.772
12	2	5600	30	85	1064.716	1051.140	1061.883
13	2	5600	90	95	1044.846	1042.245	1044.119
14	2	3600	30	95	1335.203	1331.890	1332.722
15	2	5600	30	95	1064.982	1064.182	1063.228
16	2	5600	90	85	1054.228	1050.301	1052.753
17	2	4600	60	90	1176.359	1173.438	1174.665
18	2	3600	90	95	1309.892	1306.741	1309.767
19	2	3600	90	85	1318.893	1317.619	1319.452
20	2	4600	60	90	1177.319	1175.234	1174.574
21	2	4600	60	90	1180.061	1176.274	1178.223
22	2	3600	30	85	1335.173	1333.280	1330.116

$$\text{PhTh} = 0.284146 + \frac{80691.261}{\sqrt{\text{SPS}}} + (-0.2580632)\,\text{BTI} + (-0.06735417)\,\text{BTE} \quad . (14)$$

Strictly speaking, the equipment model is only accurate in a certain time period because of "machine aging". FIGURE 8 shows the drift of the equipment model. "*" and "+" are the photoresist thickness measurement in two-week time interval. We can see a clear process drift. Considering the lifetime of an equipment model, a simple model modification is proposed as following:

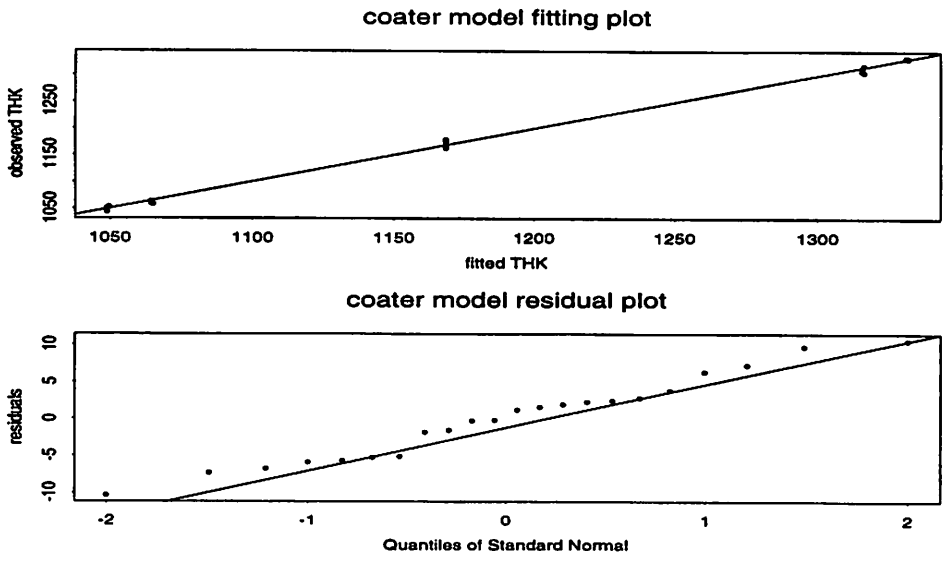


FIGURE 6 Coater model fitting and residual plot.

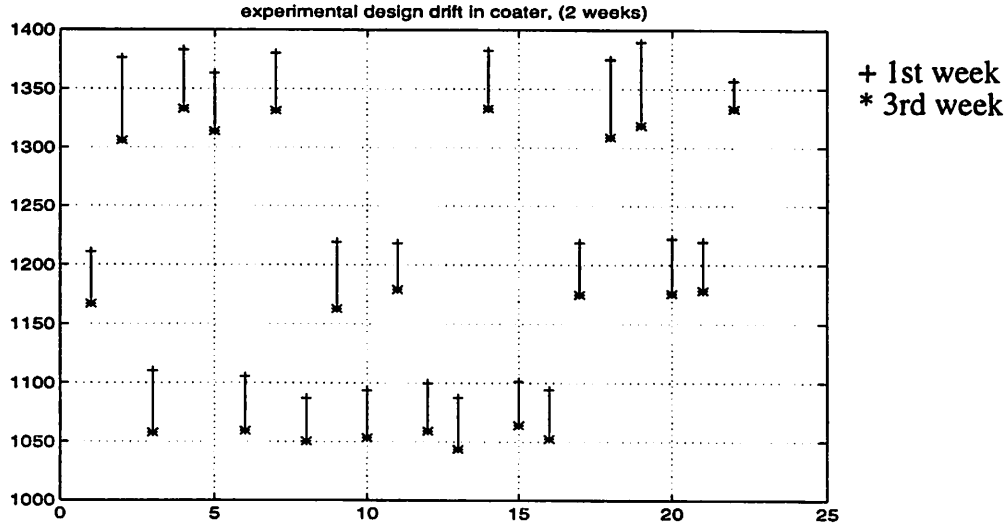


FIGURE 7 The drift of the equipment model.

Assume that we have built two models in the near past, f_1 and f_2, then the modified model is

$$f = \lambda f_1 + (1-\lambda) f_2$$
$$0 \le \lambda \le 1 \tag{15}$$

where λ is determined by several new baseline runs.

3.4 EWMA control chart

We apply EWMA control chart to a baseline run, which includes 24 wafers. FIGURE 9 represents the plot of the sum of squared one-step-ahead prediction errors for the EWMA as a function of λ. The minimum squared prediction error occurs at $\lambda = 0.6$. With this λ value, FIGURE 7 (a) shows the EWMA center line control chart, FIGURE 7 (b) shows the one-step-ahead prediction error. From both control charts, we can see two points out of control. The first alarm can be explained as lack of dummy wafer beforehand. The second alarm is a process fault. We can see the clear process drift.

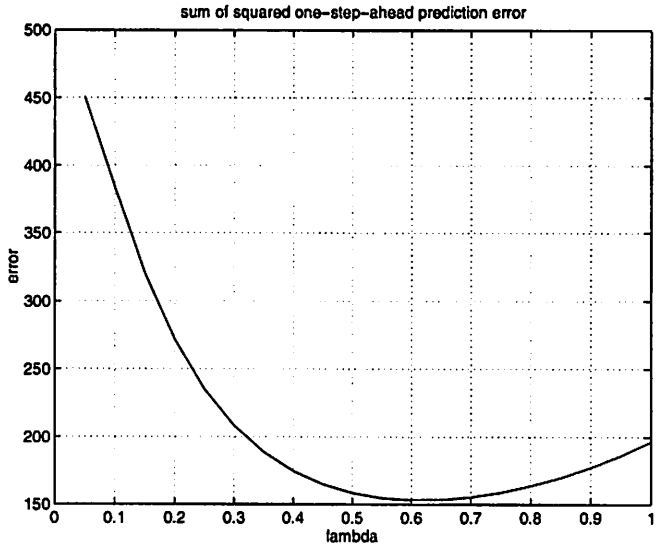


FIGURE 8 Sum of squared one-step-ahead prediction errors versus λ.

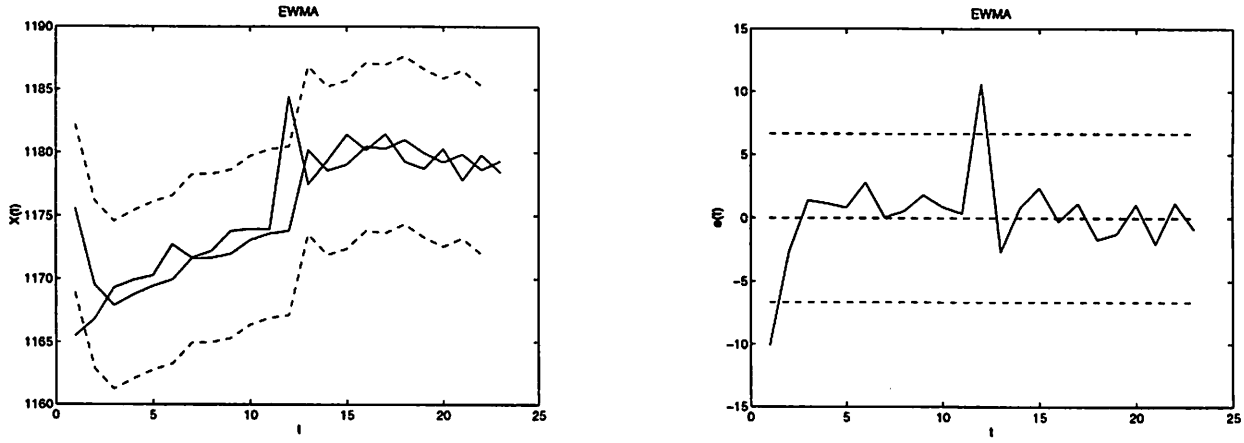


FIGURE 9 EWMA charts on the baseline process.

Based on the calculated λ and σ_p, we monitor next 13 wafers under same process recipe. FIGURE 11 are the corresponding EWMA charts, all of them are in process control.

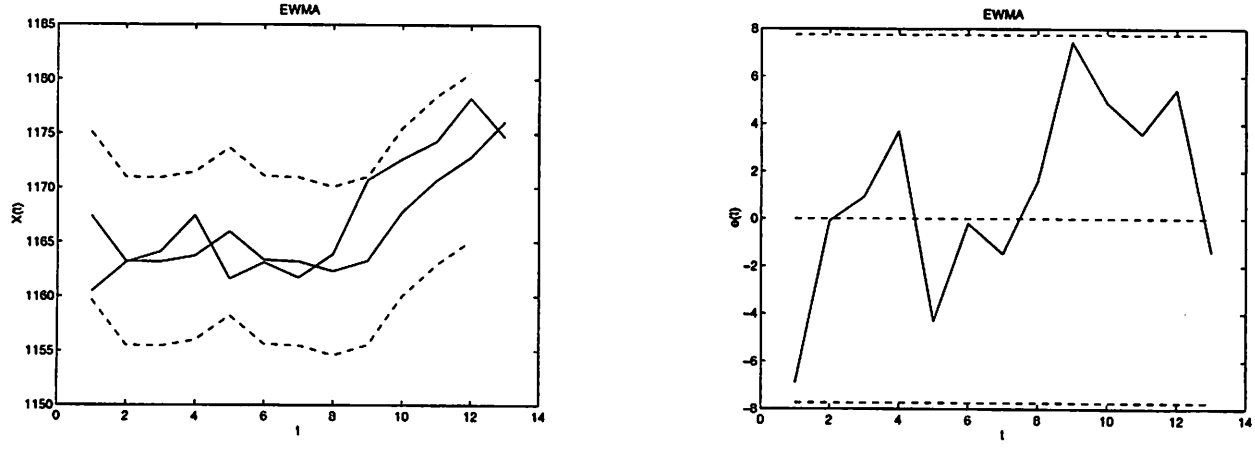


FIGURE 10 EWMA chart.

