EQUIPMENT ANALYSIS AND WAFER PARAMETER PREDICTION USING REAL-TIME TOOL DATA

Sherry F. LEE and Costas J. SPANOS

Department of Electrical Engineering & Computer Sciences
University of California, Berkeley CA 94720

We propose a system which uses real-time equipment sensor signals to automatically detect and analyze semiconductor equipment faults, and evaluate the impact of the fault on the wafer parameters. The system, which has been applied on plasma processes, consists of three modules: (1) fault detection, (2) fault analysis, and (3) prediction of final wafer parameters such as etch rate, uniformity, selectivity, and anisotropy.

1.0 Introduction

To compete in today's semiconductor industry, companies must continuously improve upon their manufacturing skills to both maintain high yield and reduce the cost of ownership of the equipment on the manufacturing line. A key element required to achieve these goals is to monitor the equipment to ensure that the semiconductor wafers are processed properly at each step. Measuring each wafer after it completes each step, however, is especially difficult in semiconductor factories producing chips with over 100 manufacturing steps. Moreover, due to throughput requirements, each wafer processed in each machine can not be measured individually. Present practice is to measure monitor wafers periodically, perhaps at the start of each work shift, after performing maintenance, or after changing the machine settings. Unfortunately, monitor wafers give no guarantee that subsequent production wafers will be processed properly. Thus, instead of detecting equipment faults causing yield loss early in the process flow, yield loss is usually found at the very end of the processing line.

We propose a novel system which uses equipment sensor signals to automatically detect equipment faults in real-time, and analyze and evaluate the effect of the fault on wafer parameters on a run-to-run basis. The three modules of the system are (1) detection of equipment malfunctions, (2) analysis (classification) of equipment faults, and (3) the prediction of output wafer parameters (Figure 1). The system impacts the semiconductor fabrication line by reducing the scrap produced by the equipment, reducing the down-time, and reducing the mean-time-to-repair. The result is a reduction in the overall cost of ownership of the equipment.

This general methodology is verified on a plasma etcher, one of the costliest pieces of equipment in the semiconductor fabrication line. Not only is the etcher usually a bottleneck piece of equipment, it is difficult to control because it is not well understood. Most importantly, each etcher can generate up to $100,000 worth of scrap per hour. The plasma etcher used in this work is a Lam Rainbow 4400 polysilicon etcher.

In this paper, the designed experiment is described first. A discussion of the fault detection and analysis modules is next, followed by a description of the wafer parameter prediction module.

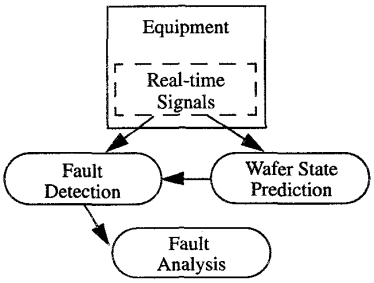

Figure 1 Proposed system

2.0 Designed Experiment

This section describes the experiment conducted to obtain the real-time data sets used to develop and verify the system. First, the wafer test structure is briefly described, followed by a discussion of both the training and the prediction experiments. Finally, the measurements taken on each wafer are described along with a discussion of the real-time signals collected during wafer processing.

2.1 Test Structure

The test structure was designed so all processes of interest are simultaneously obtained in the same etch step. Due to complex loading effects, this method results in more accurate etch rates and selectivities than etching blanket wafers individually[1]. A simplified view of the test structure indicating the etched surfaces is shown in Figure 2. First, a 600Å thermal gate oxide is grown on the 4" wafers, followed by 5500Å n+ doped polysilicon, deposited via low pressure chemical vapor deposition. After a 20 minute nitrogen anneal at 950°C, 2800Å undoped low temperature oxide (LTO) is deposited by chemical vapor deposition. A three step mask process is required to build the test structure.

