A HYBRID NON-THREADED RUN-TO-RUN CONTROL AND INCORPORATION OF MULTIPHYSICS MODELS INTO SEMICONDUCTOR VIRTUAL METROLOGY

by

Shijing Wang

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemical Engineering

The University of Utah

May 2016

Copyright © Shijing Wang 2016

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Shijing Wang**

has been approved by the following supervisory committee members:

Name	Role	Date Approved
Mikhail Skliar	, Chair	**3/10/2016**
Anthony Butterfield	, Member	**3/10/2016**
Geoffrey Silcox	, Member	**3/3/2016**
James Sutherland	, Member	**3/4/2016**
Jennifer Gray	, Member	**3/7/2016**

and by **Milind Deo** , Chair/Dean of

the Department/College/School of **Chemical Engineering**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

In order to ensure high production yield of semiconductor devices, it is desirable to characterize intermediate progress towards the final product by using metrology tools to acquire relevant measurements after each sequential processing step. The metrology data are commonly used in feedback and feed-forward loops of Run-to-Run (R2R) controllers to improve process capability and optimize recipes from lot-to-lot or batch-to-batch.

In this dissertation, we focus on two related issues. First, we propose a novel non-threaded R2R controller that utilizes all available metrology measurements, even when the data were acquired during prior runs that differed in their contexts from the current fabrication thread. The developed controller is the first known implementation of a non-threaded R2R control strategy that was successfully deployed in the high-volume production semiconductor fab. Its introduction improved the process capability by 8% compared with the traditional threaded R2R control and significantly reduced out of control (OOC) events at one of the most critical steps in NAND memory manufacturing. The second contribution demonstrates the value of developing virtual metrology (VM) estimators using the insight gained from multiphysics models. Unlike the traditional statistical regression techniques, which lead to linear models that depend on a linear combination of the available measurements, we develop VM models, the structure of which and the functional interdependence between their input and output variables are determined from the insight provided by the multiphysics describing the operation of the processing step for which the VM system is being developed. We demonstrate this approach for three different processes, and describe the superior performance of the developed VM systems after their first-of-a-kind deployment in a high-volume semiconductor manufacturing environment.

CONTENTS

ABSTRACT iii

LIST OF FIGURES viii

LIST OF TABLES xii

ACRONYMS xiii

ACKNOWLEDGEMENTS xvi

CHAPTERS

1. INTRODUCTION **1**

1.1 Process Control Systems for Semiconductor Manufacturing 1
1.2 Introduction to R2R Control 5
1.2.1 R2R Control Algorithms 7
1.2.2 Threaded and Non-threaded R2R Control 11
1.3 R2R Control and Virtual Metrology 14
1.4 Introduction to Virtual Metrology 15
1.4.1 Principle Component Analysis 16
1.4.2 Partial Least Squares 19
1.4.3 Neural Networks 22

2. SCOPE OF THIS WORK **26**

2.1 Motivations and Objectives 26
2.2 Overview of This Dissertation 28
2.2.1 Hybrid Non-Threaded R2R Control 28
2.2.2 Etch Rate Prediction of Diluted HF Solution 29
2.2.3 Thickness Profile Prediction of Diffusion Furnace 30
2.2.4 Etch Rate of Resist Descum 31

3. HYBRID NON-THREADED RUN-TO-RUN CONTROL **32**

3.1 Abstract 32
3.2 Introduction 32
3.3 Problem Statement 36
3.3.1 Bias in State Estimation 39
3.3.2 The Change of Matrices Sizes 40
3.3.3 Long Execution Time for State Estimation 41
3.4 Background 41
3.4.1 Kalman Filtering 41
3.4.2 Jade Algorithm 42

3.4.3 EWMA Method . . . 44
3.4.4 Model Regularization . . . 45
3.5 Methods . . . 46
3.5.1 Two Model-Based Non-threaded R2R State Space Representations . . . 46
3.5.2 The Hybrid Non-threaded R2R Control Methodology . . . 48
3.5.3 Handling of Varying Matrices Sizes . . . 52
3.5.4 Handling of Long Execution Strategy . . . 54
3.6 Demonstration . . . 55
3.6.1 Non-threaded Run-to-Run: CMP Mismatch Handling . . . 55
3.6.2 Non-threaded Run-to-Run: Photo Tool Mismatch . . . 58
3.7 Discussion . . . 61
3.7.1 State Estimation of the Non-threaded R2R Controller . . . 61
3.7.2 Tuning Non-threaded R2R Control . . . 63
3.7.3 Reduction of Qualification Runs Using Non-threaded R2R . . . 64
3.8 Summary . . . 64