4.0 Conclusions

In this work we first developed a powerful *in-situ* film metrology. Then we build a model for the coater based on the experimental design. Accurate equipment model is for future run-to-run process recipe design. Because of the high process throughput, we also implement an EWMA control chart to monitor a certain process for the autocorrelated measurement.

Acknowledgments:

I would like to thank Prof. Lester Ingber for his help in using ASA. I thank Prof. Spanos for feedback from several drafts. I thank Tony Miranda and Roawen Chen for their help in the wafer preparation. I thank Anna Ison for her proof-reading of this report. I thank Maria Perez for her technical support on various equipment and process in the Berkeley Microfabrication Laboratory.

References

[1] Max Born and Emil Wolf, "Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light", 6th Edition, Pergamon Press, 1980.

[2] G. E. P. Box, W. G. Hunter, and J. S. Hunter, "Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building", John Wiley and Sons, Inc., 1978.

[3] G. E. P. Box and T. Kramer, "Statistical Process Monitoring and Feedback Adjustment --- A Discussion (with discussions)", Technometrics, Vol. 34, No. 3, 251-285, Aug. 1992.

[4] G. E. P. Box and G. M. Jenkins, "Time Series Analysis: Forecasting and Control", San Francisco: Holden-Day, 1976.

[5] Lester Ingber, "Simulated Annealing: Practice versus Theory", Math. Comput. Modeling, 18, 11, 29-57, 1993.

[6] Lester Ingber, "Adaptive Simulated Annealing (ASA) v7.8", ftp://alumni.caltech.edu, /pub/ingber/, Lester Ingber Research, 1995.

[7] Sovarong Leang and C. J. Spanos, "A Novel In-line Automated Metrology For Photolithography", Private Communications, University of California at Berkeley, Nov, 1994.

[8] Shang-Yi Ma, "The Application of Dynamic Specifications in a Multistep Lithographic Sequence", Memorandum No. UCB/ERL M94/84, Oct. 1994.

[9] D. C. Montgomery, "Introduction to Statistical Quality Control", 2nd ed., New York, Wiley, 1991.

[10] D. C. Montgomery and C. M. Mastrangelo, "Some Statistical Process Control Methods for Autocorrelated Data", Journal of Quality Technology, Vol. 23, No. 3, p179-193, July, 1991.

[11] Andrew Neureuther et al, "SAMPLE User Guide v1.8", University of California at Berkeley, Memorandum No. UCB/ERL M91, 1991.

[12] Andrew Neureuther, "Simulation of Semiconductor Lithography and Topography", Springer-Verlag, (to be published) 1995.

[13] Edward D. Palik, "Handbook of Optical Constants of Solids", Vol I & II, Academic Press, Maryland, 1991.

[14] Bruce Rosen, "Function Optimization Based on Advanced Simulated Annealing", technical report, Division of Mathematics, Computer Science and Statistics.

[15] G. A. F. Seber, "Linear Regression Analysis", John Wiley & Sons, 1977.

[16] S. E. Stokowski, "Measuring Refractive Indices of Films on Semiconductors by Micro-Reflectometry", SPIE, Vol. 1261, p253-263, 1990.

[17] SC Technology, "INS-800-1 Instruction Manual", 1991.

Application of the Robust Design Method for IC Design Improvement

Jone Chen

A computer based experiment for IC design improvement is studied. By applying the Robust Design Method, the optimal and robust IC design for manufacturability can be quickly achieved. This is done with the assumption of an additive model of factor effects, and the use of an orthogonal array to define the matrix experiment and to maximize the quality metric (signal to noise ratio) of the performance criterion. Based on such a technique, a circuit designer can quickly get feedback on circuit performance and its sensitivity to manufacturing imperfections.

1.0 Introduction

Due to process variations in IC manufacturing, parameter variations may cause product to either degrade in performance or even perform outside the given specifications. Therefore, to reduce the effect of parameter variations on product performance is required. For this purpose, the Robust Design Method (Taguchi's-based method) [1] is a very useful design philosophy.

The Robust Design Method is aimed at optimizing product performance, manufacturability, and cost by varying certain decision variables in order to make the product less sensitive to manufacturing imperfections. It uses an orthogonal array to explore the decision variables with a small number of experiments, and to maximize the signal to noise ratio of the performance criterion and thus find the optimal variable settings. Though this technique is simple and powerful, the assumption of the additive model for factor effects need to be confirmed by using confirmation runs.

In this report, the application of the Robust Design Method to improve the design of a VLSI circuit building block, an adder, is illustrated. The experiment was conducted by using the HSPICE [2] circuit simulator. After the analysis of the experiment, the error and the assumption of the additive model for factor effects are investigated. The conclusion of the optimal parameter settings is also discussed.

2.0 Methodology

In order to improve the quality of a product by minimizing the effect of the process variations without eliminating the causes, a set of design parameters that cause the smallest deviation of the quality characteristic from its desired target is required. In traditional work, such an optimal set of design parameters was found by trial and error. For each parameter, the effect on the products performance has to be examined alone. When the number of control variables increase, this approach can be very costly and time consuming.

The Robust Design Method performs only a smaller number of experiments and makes conclusions based on the available results. This is done by an orthogonal array that defines the matrix

experiment. Because any two columns of an orthogonal matrix are mutually orthogonal, combinations of factor levels occur an equal number of times. Such a technique allows the effects of all variables to be determined efficiently.

An important assumption in Robust Design Method is the additive model for factor effects of variables. This means that no interaction terms of factor effects exist. One might doubt this rather optimistic assumption. Fortunately, in the real case, the interaction terms usually can be negligible in comparison with the main effects. Besides, the definition of the signal to noise ratio (as will be stated later) as a logarithm function can further improve this drawback. However, after the analysis of factor effects, this additive model still has to be validated by confirmation runs.

The way to find the optimal setting of each variable under a specific performance criterion is achieved by maximizing the quality metric called signal to noise ratio (SN). The definition of the signal to noise ratio, for the speed of an adder, is the following:

$SN_{speed} = -10 \log_{10}(\text{delay})$ (dB)

The factor effect of variable A set at level 1 is calculated as the following:

$$m_{A1} = \overline{SN_{A1}} - \overline{SN} = \sum_{i=1}^{m} SN_{A1} - \sum_{i=1}^{n} SN$$

where $\overline{SN}$ is the mean of all n experiments and $\overline{SN}_{A1}$ is the mean of m experiments where variable A is set to level 1. By plotting the factor effects for all decision variables, the optimal setting of variables under a certain performance criterion can be determined.

3.0 Implementation

An adder was used to illustrate the application of Robust Design Method for IC design improvement. The decision variables are oxide thickness (T_{ox}), the width of the carry input buffer transistors (W_{in}), and the width and length of the carry output buffer transistors (W_{out}, L_{out}). The levels of each factor are listed in Tab. 1.

Factor	level1	lever2	lever3
T_{ox} (nm)	15	20	25
L_{out} (μm)	1.0	1.2	1.4
W_{in} (μm): NMOS, PMOS	4, 8	6, 12	8, 16
W_{out}(μm): NMOS, PMOS	4, 8	6, 12	8, 16

TABLE 1. Definition of decision variables and their factor levels

The performance criteria to be optimized are speed (delay from the carry input to carry output), speed degradation due to hot-carrier effects, area (the sum of the product of channel length and channel width of each transistor), sensitivity of speed to variations of the channel length, and the sensitivity of power to variations of the channel length.

The speed degradation due to hot-carrier effects was estimated by two HSPICE simulation runs. During the first HSPICE run, the hot-carrier induced damage (quantified as a Age parameter [3]) of each transistor under a specified circuit operation time was calculated from the waveform of gate and drain voltages. The second HSPICE run was used to simulate the degraded circuit behavior by adding a voltage controlled current source in parallel with each degraded transistor as shown in Fig. 1. This voltage controlled current source is used to model the hot-carrier induced degradation in drain current and has been found to be a function of the Age parameter, effective channel length L, gate to source voltage Vgs, and drain to source voltage Vds [4]

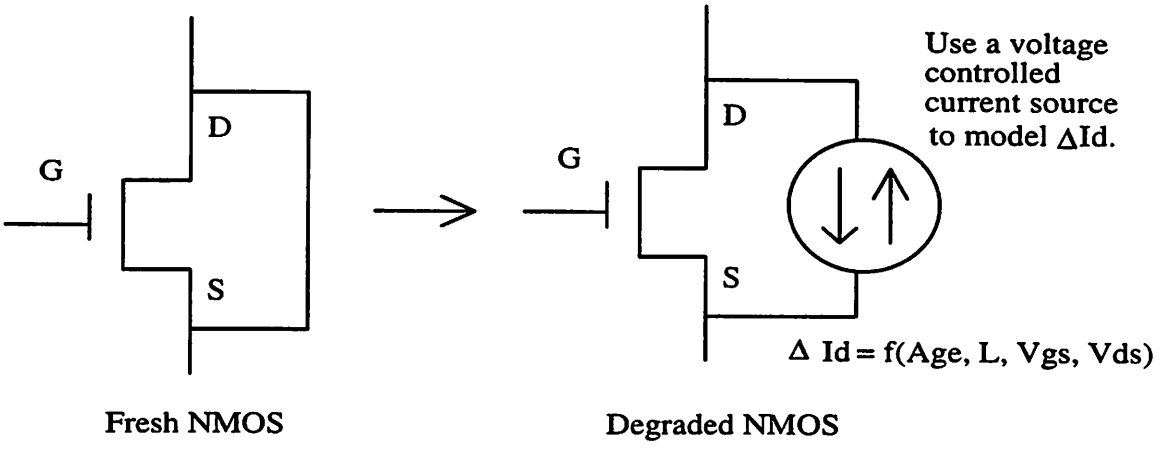


FIGURE 1. Hot-carrier effect can be modeled by a voltage controlled current source

The definition of the signal to noise ratio of each performance criterion is the following:

$SN1 = SN_{speed} = -10 \log_{10}(delay_0)$ (dB)

$$SN2 = SN_{deg(speed)} = -10 \log_{10}\left(\frac{delay_{deg} - delay_0}{delay_0}\right) \text{ (dB)}$$

$SN3 = SN_{area} = -10 \log_{10}(area)$ (dB)

$SN4 = SN_{power} = -10 \log_{10}(power)$ (dB)

$SN5 = SN_{sen\_of\_speed\ to\ \Delta L} = -10 \log_{10}(sen\_of\_speed\ to\ \Delta L * delay_0)$ (dB)

$SN6 = SN_{sen\_of\_power\ to\ \Delta L} = -10 \log_{10}(sen\_of\_power\ to\ \Delta L * power)$ (dB)

where $delay_0$ is the delay of a fresh adder, $delay_{deg}$ is the delay of a degraded adder, and

$$sen\_of\_speed\ to\ \Delta L = \frac{delay_{L+\Delta L} - delay_L}{\Delta L}$$

$$sen\_of\_power\ to\ \Delta L = \frac{Power_{L+\Delta L} - Power_L}{\Delta L}$$

where L=1.2μm, ΔL=0.05μm. $delay_{L+\Delta L}$ and $power_{L+\Delta L}$ are the delay and power when the channel length of each transistor increases by 0.05μm. The sensitivity of speed and power to the channel length is important when considering the process variation causing the channel length of all transistors to shift out of their designed target values.