2.2 Training and Prediction Experiments

In both the training and prediction experiments, a fixed pre-etch recipe was used for all runs. The main etch recipe was modified according to a designed experiment described below. To obtain accurate etch rates the main etch was a timed etch, so no overetch was performed. The input parameters varied in the experiment are the chamber pressure, RF forward power, electrode gap spacing, the ratio of Cl_2 to He, and the total gas flow of Cl_2 and He. Because the ratio and total gas flows are more significant to the etch results, they were varied in the experiment instead of the individual gas flows. The output wafer parameters of interest are the etch rate of polysilicon, selectivity of polysilicon to oxide and I-line positive photoresist, polysilicon wafer uniformity, and anisotropy.

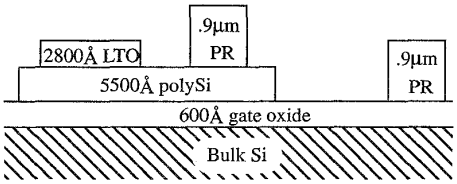

Figure 2 Test structure for the experiment.

2.2.1 Training Experiment

The training experiment consisted of two phases. Phase I, the variable screening stage, determined the statistically significant variables in the models. Phase II assessed the quadratic nature of the system via a star design. The input values used for all experiments are listed in Table 1, in terms of percent offset from the nominal values. The particular values were chosen to cover a wide range of operating conditions of the machine. Of the 37 runs in both phases of the training experiment, including replicated runs, 10 were eliminated before modeling due to unstable real-time signals or misprocessing.

Table 1: Change in % From Nominal

Parameter	Phase I	Phase II	Prediction
Pressure	15%	22.5%	10%
Power	15%	22.5%	10%
Gap	11%	17%	10%
Flow Ratio	19%	23%	10%
Total Flow	11%	22%	10%

Phase I consists of a two-level, 16 run fractional 2^{5-1} factorial design and 4 center points. The design is resolution V with no blocking, but drops to resolution III when blocked for time and split lots. The design is essentially resolution V because blocking was not a factor in any of the phase I response surface models. Assuming that four factor interactions are negligible, this experiment provides a good estimate of the main effects.

The variable screening analysis was performed by building models from the phase I data. The statistical significance of each parameter was determined via the student-t test at the 0.05 significance level. Results of the analysis show that although input settings are not all statistically significant in every model, all are required to model the four output characteristics of interest.

Additional runs were performed in phase II to estimate the quadratic behavior of the system. The models are limited to quadratic terms to limit complexity. The phase II runs consisted of center points and "star" points, arranged symmetrically along the axis of each variable[2]. Two star points were run for each variable. Two center points were also run, for a total of 12 additional runs.

2.2.2 Prediction Experiment

The purpose of the prediction experiment is to collect another data set used to simulate equipment faults and test the prediction capability of the models. The prediction experiment was run approximately four weeks after the phase II experiment. The input settings for this experiment were varied one at a time.

2.3 Wafer Measurements

In both experiments, film thickness measurements were taken by a Nanometrics Nanospec AFT system on 9 die per wafer. Four points were measured on the outer perimeter of the wafer, four were measured half-way from the edge of the wafer, and one point was measured at the center. Measurement error was approximately 10Å. The Alphastep 200 Automatic Step Profiler was used to confirm the Nanospec measurements. Film thicknesses were measured before and after etching; etch rates at each measured point are calculated by subtracting the post-etch from the pre-etch measurements, and dividing by the etch time. Wafer etch rates are averaged over the 5 inner points. Uniformity is calculated by scaling the difference between the etch rates of the outer and the inner rings by the etch rate of the inner ring.

2.4 Real-time Data

The real-time data collected from the plasma etcher are comprised of various electrical and mechanical signals. Between six and thirteen signals are collected. Six signals are collected via a Comdel Real Power Monitor (RPM-1), placed directly above the upper electrode[3]. The remaining seven signals are collected via the LamStation software, which reads the signals from the SECSII serial port on the etcher[4].