4. ETCH RATE PREDICTION OF SILICON DIOXIDE FILM IN A DILUTED HF SOLUTION . . . 66

4.1 Abstract . . . 66
4.2 Introduction . . . 66
4.3 Motivations . . . 69
4.4 Background . . . 70
4.4.1 Virtual Metrology Using Statistical Models . . . 70
4.4.2 Etch Rate Prediction Using PLS . . . 75
4.4.3 Limitations of Statistical Models . . . 77
4.4.4 Scope of This Work . . . 79
4.5 Virtual Metrology Using a Multiphysics Model . . . 79
4.5.1 Etch Rate Prediction Background . . . 79
4.5.2 HF Etch Rate and Its Mechanism . . . 80
4.5.3 Etch Rate Prediction Using the Multiphysics-Based Model . . . 83
4.5.4 Etch Rate Prediction Reliance Index . . . 84
4.5.5 Etch Rate Virtual Metrology Design . . . 87
4.5.6 Virtual Metrology Results Analysis and Discussions . . . 87
4.6 Integration of Virtual Metrology into Run-to-Run Control . . . 90
4.7 Benefit Analysis . . . 94
4.7.1 Excursion Prevention . . . 95
4.7.2 Process Capability Improvement . . . 95
4.7.3 Yield Improvement . . . 97
4.7.4 Cycle Time Reduction . . . 97
4.7.5 Cost Reduction . . . 97
4.8 Summary . . . 98

5. A GENERIC DIFFUSION FURNACE VIRTUAL METROLOGY . . . 100

5.1 Abstract . . . 100
5.2 Introduction to the Diffusion Furnace . . . 100
5.2.1 Introduction to Furnace Process . . . 100
5.2.2 Introduction to Furnace Equipment . . . 104

5.3 Introduction to Diffusion Furnace R2R Control 107
5.3.1 Furnace R2R Control Model 107
5.3.2 State Space Representation 108
5.3.3 State Estimation of Furnace R2R Control 110
5.3.4 Furnace R2R Control and Its Performance 112
5.4 Motivations for VM of Diffusion Furnace 113
5.5 Incorporating Multiphysics into Furnace VM 114
5.5.1 Background 114
5.5.2 Design of Experiment 116
5.5.3 Curve Fitting Results 118
5.5.4 VM Model Update Methods 120
5.5.5 Queue Time and Metrology 123
5.5.6 Effects of R2R Control and Its Compensation 124
5.6 VM Results and Conclusions 125
5.7 Summary and Future Work 128

6. THE OXYGEN PLASMA RESIST DESCUM VIRTUAL METROLOGY 129

6.1 Abstract 129
6.2 Introduction 130
6.2.1 Introduction to Plasma Dry Etching 130
6.2.2 Introduction to O_2 Plasma Photo Resist Etching 131
6.2.3 Chemical Reactions in Photo Resist Descum 134
6.2.4 Physical Etching Models 137
6.2.5 Model Parameters Candidates 138
6.3 Partial Least Squares Model 139
6.3.1 Process Data and Model Parameters Selections 139
6.3.2 PLS Model and Validation 141
6.3.3 "Zonal" Data Analysis 142
6.3.4 Model Validation Methods 143
6.3.5 Model Update Methods 144
6.3.6 PLS Model Discussions 146
6.4 New Methods 148
6.4.1 Introduction of New Model Parameters 148
6.4.2 The VM Model Based on DOE and Its Results 150
6.5 Challenges and Discussions 151
6.6 Conclusions and Future Work 152

7. CONCLUSIONS AND PROPOSED FUTURE WORK 154

7.1 Conclusions 154
7.1.1 Hybrid Non-threaded Run-to-Run Control 154
7.1.2 Etch Rate Prediction of Silicon Dioxide Film in Diluted HF Solution 155
7.1.3 A Generic Diffusion Furnace Virtual Metrology 156
7.1.4 Oxygen Plasma Resist Descum Virtual Metrology 157
7.2 Future Work 158
7.2.1 Context Matching and Relaxation 158
7.2.2 Auto Tuning of Non-threaded R2R Control 159
7.2.3 VM and Wafer Level Implant R2R 161