4.0 Results and Discussion

According to the orthogonal array, the matrix experiment and the results of each quality metric (in dB) are shown in Tab. 2. The factor effect of each decision variable has been calculated and shown in Tab. 3 and in Fig. 2

	Level				Result					
Exp. #	T_{ox}	L_{out}	W_{in}	W_{out}	**SN1**	**SN2**	**SN3**	**SN4**	**SN5**	**SN6**
1	1	1	1	1	85.63	13.35	94.41	26.07	121.70	9.75
2	1	2	2	2	86.05	14.43	94.10	25.84	122.47	9.28
3	1	3	3	3	86.20	15.34	93.77	25.59	122.85	8.63
4	2	1	2	3	85.84	26.03	94.05	25.98	122.09	10.05
5	2	2	3	1	85.03	23.83	94.18	26.31	121.58	15.60
6	2	3	1	2	85.02	23.68	94.10	26.14	120.93	10.28
7	3	1	3	2	85.12	37.89	94.10	26.30	120.96	15.89
8	3	2	1	3	84.90	35.10	94.02	26.17	120.73	13.85
9	3	3	2	1	84.28	25.12	94.21	26.71	121.04	17.63
10	1	1	1	3	86.24	14.39	94.13	25.71	122.39	10.81
11	1	2	2	1	85.60	14.55	94.27	26.03	122.46	10.85
12	1	3	3	2	85.97	15.26	93.94	25.77	122.60	10.33
13	2	1	2	2	85.59	25.26	94.18	26.14	121.58	9.37
14	2	2	3	3	85.74	35.57	93.87	25.93	122.31	9.31
15	2	3	1	1	84.61	23.76	94.29	26.41	120.68	13.27
16	3	1	3	1	84.67	35.74	94.24	26.54	120.75	18.76
17	3	2	1	2	84.65	30.77	94.18	26.36	120.43	14.20
18	3	3	2	3	84.91	24.22	93.84	26.15	120.84	11.26
Average					85.34	24.13	94.11	26.12	121.58	12.17
Error(exp)					0.240	2.153	0.114	0.143	0.239	1.785
Confirm										
c1	2	2	2	2	85.39	24.33	94.10	26.13	121.42	9.60
predicted					85.18	26.29	94.20	26.25	121.58	11.44
error					0.206	-1.96	-0.09	-0.13	-0.167	-1.83
c2	1	1	3	3	86.59	17.66	93.97	25.65	123.03	10.13
predicted					86.44	19.52	94.01	25.67	122.81	10.02
error					0.150	-1.86	-0.04	-0.02	0.218	0.116

TABLE 2. Experimental matrix and the experimental results

Factor	level	Effect					
		SN1	SN2	SN3	SN4	SN5	SN6
T_{ox}	1	0.613	-9.573	-0.003	-0.285	0.835	-2.232
	2	-0.032	2.227	0.008	0.033	-0.050	-0.859
	3	-0.582	7.346	-0.005	0.252	-0.785	3.090
L_{out}	1	0.178	1.316	0.080	0.003	0.002	0.263
	2	-0.007	1.582	-0.001	-0.011	0.087	0.008
	3	-0.171	-2.898	-0.079	0.008	-0.089	-0.271
W_{in}	1	-0.160	-0.621	0.085	0.024	-0.434	-0.147
	2	0.041	-2.525	0.004	0.022	0.170	-0.766
	3	0.119	3.146	-0.088	-0.046	0.264	0.913
W_{out}	1	-0.365	-1.403	0.161	0.224	-0.209	2.136
	2	0.056	0.422	-0.001	-0.027	-0.082	-0.615
	3	0.309	0.981	-0.159	-0.197	0.292	-1.522
Error(eff)		0.098	0.879	0.047	0.059	0.098	0.729

TABLE 3. Factor effect of each decision variable for each performance criterion

Before drawing any conclusions about the experimental results, the errors associated with the experiment needs to be evaluated. There are two kinds of errors that have to be considered. First, the experimental error (σ_{exp}), which determines whether the model is good or not. The second one is the effect error (σ_{eff}), which determines whether the factor effect is significant or not.

For computer-based experiments, care must be taken concerning the estimation of experimental error. In real-world experiments, two kinds of model error exist. One is the random noise error, which accounts for the error when the identical inputs during different time space create different results. The other is the lack of fit error, which is the error between model prediction and the mean value of experimental results. It is clear that only lack of fit error could exist for computer-based experiments because the same inputs always result in the same output. Therefore, the root-mean-squared (r.m.s.) error [5] of the data was used to estimate the computer-based experimental error (σ_{exp}). The root-mean-squared error is calculated as follows:

$$rms = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

where n is the number of experimental points and y_i and $\hat{y}_i$ are the measured and model prediction values of the output at the ith experiment, respectively. The other concern is that such estimation of experimental error rely on the lack of fit error. However, for small number of experimental runs, the degree of freedom left for error is zero. This causes no lack of fit error and the above

experimental error estimation will fail. In such a case, the number of experiments need to be increased so the lack of fit error can be estimated.