Because a number of the measurements are related electrically or mechanically, many of the signals are highly correlated. A few signals are collected from different places in the equipment by the two different monitoring systems. Although correlated, these signals are not identical. The important signals monitored are RF Power, RF Voltage (rms), RF Current (rms), Load Impedance, RF Phase Error, Tune Vane Position, Load Coil Position, Peak-to-Peak Voltage, and End Point Data.

3.0 Fault Detection and Analysis

Since the data are collected sequentially at a sampling rate of 1 Hz the real-time signals are correlated in time, demonstrating time series behavior. The real-time fault detection module utilizes time series models to analyze the real-time signals. The objective is to use these automatically collected signals to establish the baseline behavior of a complex tool and later detect deviations from this baseline. The fault detection algorithm is implemented through RTSPC, a software utility which automatically collects real-time sensor data and generates real-time alarms[7]. Examples of faults include shifts in the process parameters, such as changes in chamber pressure, RF power, or gas flows.

If no equipment faults are detected, normal operation of the machine continues. When a malfunction is detected, the diagnostic routine is triggered, and an alarm is generated to alert the operator. After being filtered in RTSPC, the real-time residual data form distinct signatures which can be traced back to a specific equipment fault or group of faults. Initially, a training set of faults must be generated to teach the diagnosis module fault signatures, creating a library of signatures. Discriminant analysis techniques are employed to analyze the equipment faults and train the system. Once training is complete, equipment faults are detected and analyzed on a run-to-run basis[5][6][7].

4.0 Wafer Parameter Prediction

Empirical models are used to predict the outcome of each wafer immediately after it is processed by the equipment. To provide useful prediction capabilities, robust prediction models of the machines are required. The industry standard is to use response surface methodology to build models relating the input settings of the machine to the output characteristics. Response surface models, however, become unusable in time due to machine drifts, rendering them ineffective for prediction.

We propose that using real-time signals to build the models results in better prediction capabilities. Four types of regression methods were explored: ordinary least squares (OLS) regression, ridge regression, principal component regression (PCR), and partial least squares regression (PLSR). No one modeling method was overwhelmingly better than the others, although OLS regression resulted in slightly better prediction models for polysilicon etch rate. PCR and PLSR models, however, were less sensitive to overfitting.

The time series nature of the signals is not exploited in the prediction module. Instead, each signal is averaged over the duration of the main etch step, which lasts approximately 30 seconds. Approximately 30 points are collected per signal per wafer etch. Since the wafer-to-wafer variance of the real-time signals is much larger than the within wafer variance, the average values per signal across each wafer are used as the input for the prediction models built with the real-time signals. The training model is built from data collected during the training experiment. The final prediction metric is based on how well the training model predicts the outcome of the data collected during the prediction experiment. By using this prediction data, the true prediction capability of the models can be gauged. The two metrics used are the average prediction error (PE) and the standard error prediction (SEP), where Y_i is the ith observation, $\hat{Y}_i$ is the predicted value of the ith point, and n is the number of observations:

$$PE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|, \quad SEP = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-1}}.$$

It must be emphasized that models with the best adjusted R^2 value are not necessarily good for prediction. The best models built with real-time data have adjusted R^2 values of 0.95 or greater. These models, however, can potentially have huge prediction errors even if all the terms in the model are statistically significant. In some cases, especially for OLS models, the prediction error is huge. To allow for better prediction a few terms in the models are eliminated, at the cost of reducing the adjusted R^2 value.

Two sets of models were built for each of the wafer characteristics to show that the real-time signals are better suited for prediction than the machine input settings. The first uses the real-time signals and the second uses the input settings. Due to the small ranges of selectivities across the design space, models are created for the individual etch rates of gate oxide and photoresist instead of modeling the selectivities. The ranges of each output parameter are listed to give a relative measure of the accuracy of the prediction. The resulting PE and SEP values for the etch rate and uniformity models are listed in Table 2 in terms of percent error of the range. In all cases, the models built with the real-time data are superior to those built with input settings. The best models for uniformity have 25% prediction error, indicating that uniformity can not be successfully modeled with this data set.