APPENDICES

A. CURVE FITTING IN MATLAB 163

B. DATA COLLECTION OF ETCH RATE AND HF WEIGHT PERCENT . 168

C. DESIGN OF EXPERIMENT FOR DESCUM 171

REFERENCES 175

LIST OF FIGURES

Figure | Page

1.1 Sample SPC chart 2

1.2 Sample FD trace data of the ESC inner and outer temperature 3

1.3 CD process capability is improved by 40% with a R2R controller 4

1.4 Process monitoring and control through SPC Chart 5

1.5 A new architecture for process monitoring and control 6

1.6 A framework of fab-wide R2R control [1]. 6

1.7 Two model update methodologies for R2R control: model update through estimating intercept (left graph); model update through estimating slope (right graph) 8

1.8 VM and W2W R2R control [2] 15

1.9 A principle component in the case of two variables: A. The loading for the principle component, also the direction of principle component B. The scores of the principle component, which are the projections of the sample points (1-6) on the principle component direction [3] 17

1.10 Partial least square regression [4] 21

1.11 Back-propagation neural network architecture 23

1.12 The simplest neural network: x is the input, z is the hidden layer, y is the output, d is the desired output, w_1 and w_2 are the weighting factors and J is the objective function or performance function. 24

3.1 Run-to-Run controller. 33

3.2 Metrology data are diluted due to increased control threads. 37

3.3 Non-threaded R2R control architecture 49

3.4 Evaluation of state estimation for threaded and non-threaded R2R control 53

3.5 Load balancing among different servers 55

3.6 Recess oxide before and post dry etch step. 56

3.7 CMP tool to tool mismatch post dry etch step. 57

3.8 Wafer level multiple inputs multiple outputs feed-forward and feedback non-threaded dry etch R2R control schematic. 57

3.9 One chamber (red) was deployed with non-threaded R2R control and the other three chambers (blue) remained threaded R2R control. 58

3.10 Out of control events were reduced after non-threaded Run-to-Run deployment. 59
3.11 Photo tool to tool mismatch in the dry etch R2R control. 59
3.12 Comparison of threaded and non-threaded intercept states updating frequency. 62
3.13 It took about 10 runs from zero initial state to steady state for CVD tool states. 62
3.14 Benchmarking performance between threaded and non-threaded R2R control. 63
4.1 Place of execution of VM in a process flow 67
4.2 Neural network based VM model for CVD thickness [5] 72
4.3 20-liter ONB tank schematics 76
4.4 Chemical charge window of HF flow rate 76
4.5 PLS evaluation results 78
4.6 Composition changes of each species for initial concentration [HF] equal to 0.05 mol/liter 81
4.7 The calculated fraction of each component in an HF solution as a function of total fluoride concentration [6] 82
4.8 Virtual metrology design architecture 88
4.9 Initial results $r^2 = 0.83$: correlation between actual measured etch rate and predicted etch rate 88
4.10 Long-term results $r^2 = 0.64$: correlation between actual measured etch rate and predicted etch rate 89
4.11 Shape and topography in a semirecessed and fully-recessed LOCOS structure 92
4.12 LOCOS process steps 92
4.13 Run-to-Run controller model update through virtual metrology 94
4.14 SPC performance for “Post Nitride Strip Oxide Thickness”: VM R2R vs. no VM R2R. The sampling rate is 100% of the lots. 96
4.15 SPC performance for “Post Diffusion Thin Oxide Thickness”: VM R2R vs. no VM R2R. The sampling rate is 10% of the lots. 96
5.1 Growing oxide in diffusion furnace consumes some of the silicon substrate 102
5.2 Depositing poly-silicon in diffusion furnace does not consume any silicon substrate 103
5.3 The damaged lattice and its repair: (A) The lattice is damaged by an ion implant process (B) The lattice is repaired by a rapid thermal process . . . 103
5.4 Kokusai vertical diffusion furnace 104