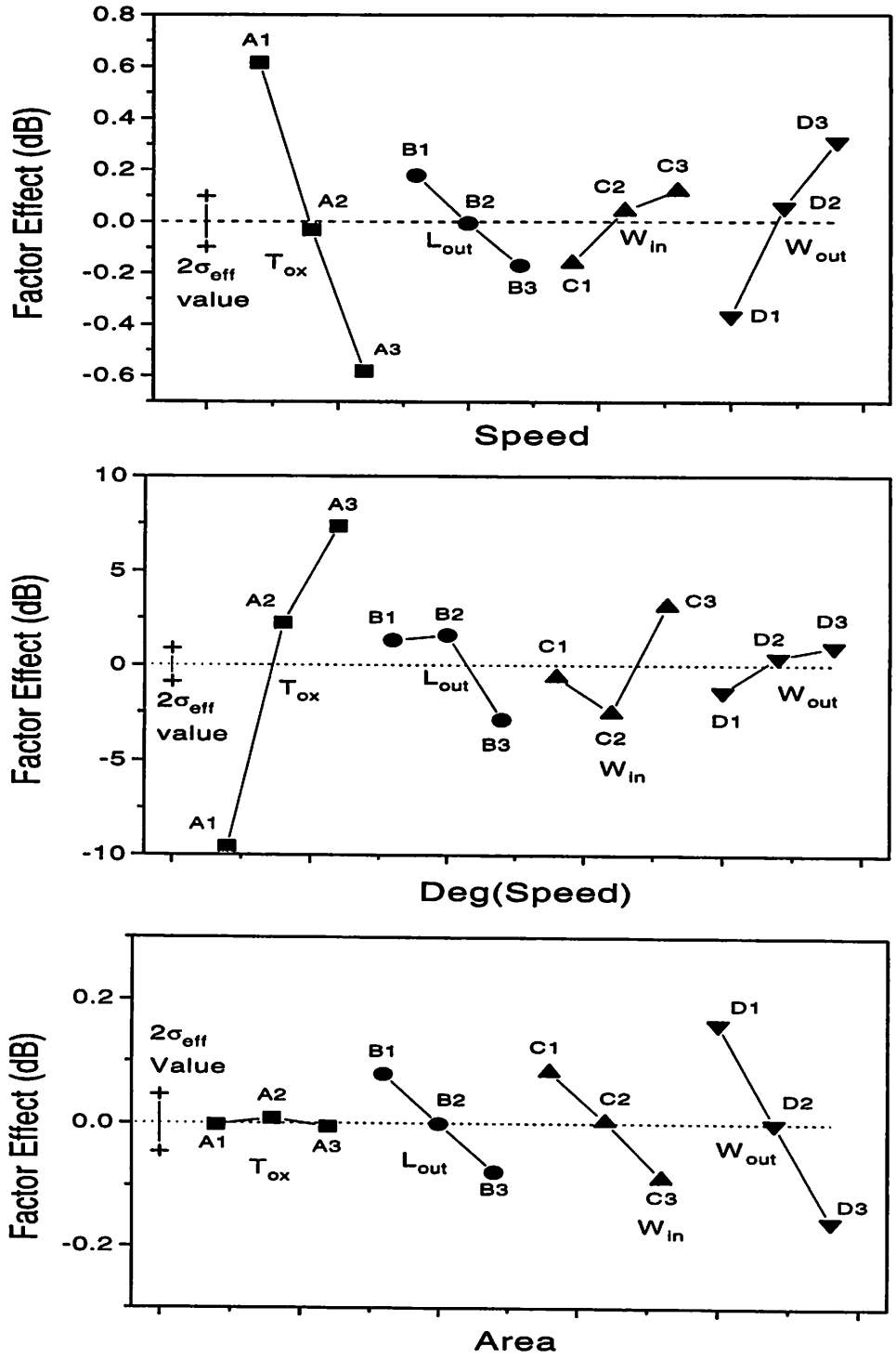


FIGURE 2. Factor-effect plots for each performance criterion

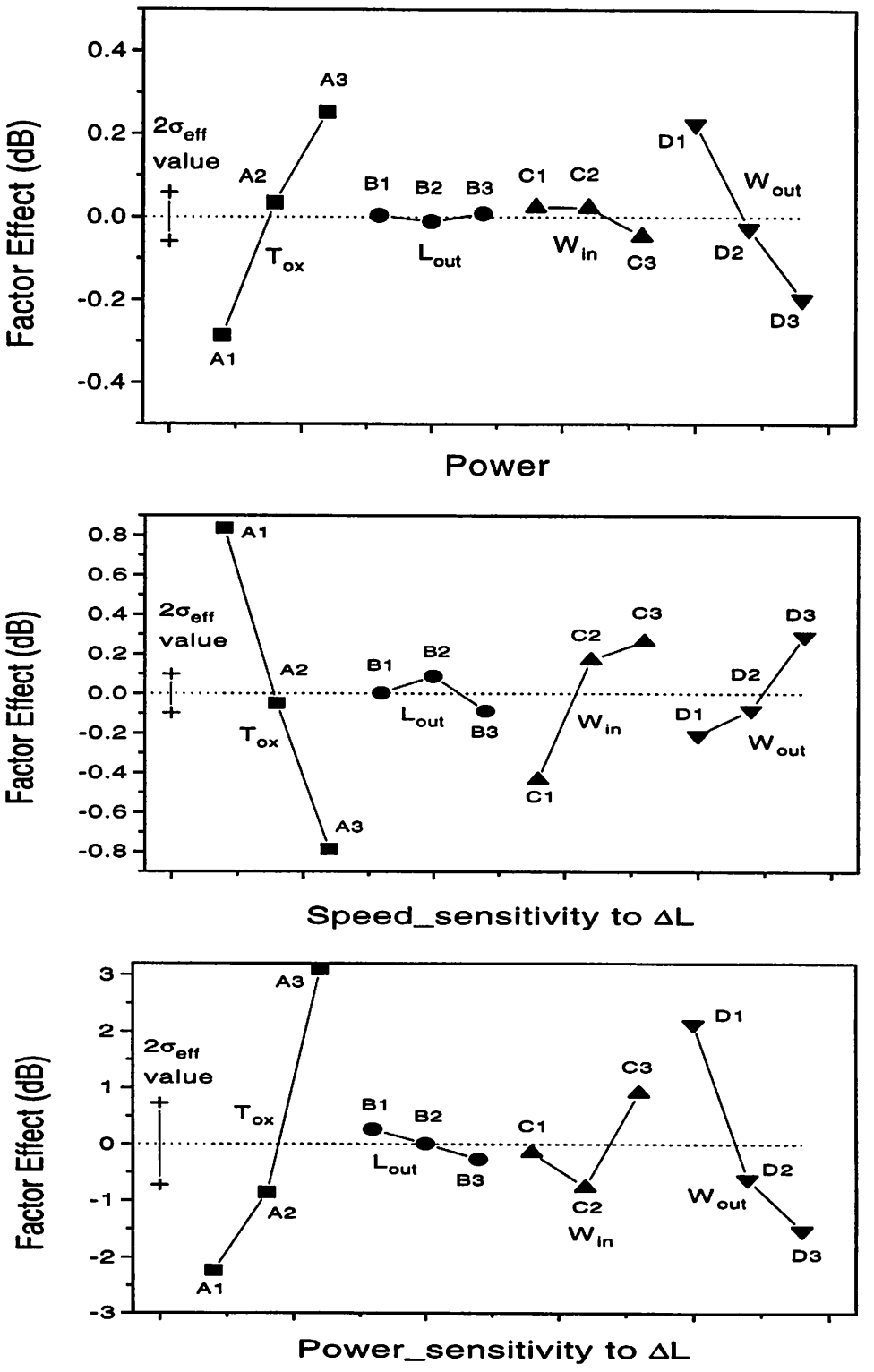


FIGURE 3. Factor-effect plots for each performance criterion

Fig. 2 and 3 show the factor effect of each performance criterion. The decision of a factor setting at level *i* is better than at level *j*, judged by whether their output difference is more than twice the effect error (σ_{eff}) or not. The effect error is calculated from the experimental error as the following:

$$\sigma_{eff}^2 = \frac{1}{m}\sigma_{exp}^2$$

where *m* is the number of experiments that the decision variable is set at level *i* (for the adder example in this report, *m*=6). If the output difference is less than twice σ_{eff}, setting at level i or level *j* is considered to have no significant difference in output. It means that such a conclusion is correct for about 95% of probability.

Tab. 4 shows the optimal parameter setting for each performance criterion.

Factor level	Speed	Deg(Speed)	Area	Power	Sen_of_speed to ΔL	Sen_of_power to ΔL
T_{ox}	1	3	2,1,3	3	1	3
L_{out}	1	2, 1	1	3,1	2,1	1,2,3
W_{in}	3, 2	3	1	1,2,3	3,2	3,1
W_{out}	3	3, 2	1	1	3	1

TABLE 4. The optimal setting of each decision variable for each performance criterion

It is evident that for different performance criteria, the optimal parameter settings are different. Therefore, designers have to pick the parameter setting based on what performance criterion are considered to be important.

Finally, two confirmation runs were performed. First, the typical case of each level setting was used. Second, the setting was picked to aimed at highest speed. In both cases, the lack of fit errors are all less than three times of the estimated experimental error. This indicates that the assumption of additivity of factor effects is valid in this study.

5.0 Conclusions

The fundamentals of the application of the Robust Design Method for IC design improvement are presented in this report. It shows that through the Taguchi's design philosophy, the optimal setting of parameters for product performance and IC manufacturability can be quickly found. The orthogonal array for experiment matrix, the assumption of the additive model of factor effect, the maximizing the signal to noise ratio, and the confirmation runs form the core concept in this design technique. The circuit designer can quickly get feedback from such a methodology on circuit performance and its sensitivity to variation in IC manufacturing.

Acknowledgments

The author would like to thank Prof. C. Spanos for his valuable suggestions and providing the DORIC (Design of Optimized and Robust Integrated Circuits) software.

References

[1] M. S. Phadke, "Quality Engineering using Robust Design," Prentice Hall, AT&T Bell Laboratories, 1989.

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

[3] R. H. Tu, E. Rosenbaum, C. C. Li, W. Y. Chan, P. M. Lee, B.-K. Liew, J. D. Burnett, P. K. Ko, and C. Hu, "BERT - Berkeley Reliability Tools," Memorandom No. UCB/ERL M91/107, University of California, Berkeley, 1991.

[4] W. Y. Chan, "Simulation Model for Hot-Electron-Induced Degradation of CMOS Analog Circuits," M.S. Thesis, University of California, Berkeley, 1994.

[5] Z. Daoud, "DORIC: Design of Optimized and Robust Integrated Circuits," University of California, Berkeley, 1993.

Transistor Matching Properties of the UC Berkeley CMOS Process

Manolis Terrovitis

The subject of this project is the examination of mismatch among transistors manufactured in the UC Berkeley microfabrication line. Although the data is corrupted by processing and measurement noise, basic patterns about the matching behavior can be found with the aid of statistical processing.

1.0 Introduction

Because of IC processing imperfections there are small differences in the behavior of identically designed devices. Two kinds of variation can be considered. Global variation accounts for the total variation of a process parameter over a wafer or a lot. Local variation, or mismatch is the difference in a process parameter between adjacent devices that are designed to be used together in a circuit. Mismatch is critical in Analog Designs whose operation is often based on the ratio rather than the absolute values of the components. Mismatch, in these cases, causes random input offset in amplifiers, and determines the nonlinearity errors of most kinds of A/D converters.

Our study is focused on transistor mismatch. Different drain currents under identical bias conditions are the result of different process parameters, mainly the threshold voltage, the current factor (the product of mobility, thickness of the oxide, and ratio of the effective width over the effective length of the channel), and the substrate factor. We will study measurements of the threshold voltage and we will investigate the factors that affect the threshold voltage mismatch of NMOS and PMOS devices. The results for one wafer will be presented and discussed next. The results of a second wafer will be summarized in the appendix.

2.0 Measurements

In this section we describe the transistors from which the measurements were collected. We will also discuss a manufacturing problem that affects the quality of our results. Finally, we will describe the available equipment and the procedure we followed in taking the measurements in order to achieve certain accuracy.

2.1 Test Structures

Our measurements are collected from four by four arrays of transistors, shown in FIGURE 1. The transistors have a width of 20μm and a length of 2μm. Every four transistors in one row share a common source and every four in a column share a common drain. All 16 share a common gate. There are 52 dies on each 4" wafer, and each die has six arrays of NMOS and six of PMOS transistors.

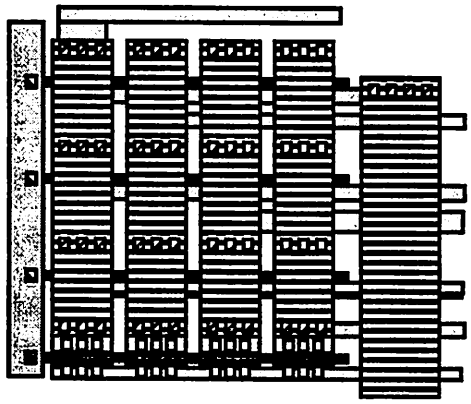


FIGURE 1 Four by four array of transistors.

Metal 2 is used to connect the common drains as well as for the connections from the common sources to the pads as can be seen in FIGURE 1. These wafers have been fabricated in the Berkeley Microfabrication Laboratory where the baseline process is still under development and the vias are not reliable. Some residual oxide on metal1 creates an open-circuit, but we found that applying a high voltage, breaks down the residual oxide and establishes contact. Therefore, in order to enable the experiment, before we took any measurement from each device, we applied 8V on the Gate and the Drain for 2 seconds.

However, even the "enabled" contact has possibly high resistance, is nonlinear and generally has unpredictable behavior. An electric schematic of the array including the via resistances is shown in FIGURE 2. Because of the remaining high via resistance, our measurements are corrupted, and so the conclusions drawn from the subsequent analysis cannot be generalized.

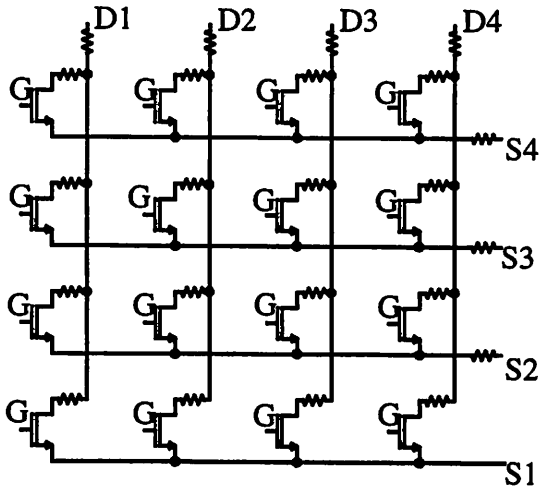


FIGURE 2 Electric schematic of the array including the "enabled" high resistance vias.