One reason for the better prediction is that models built with input settings generally include statistically significant

blocking terms to account for differences in the machine between sets of runs. In this experiment, the chlorine bottle was refilled between phases I and II, causing a slight shift in the baseline behavior of the machine. This shift is accounted for in the models built with input settings through a blocking parameter. Blocking terms, however, can not be used in prediction models. Without the blocking parameter, the prediction capability of the model built with input settings suffers.

Table 2 Comparison of Models Built Using Input Settings vs. Real-Time Data

Wafer Characteristic	Metric	Input Settings: Error in % of Range	Real-Time Signals: Error in % of Range
Polysi Etch Rate	PE	15%	7%
(2000 - 3200Å)	SEP	23%	9%
Oxide Etch Rate	PE	34%	4%
(160 - 230Å)	SEP	82%	6%
PR Etch Rate	PE	31%	3%
(1280 - 1750Å)	SEP	70%	5%
Uniformity	PE	58%	25%
(4 - 20%)	SEP	83%	25%

Unlike the fixed input settings the real-time signals change with the state of the machine, eliminating the need for blocking terms. Figure 3 compares the modeling results of 6 centerpoint wafers. The model built with the fixed input settings predicts a constant etch rate, while the real-time model adjusts the prediction as a result of small changes in the machine state. Thus, from the results listed in Table 2 and the above argument, we conclude that models built using real-time data predict etch rates with more accuracy than those built with input settings.

5.0 Conclusions

The three-module system presented is especially powerful because it does not depend upon monitor wafers or expensive metrology; rather, it uses non-invasive real-time signals collected automatically from the tool while the wafer is processing. These signals are used effectively to detect and analyze equipment faults. Prediction models have also been developed using real-time signals. Since the wafer parameters are predicted immediately after the wafer has finished processing in the machine, important yield information is obtained on a run-to-run basis. In addition to catching problems with the machine, the real-time data can be used to assess the quality of the wafer immediately after processing, making it possible to ensure that only wafers worth processing continue down the line.

Another consequence of the prediction capability of the real-time models is that inexpensive run-to-run control is possible. Future work includes pursuing such a run-to-run control scheme of plasma etch equipment which will bring the specified output parameters back to their target value in the case of equipment drift.

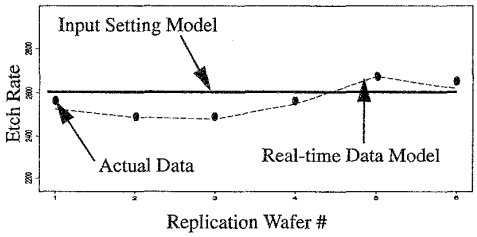

Figure 3 Comparison of the model built with input settings versus the model built with real-time signals.

6.0 References

1) G.S. May, J. Huang, C.J. Spanos, "Statistical Experimental Design in Plasma Etch Modeling," *IEEE Trans. Semiconductor Manufacturing*, vol. 4, no. 2 (1991).

2) G.E.P. Box, N.R. Draper, Empirical Model-Building and Response Surfaces, Wiley (1987).

3) Real Power Monitor (RPM-1), Comdel Inc.

4) *LamStation Rainbow*, v 3.6, Brookside Software (1991).

5) C.J. Spanos, S. Leang, S.F. Lee, "A Control & Diagnosis Scheme for Semiconductor Manufacturing," *Proc. American Control Conference*, San Francisco (1993).

6) S. F. Lee, C. J. Spanos, "Real-time Diagnosis for Plasma Etch Equipment," *Techcon*, pp. 16-18 (1993).

7) S.F. Lee, E.D. Boskin, H.C. Liu, E. Wen, C.J. Spanos, "RTSPC: A Software Utility for Real-Time SPC and Tool Data Analysis," to appear in *Trans. Semiconductor Mfg.*