5.5 The configuration of boat, liner and tube . . . 105
5.6 Heater center and monitoring wafer location in the boat drawing and their misalignment . . . 106
5.7 The filter horizon and predictive horizon for model predictive control. . . 113
5.8 A furnace R2R control (6×9) improved process capability in terms of C_{pk} by more than 90%. . . . 114
5.9 Design of experiment: deposition rate change by tuning knob at each boat slot (or furnace position) . . . 117
5.10 Accumulation effect and its contributions of a diffusion furnace thickness profile . . . 119
5.11 Overlay of actual metrology and fitted curve . . . 120
5.12 Model update by changing intercept state only . . . 121
5.13 Metrology queue time . . . 123
5.14 VM prediction gap at different times, furnace step vs. metrology step . . . 124
5.15 VM prediction gap at different times is improved between furnace step and metrology step . . . 126
5.16 The overlay plot of actual metrology and VM prediction at furnace step . 127
5.17 The correlation of actual metrology and VM prediction at furnace step . . 127
6.1 A simple plasma reactor . . . 131
6.2 O2 plasma reactions with photo resist [7] . . . 132
6.3 The chamber pressure influences the plasma etching types [8] . . . 133
6.4 Resist etch rates in downstream plasma using O_2 or O_2/N_2O chemistries with different temperatures [9] . . . 134
6.5 DNQ-Novolac photoresists . . . 135
6.6 The volume of resist is stripped by time by monitoring CO^* band [10] . . . 137
6.7 Spectral response of atomic oxygen as a function of chamber pressure [11] 139
6.8 The plasma power effect on photo resist removal rate [11] . . . 140
6.9 Chuck temperature of fifty consecutive runs . . . 141
6.10 PLS model evaluation results . . . 142
6.11 Metrology data are classified into high and low zones . . . 143
6.12 Improved PLS model evaluation result is obtained after excluding data in interquartile range . . . 144
6.13 Actual metrology and predicted metrology overlay for Case 1. . . . 147
6.14 Actual metrology and predicted metrology overlay for Case 2 and Case 3. 147
6.15 Etch rate and relative O atom concentration changes with hydrogen and oxygen mixing ratio [12]. . . . 148

6.16 Etch rate and relative O atom concentration changes with various percentages of nitrogen in the oxygen plasma [13] 149
6.17 PLS analysis of descum DOE data. 150
6.18 Etch rate response with the gas ratio factor 151
6.19 The prediction and measurement correlation of the VM model based on DOE. 152
7.1 Prediction error bar grows when context matching level is relaxed. 159
7.2 The process of auto-tuning state estimation Q and R for a non-threaded R2R control. 160
7.3 Poly-silicon resistor [14] 161
7.4 The relationship between the poly-silicon resistance and the gain size of the poly-silicon [14] 162
A.1 Data selection for curve fitting application 164
A.2 Data fitting in the curve fitting application 165
A.3 Curve fitting results 166
A.4 The plot of the curve fitting output 167
C.1 Etch rate response with the O_2 gas factor 173
C.2 Etch rate response with the H_2N_2 gas factor 173
C.3 Etch rate response with the RF forward power factor 174
C.4 Etch rate response with the chamber pressure factor 174

LIST OF TABLES

Table | Page

1.1 Control thread definition example: contexts of three chambers and two devices 12
3.1 Execution time of non-threaded state estimation 54
3.2 Head to head non-threaded R2R control comparisons 61
4.1 The measured and the predicted weight percent of HF solution 84
6.1 Process data items of a descum process 140
7.1 The context match and relaxation 158
B.1 1000:1/T2 HF etch rate and weight percent 169
B.2 500:1/T2 HF etch rate and weight percent 169
B.3 500:1/T4 HF etch rate and weight percent 170
B.4 100:1/T4 HF etch rate and weight percent 170
C.1 DOE of O_2 plasma descum 172