2.2 Equipment accuracy

For the measurements we used an automatic wafer tester, consisting of the Electroglas Autoprobe 2001X, the HP4084A Switching Matrix Controller, the HP4085A Switching Matrix and the HP4141A DC Source/Monitor. For the range of voltages that we apply in order to measure threshold voltage, the advertised resolution of the DC Source/Monitor is 1mV, and for the range of currents that we measure the advertised accuracy is 0.17μA.

In order to test the accuracy of this equipment in the threshold voltage measurement, we measured repeatedly the threshold voltage of one device, first without moving the pins from the pads and then by raising and reapplying the pins before each measurement. The scatter plot of the measurement showed only random error for the first 40 points. In the case that the pins do not move from the wafer the standard deviation of the measurements was 0.247mV and in the case that we raise the pins before each measurement, for the set of the 40 first points, 0.484mV.

This means that the in series resistance of the pin-pad contact is a random variable and we should compensate for this also in order to measure in a common way all the devices. We decided to take each measurement 5 times by raising the chuck between successive measurements and consider the averages. This way, the above accuracy is divided by the square root of 5 and our final measurement has a standard deviation of 0.216mV.

3.0 Analysis

In this section, the results of the measurements are analyzed using statistical processing software (Splus) in order to characterize the mismatch behavior. We examine the pattern of the threshold voltage across the array, we look for a pattern of the mismatch across the wafer, we quantify mismatch within the array, the die, and the wafer and, finally, we investigate the effect of distance on mismatch between two transistors in the array.

3.1 Superposition of the arrays

As a first test we superimposed the threshold voltage of all the arrays across the wafer and the results for the NMOS case are shown in FIGURE 3. The sample size is 52*6=312 arrays. We observe that the average of the threshold voltage of the transistors across the one diagonal is higher. In order to check if this effect is significant we plot the standard deviation of the measured threshold voltage across the array also. The result which is shown in FIGURE 3, shows that this pattern is significant. The corresponding plots for the PMOS transistors are shown in FIGURE 4. We observe the same phenomenon although not as intense as in the NMOS case.

There is not straightforward explanation for this pattern. In order to look into this effect and acquire some insight from a specific example, we plot the measured threshold voltage across two sample NMOS arrays in FIGURE 5. From this plot we observe that we have better matching across one dimension. This probably happens because the transistors share common sources and common drains. For example high resistance in series with a common drain modifies in a similar way the measurement of all four transistors that share this drain as can be seen in FIGURE 2. This does not explain the pattern observed in FIGURE 3 and FIGURE 4, which may be due to the unpredictable behavior of the via resistances.

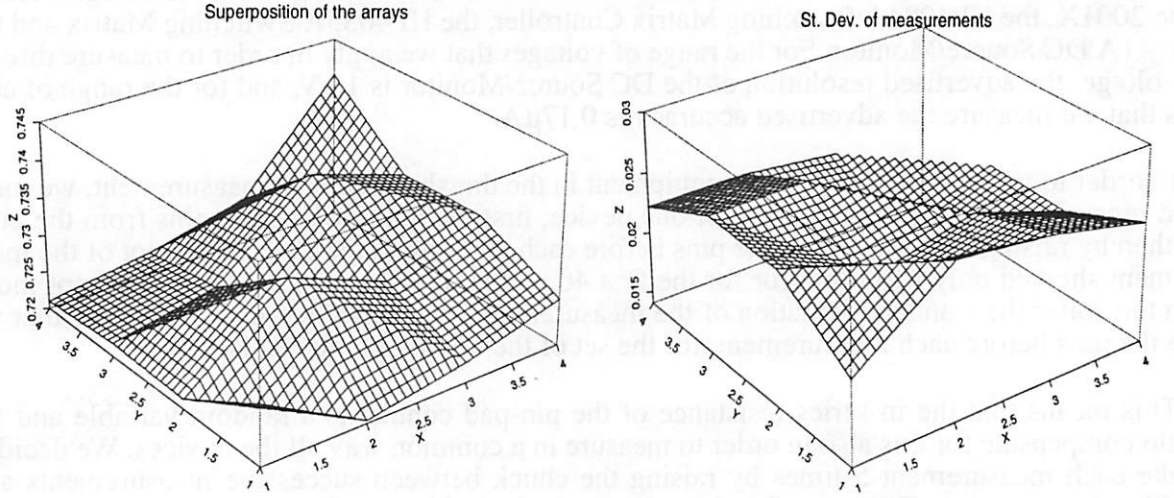


FIGURE 3 Average and Standard Deviation across 312 NMOS arrays.

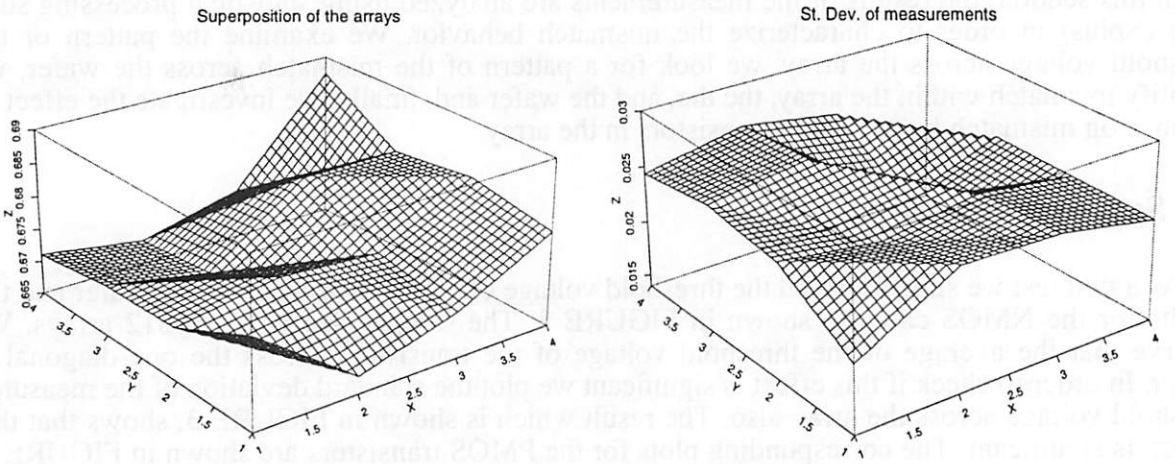


FIGURE 4 Average and Standard Deviation across 312 PMOS arrays.

3.2 Wafer plot of mismatch

In order to examine mismatch across the wafer we need a metric of local mismatch for the whole array. We decided to use the variance of the measured threshold voltages. The resulting maps are shown in FIGURE 6 for the NMOS and the PMOS cases.

From this figure we suspect that for the NMOS case matching is better in the center of the wafer than in the periphery. In order to check this hypothesis we plot the scatter plot of the value of mismatch with respect to the distance from the center in FIGURE 7.

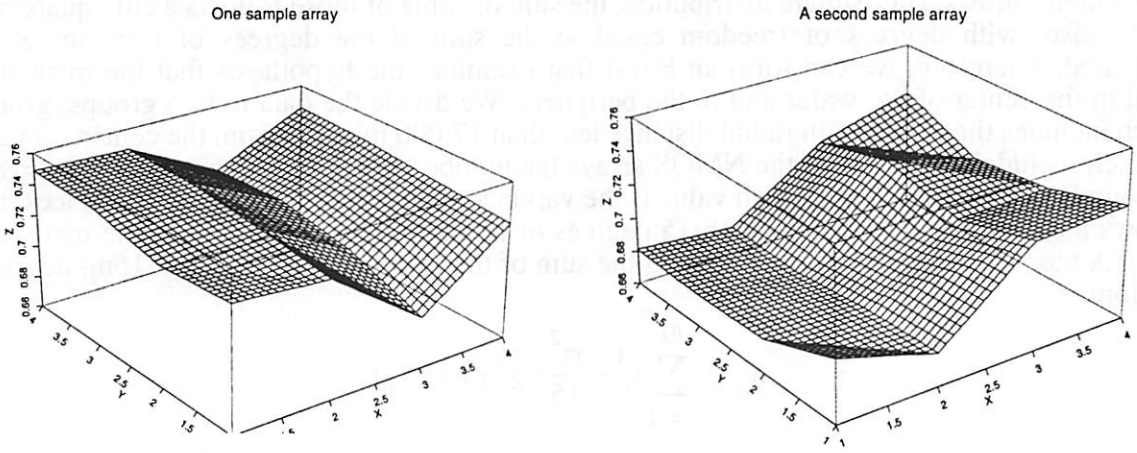


FIGURE 5 Threshold voltage in two sample NMOS arrays.

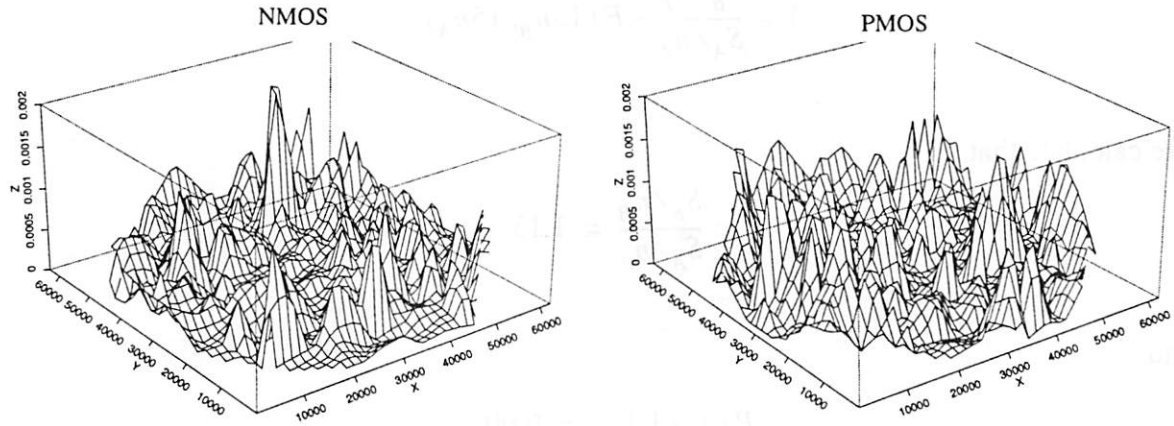


FIGURE 6 Perspective map of array V_t mismatch across the wafer for NMOS (left) and PMOS (right) transistors.

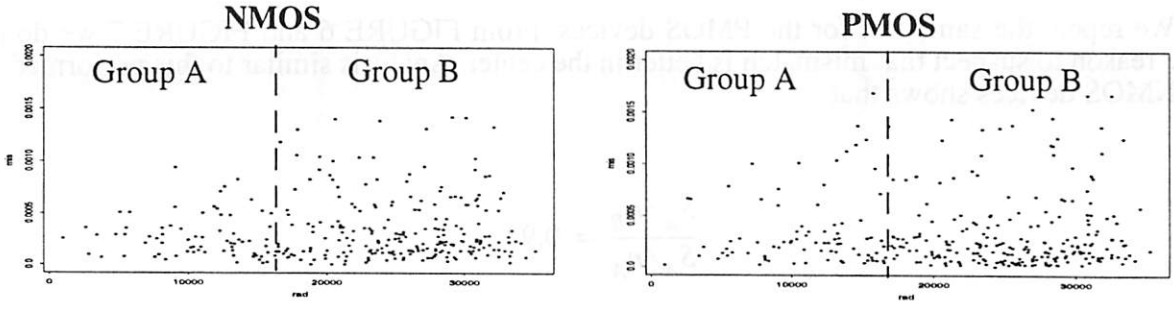


FIGURE 7 Scatter plot of array V_t mismatch with respect to radial distance for NMOS (left) and PMOS (right) transistors.

It is hard to draw a conclusion from this graph. Since each one of the points represents variance which follows a chi-square distribution, the sum of some of those follows a chi-square distribution also, with degrees of freedom equal to the sum of the degrees of freedom of each individual. Therefore, we can form an F-test that examines the hypothesis that the mismatch is equal in the center of the wafer and in the periphery. We divide the data in two groups, group A, which includes the arrays with radial distance less than 17,000 microns from the center, and group B which includes the rest. For the NMOS arrays the number of samples in group A is n_A=78 and in group B n_B=221. Each mismatch value is the variance of threshold voltage of 16 devices, and it follows a chi-square distribution with 15 degrees of freedom. Hence the sum of the members of group A has $15n_A$ degrees of freedom and the sum of the members of group has $15n_B$ degrees of freedom.

$$S_A = \sum_{i=1}^{n_A} s_i^A \sim \frac{\sigma^2}{15} \cdot \chi^2(15 \cdot n_A)$$

$$S_B = \sum_{i=1}^{n_B} s_i^B \sim \frac{\sigma^2}{15} \cdot \chi^2(15 \cdot n_B)$$

$$Y = \frac{S_B/n_B}{S_A/n_A} \sim F(15n_B, 15n_A)$$

We calculate that

$$\frac{S_B/n_B}{S_A/n_A} = 1.13$$

and

$$P(Y > 1.13) = 0.005$$

The F-test shows that the null hypothesis of equality of the mismatch can be rejected against the alternative hypothesis of lower mismatch in the center with probability of type I error 0.5%. Therefore there is evidence that the matching is better in the center, but not by much since the ratio of the average of the mismatch is only 1.13.

We repeat the same test for the PMOS devices. From FIGURE 6 and FIGURE 7 we do not have reason to suspect that mismatch is better in the center. Analysis similar to this performed for the NMOS devices shows that

$$\frac{S_B/n_B}{S_A/n_A} = 0.92$$

where n_A=83 and n_B=228. Testing the null hypothesis that the mismatch is equal in the two groups against the alternative that it is higher in the center (the opposite than for NMOS) shows that the null hypothesis can be rejected with probability of type I error 2.96%. Therefore here

there is evidence that the mismatch is higher in the center, but not by much since the ratio of the average mismatch in the two groups is only 0.92.

In addition, in order to check if there are any vertical or horizontal mismatch patterns, we calculated horizontal and vertical mismatch in each array. For the horizontal mismatch, this was done by calculating the standard deviation of the threshold voltage of the transistors in each of the four rows and averaging them. This was done because four values of mismatch in approximately the same point cannot be depicted on the wafer map. The vertical mismatch was calculated in a similar way by calculating the standard deviation in each column and averaging the four columns. The results are shown in FIGURE 8.

Horizontal and vertical mismatch have different shape, but not specific pattern. There are

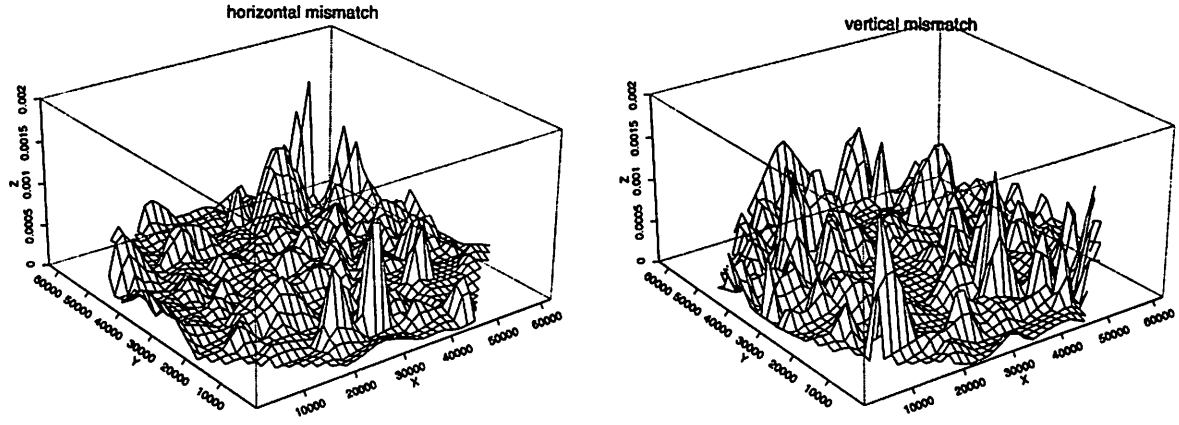


FIGURE 8 Horizontal and vertical mismatch across the wafer for the NMOS devices.

areas where the vertical mismatch dominates and others where the horizontal dominates. The average of the overall mismatch is 3.4e-4 V^2, the horizontal is 2.1e-4 V^2 and the vertical is 2.6e-4 V^2.

The corresponding wafer perspective maps for the PMOS devices are shown in FIGURE 9.

These graphs lead to the same conclusions as for the NMOS devices. The average of the overall mismatch is 3.4e-4 V^2, the horizontal is 1.9e-4 V^2, and the vertical is 2.7e-4 V^2.

3.3 Within array, die, and wafer variability

The purpose of the next step was to calculate the average variability within the array, within the die, and within the wafer. The first is just the average of the mismatch of the arrays and has been calculated above. The variability across the die can be calculated by averaging the threshold voltage in each array, then taking the variance over each die and finally averaging the variances of all the dies. The variation across the wafer (between dies) can be calculated by averaging the

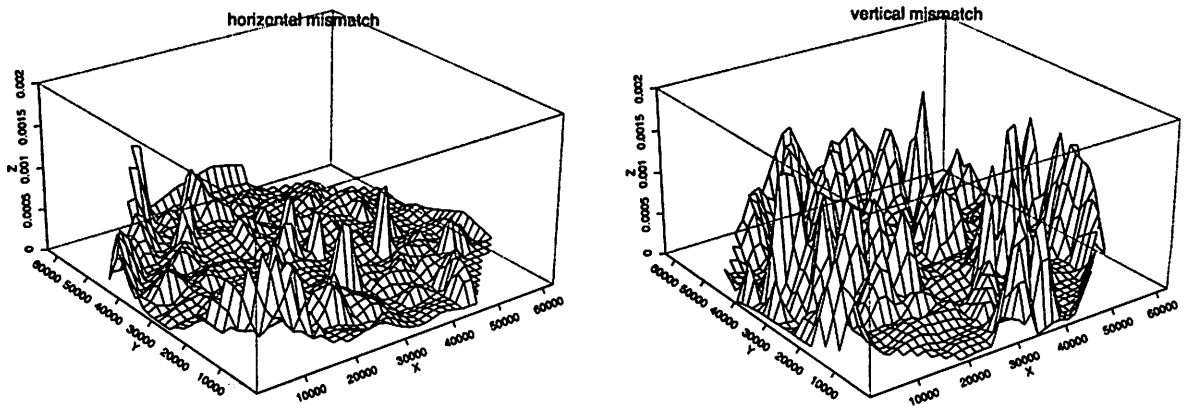


FIGURE 9 Horizontal and vertical mismatch across the wafer for the PMOS devices.

threshold voltage across each die and then taking the variance for all dies. The square root of the above variations for the NMOS transistors are depicted in Table 1.

Table 1: Within array, die and wafer mismatch for NMOS devices

Source of mismatch	Average Mismatch
within array	18.4 mV
within die (between array averages)	12.0 mV
within wafer (between die averages)	9.5 mV
between trans. in different arrays of the same die	22.0 mV
between trans. in different dies	23.9 mV

The mismatch between individual transistors in different arrays of the same die can be considered as the sum of two random processes: one that gives the mismatch between the means of the two arrays and one that gives the mismatch within one array. Therefore the variance is given as the sum of the variances of the two processes. The mismatch between individual transistors in different dies can be calculated in a similar way.

The above results agree with our intuition. Averaging the threshold voltage across each array in order to calculate the within die variability smooths the local variations. This is the reason that large devices (such as the array average) exhibit exist smaller mismatch compared to that of small devices. It is well known that the mismatch (expressed as the standard deviation of a parameter) is inversely proportional to the square root of the area of the devices.

We can form a statistical test to verify the (obvious) inequality of mismatch within the array, within the die and within the wafer. The test is depicted in table 2.

Table 2: Testing of equality of mismatch from different sources for NMOS devices

Source of variation	mean square ($Volt^2$)	degrees of freedom	ratio
within array	s_a=0.0003403383	52*6*(16-1)	s_a/s_d=2.350094
within die	s_d=0.000144819	52*(6-1)	s_d/s_w=1.597082
within wafer	s_w=9.067722e-5	(52-1)	

The test of equality of mismatch within the array and within the die gives that we can reject the null hypothesis of equality with probability of error essentially 0. Similar test of equality of mismatch between the within die and the within wafer variation confirms that we can reject the null hypothesis at a level of confidence 2.29%.

The corresponding values for the PMOS devices are shown in Table 3 and Table 4.

Table 3: Within array, die and wafer mismatch for PMOS devices.

Source of mismatch	Average Mismatch
within array	18.4 mV
within die (between arrays)	13.4 mV
within wafer (between dies)	10.9 mV
between trans. in different arrays of the same die	22.8 mV
between trans. in different dies	25.2 mV

Table 4: Testing of equality of mismatch from different sources for PMOS devices.

Source of variation	mean square ($Volts^2$)	degrees of freedom	ratio
within array	sa=0.0003385738	52*6*(16-1)	sa/sd=1.867708
within die	sd=0.0001812777	52*(6-1)	sd/sw=1.531576
within wafer	sw=0.0001183602	(52-1)	

Test of equality of mismatch within the array and within the die gives that we can reject the null hypothesis of equality with probability of error 1.4e-10. The test of equality of mismatch between the within die and the within wafer variation confirms that we can reject the null hypothesis with probability of error 3.42%.

3.4 Dependence of mismatch on distance within the array

From each raw (or column) of transistors in the arrays we can calculate

- 3 differences of threshold voltage between devices with distance 1 (adjacent transistors),
- 2 differences of threshold voltage between devices with distance 2 (there is one transistor in between) and
- 1 difference of threshold voltage between devices with distance 3 (there are two transistors in between).

Horizontal distance of 1 corresponds to 25μm and vertical distance of 1 to 22μm. In this analysis we will use the absolute value of these differences. In order to simplify the analysis we average the differences from each raw (column) that correspond to the same distance. Then we average the differences that correspond to the same value of distance, and we keep three values form each array, one for each value of the distance. We can consider the distance as a "treatment" and perform the analysis of variance in order to check if the mean value of mismatch is different for the three distances.

The data now consists of either absolute values of difference of threshold voltage or average of those and does not follow a Gaussian distribution. We have three sets of data, one for each value of the distance. We can divide each set in smaller groups, for example each consisting of 10 arrays, and average the data, thus obtaining one value from each group. Then the central limit theorem can guarantee the normality of the distribution of the averaged data. Applying the above method to the horizontal mismatch of the NMOS transistors we obtain the following ANOVA table:

Table 5: Mismatch versus horizontal distance for NMOS transistors

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01558989	s1=8.875025e-06	29	
trn with dist. = 2	0.01576097	s2=9.244643e-06	29	
trn with dist. = 3	0.0156616	s3=2.154695e-05	29	
within treatment		s_R=1.322221e-05	87	s_T/s_R=0.01617
between treatment		s_T=2.214264e-07	2	

$P(F(2,87) > s_T/s_R) = 0.98$

Therefore there is serious evidence that the means are equal and that the mismatch between the devices does not depend on distance. This again is a result that we expect. In [2] we find: "The relative effect of the mismatch due to the distance is only significant for large area devices with considerable spacing."

Repeating the same procedure for vertical mismatch of the NMOS transistors we obtain the following results:

Table 6: Mismatch versus vertical distance for NMOS transistors

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01468043	s1=1.124786e-05	29	
trn with dist. = 2	0.01555932	s2=1.00454e-05	29	
trn with dist. = 3	0.01548939	s3=1.510556e-05	29	
within treatment		s_R=1.213294e-05	87	s_T/s_R=0.59002
between treatment		s_T=7.158657e-06	2	

$P(F(2,87) > s_T/s_R) = 0.56$

Therefore we conclude that the mismatch within the arrays does not depend on distance.

The results for the PMOS devices are shown in Table 7 and Table 8, and lead to the same results as for NMOS. The level of confidence at which we can reject the null hypothesis of equality of the means remains high.

Also the fact that the mismatch does not increase monotonically with distance in some of these cases supports the previous conclusion.

Table 7: Mismatch versus horizontal distance for PMOS transistors

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01451119	s1=1.988192e-05	30	
trn with dist. = 2	0.01448647	s2=1.90976e-05	30	
trn with dist. = 3	0.01529769	s3=2.501969e-05	30	
within treatment		s_R=2.133307e-05	90	s_T/s_R=0.30935
between treatment		s_T=6.599281e-06	2	

$P(F(2,87) > s_T/s_R) = 0.7347046$

Table 8: Mismatch versus vertical distance for PMOS transistors

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.0142865	s1=1.053066e-05	29	
trn with dist. = 2	0.0143878	s2=1.010334e-05	29	
trn with dist. = 3	0.01292618	s3=9.30914e-06	29	
within treatment		s_R=9.973412e-06	87	s_T/s_R=2.00385
between treatment		s_T=1.998517e-05	2	

$P(F(2,87) > s_T/s_R) = 0.14$

4.0 Conclusions

Threshold voltage measurements have been taken from four by four transistor arrays fabricated in the baseline of the microfabrication lab at UC Berkeley. Our data is corrupted by problematic metal1 to metal2 via contacts. However, we are able to statistically process the measurements and draw some conclusions about the mismatch behavior.

The first result is that devices along one diagonal tend to have higher threshold voltage. There is not a good explanation for this phenomenon.

Second, for the NMOS devices we observe that mismatch is lower in the center of the wafer while for the PMOS is lower in the periphery. In both cases the difference is statistically significant, but small.

Third, we calculated the average variation of the threshold voltage across the array, across the die (between arrays) and across the whole wafer (between dies). We found, as we expected, that as we go from smaller to larger structures the variation is reduced.

Last, we considered the effect of the distance between transistors in the mismatch. We found that distance is not a significant factor when the devices have small size and are close to each other.

Appendix

In this appendix we summarize the results from a second wafer. For convenience the figures and the tables have the same number as the corresponding for the first wafer.

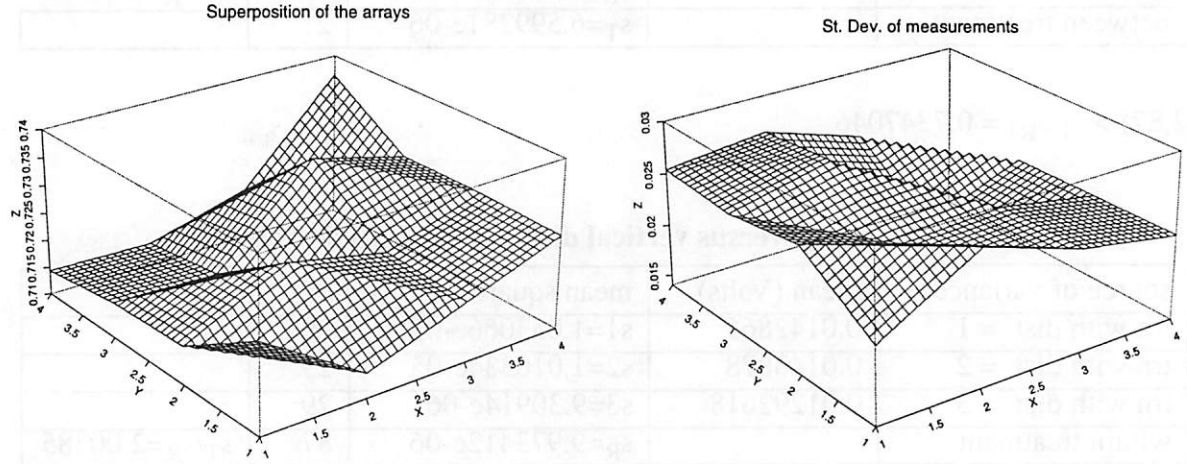


FIGURE A3. Average and Standard Deviation across 312 NMOS arrays.

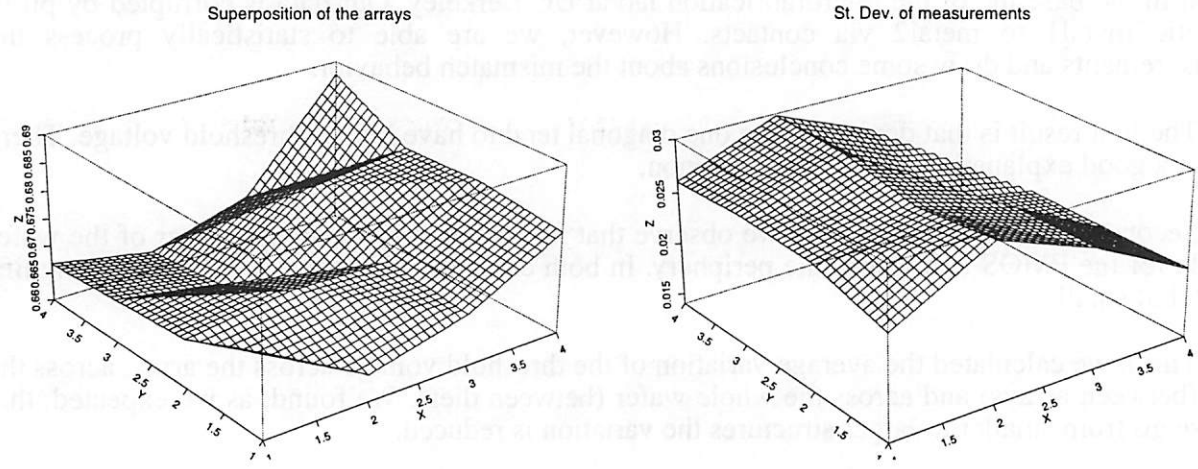


FIGURE A4. Average and Standard Deviation across 312 PMOS arrays.

From FIGURE 3 and FIGURE 4 we observe that the threshold voltage is higher across the one diagonal for this wafer also. There is very close similarity to the corresponding pattern of the first wafer that we examined.

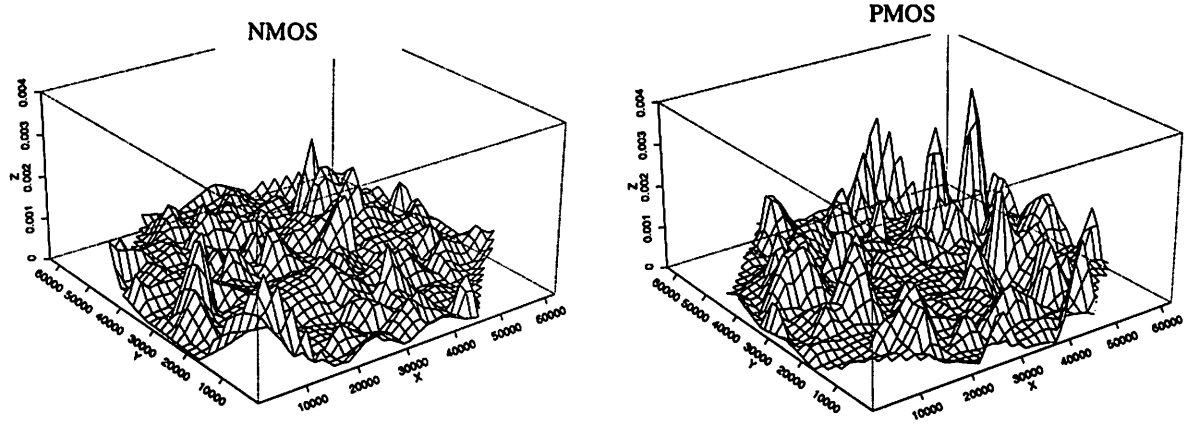


FIGURE A6. Perspective map of mismatch across the wafer for NMOS (left) and PMOS (right) transistors.

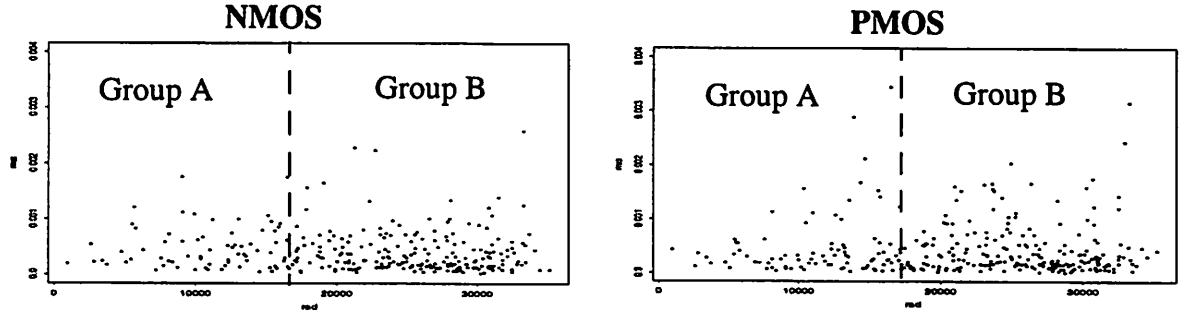


FIGURE A7. Scatter plot of mismatch with respect to radial distance for NMOS (left) and PMOS (right) transistors.

For the NMOS arrays n_A=72, n_B=211

$$\frac{S_B/n_B}{S_A/n_A} = 0.86$$

The hypothesis of equality of mismatch in the two groups against the alternative that the mismatch is higher in the center can be rejected with type I error: 0.1%

For the PMOS arrays n_A=83, n_B=227

$$\frac{S_B/n_B}{S_A/n_A} = 0.94$$

The hypothesis of equality of mismatch in the two groups against the alternative that the mismatch is higher in the center can be rejected with type I error: 10.3%

From the above we conclude that there is not significant difference of mismatch in the center and the periphery. The differences that appear are rather random than systematic.

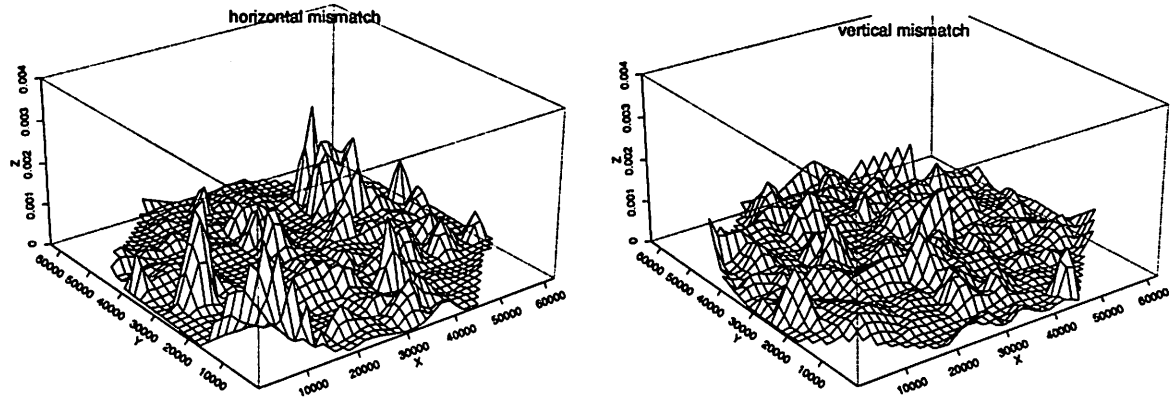


FIGURE A8. Horizontal and vertical mismatch across the wafer for the NMOS devices.

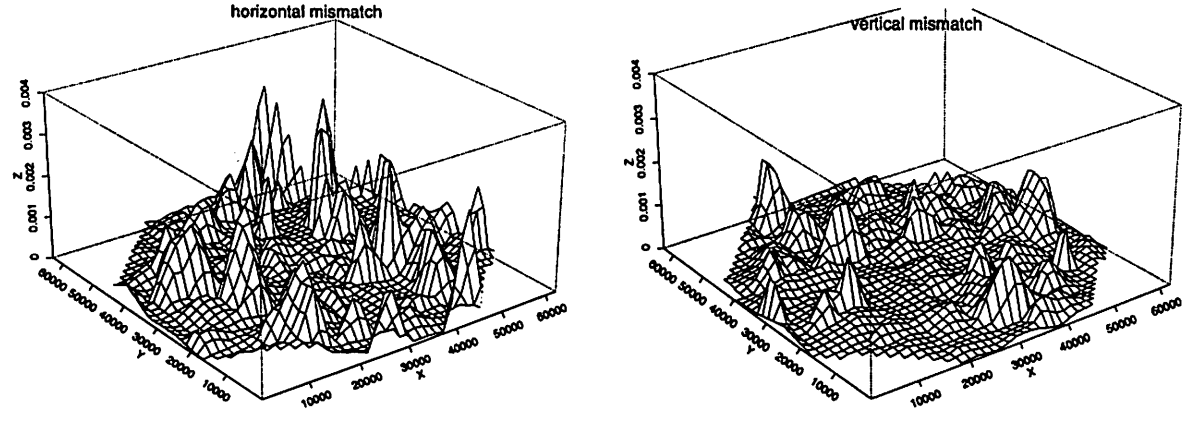


FIGURE A9. Horizontal and vertical mismatch across the wafer for the PMOS devices.

Table A1. Within array, die and wafer mismatch for NMOS devices

Source of mismatch	Average Mismatch
within array	20.4 mV
within die (between arrays)	11.5 mV
within wafer (between dies)	9.8 mV
between trans. in different arrays of the same die	23.4mV
between trans. in different dies	25.4 mV

Table A2. Testing of equality of mismatch from different sources for NMOS devices

Source of variation	mean square ($Volt^2$)	degrees of freedom	ratio
within array	s_a= 0.0004163085	52*6*(16-1)	s_a/s_d= 3.173168
within die	s_d= 0.0001311965	52*(6-1)	s_d/s_w= 1.364418
within wafer	s_w= 9.615566e-05	(52-1)	

Test of equality of mismatch within the array and within the die gives that we can reject the null hypothesis of equality with probability of error essentially 0. The test of equality of mismatch between the within die and the within wafer variation gives that we can reject the null hypothesis with probability of error 9.2%.

Table A3. Within array, die and wafer mismatch for PMOS devices.

Source of mismatch	Average Mismatch
within array	21.2mV
within die (between arrays)	12.7 mV
within wafer (between dies)	10.6 mV
between trans. in different arrays of the same die	24.7mV
between trans. in different dies	26.7 mV

Table A4. Testing of equality of mismatch from different sources for PMOS devices.

Source of variation	mean square ($Volt^2$)	degrees of freedom	ratio
within array	sa=0.0004516236	52*6*(16-1)	sa/sd=2.784553
within die	sd=0.0001621889	52*(6-1)	sd/sw=1.437304
within wafer	sw=0.0001128425	(52-1)	

Test of equality of mismatch within the array and within the die gives that we can reject the null hypothesis of equality with probability of error essentially 0. The test of equality of mismatch between the within die and the within wafer variation gives that we can reject the null hypothesis with probability of error 6.0%.

Table A5. Analysis of variance to examine dependance of mismatch on distance in horizontal direction for NMOS transistors.

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1		s1=8.966728e-06	27	
trn with dist. = 2	0.01717022	s2=9.311539e-06	27	
trn with dist. = 3	0.01654502	s3=1.980449e-05	27	
within treatment		s_R=1.269425e-05	81	s_T/s_R=0.21664
between treatment		s_T=2.75008e-06	2	

$P(F(2,87) > s_T/s_R) = 0.81$

Table A6. Analysis of variance to examine dependance of mismatch on distance in vertical direction for NMOS transistors.

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01705387	s1= 7.400891e-06	27	
trn with dist. = 2	0.01767579	s2= 1.156053e-05	27	
trn with dist. = 3	0.01790014	s3= 2.236883e-05	27	
within treatment		s_R= 1.377675e-05	81	s_T/s_R= 0.3907
between treatment		s_T= 5.381981e-06	2	

$P(F(2,87) > s_T/s_R) = 0.68$

Table A7. Analysis of variance to examine dependance of mismatch on distance in horizontal direction for PMOS transistors.

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01353845	s1=7.399386e-06	30	
trn with dist. = 2	0.01403697	s2=5.92278e-06	30	
trn with dist. = 3	0.01534171	s3=2.429329e-05	30	
within treatment		s_R=1.253849e-05	90	s_T/s_R=2.1438
between treatment		s_T=2.688005e-05	2	

$P(F(2,87) > s_T/s_R) = 0.12$

Table A8. Analysis of variance to examine dependance of mismatch on distance in vertical direction for PMOS transistors.

source of variance	mean (Volts)	mean square ($Volt^2$)	df.	ratio
trn with dist. = 1	0.01939411	s1=2.051774e-05	29	
trn with dist. = 2	0.01927744	s2=1.385921e-05	29	
trn with dist. = 3	0.01692616	s3=2.484398e-05	29	
within treatment		s_R=1.979831e-05	87	s_T/s_R=2.93785
between treatment		s_T=5.816451e-05	2	

$P(F(2,87) > s_T/s_R) = 0.058$

The last case shows some difference in the means, but it is a rather a random phenomenon since the mean that corresponds to longer distance is smaller.

Again, we conclude that there is not significant dependance of mismatch on distance within the array.

References

[1] Kadaba R. Lakshmikumar, Rubert A. Hadaway, Miles A. Copeland, "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design", IEEE Journal of Solid State Circuits, December 1986.

[2] Marcel J. M. Pelgrom, Aad C. J. Duimaijer, Anton P. G. Welbers, "Matching Properties of MOS Transistors", IEEE Journal of Solid State Circuits, October 1989.

[3] K. A. Brownlee, "Statistical Theory and Methodology in Science and Engineering", Second Edition, John Wiley & Sons.

[4] George E.P Box, William G. Hunter, J. Stuart Hunter, "Statistics for Experimenters", John Wiley & Sons.