ACRONYMS

ANOVA - Analysis of Variance

APC - Advanced Process Control

B2B - Batch to Batch

BLUE - Best Linear Unbiased Estimation

BPNN - Back-propagation Neural Network

CD - Critical Dimension

CMP - Chemical Mechanical Polishing/planarization

CVA - Canonical Variate Analysis

CVD - Chemical Vapor Deposition

DE - Dry Etch

DI - Deionized

DIW - Deionized Water

DNQ - Diazonaphthoquinone Photo Resist

DOE - Design of Experiment

DQ - Data Quality

ECD - Electrical Chemical Deposition

ER - Etch Rate

ESC - Electrostatic Chuck

EWMA - Exponential Weighted Moving Average

FD - Fault Detection

FNN - Fuzzy Neural Networks

FOUP - Front Opening Unified Pod

GOF - Goodness of Fit

GSI - Global Similarity Index

HF - Hydrofluoric Acid

IQR - Interquartile Range

ISI - Individual Similarity Index

ITRS - International Technology Roadmap for Semiconductors

JADE - Just-in-Time Adaptive Disturbance Estimation

L2L - Lot to Lot

LCL - Lower Control Limit

LMPC - Linear Model Predictive Control

LOCOS - Local Oxidation of Silicon

LPCVD - Low Pressure Chemical Vapor Deposition

LSL - Lower Spec Limit

LV - Latent Variables

MES - Manufacturing Execution System

MIMO - Multiple Inputs and Multiple Outputs

MPC - Model Predictive Control

MSE - Mean Square Error

MVA - Multivariate Analysis

NIPALS - Non-linear Iterative Partial Least Squares

NN - Neural Networks

NPW - Non-Process Wafers

OCAP - Out of Control Action Plan

ODE - Ordinary Differential Equations

OLS - Ordinary Least Squares

ONB - One Bath System

OOC - Out of Control

OOS - Out of Spec

PAC - Photoactive Compound

PCA - Principle Component Analysis

PCS - Process Control Systems

PLNN - Piecewise Linear Neural Networks

PLS - Partial Least Squares

PRESS - Predicted Residual Error Sum of Squares

PVD - Physical Vapor Deposition

R2R - Run-to-Run Control

RBFN - Radial Basis Function Neural Networks

RDA - Real-time Defect Analysis

REG - Registration

RF - Radio Frequency

RI - Reliance Index

RLS - Recursive Least Square

RMSECV - Root Mean Square Error Cross Validation

RPT - Rapid Thermal Process

SEE - Software Execution Engine

SEMI - Semiconductor Equipment and Materials International

SISO - Single Input and Single Output

SPC - Statistical Process Control

SRNN - Simple Recurrent Neural Networks

SVD - Singular Value Decomposition

SVR - Support Vector Regression

UCL - Upper Control Limit

USL - Upper Spec Limit

UVA - Univariate Analysis

VM - Virtual Metrology

W2W - Wafer to Wafer

ACKNOWLEDGEMENTS

First I want to sincerely thank Dr. Skliar for providing me an opportunity to pursue my PhD degree at the University of Utah. During the past four years, he has spent many long working with me on research meetings, and he has been very kind to accommodate my working schedule at IM Flash. I would like to express my appreciation for his patience, help and support. He has been a great mentor throughout my research life at the University of Utah and I will remember the many things that I learned from Dr. Skliar. His guidance will have a life-long influence on me.

I would also like to thank Dr. Silcox, Dr. Sutherland, Dr. Butterfield and Jennifer Gray for serving as my committee members. Thank you all for the useful discussions and for your advice. Additionally, I want to thank Jennifer Gray, my former manager, for providing me the opportunity to work on these research projects towards my PhD degree.

The etch rate prediction projects of a diluted HF solution and O2 plasma virtual metrology would not been possible to complete without the support from the Wet Process, Run-to-Run, Fault Detection, Chemical Lab and Diusion teams at IM Flash. I want to thank Matthew Willford and Kendel Saunders of the Wet Process team for their support on data collection, design of experiments (DOE) and process knowledge sharing. Thanks to Robert Correa of the Fault Detection team for the data collection and strategy troubleshooting on the Fault Detection system. I want to thank Thad Parry and Dave Zaragoza for the tool software upgrade, so that I could integrate the virtual metrology with the Run-to-Run control. I also want to thank my many colleagues in the Chemical Lab for analyzing the chemical samples. Ben Taylor was my former colleague on the Run-to-Run team, and I also want to thank him for introducing me to Dr. Skliar, which opened the door for me to earn my PhD.

I thank Andy Beemer, Cyrus Fox and Karl Disher for their support in developing the diusion furnace virtual metrology system. Andy Beemer provided the DOE data. Thanks Andy for the meaningful discussions on this topic. Thanks to Karl Disher and Cyrus Fox for providing accurate boat slot dimensions and drawings. I also thank Jeremy Stout for his help and advice on the dry etch non-threaded R2R control development, so that we could deploy this new technology on one of the most critical processes in production.

Lastly, I would like to thank my parents and my brother for their love, support and encouragement. In particular, I want to thank my wife, Bing Xiao. It would not have been possible to complete this dissertation without her love and inspiration. During the past four years, she has shouldered many responsibilities so that I could spend the necessary time pursuing this PhD and chasing my dreams. Finally, I would like to thank my daughter, Joanna Wang, for her love. I wish for her that all her figure skating dreams will come true!

CHAPTER 1

INTRODUCTION

1.1 Process Control Systems for Semiconductor Manufacturing

The state of the art process control systems (PCS) in modern semiconductor manufacturing typically include statistical process control (SPC), fault detection (FD) and Run-to-Run control (R2R).

SPC is a continuous improvement methodology that utilizes statistical tools to monitor and control a process. The acronym SPC consists of three components: statistical, process and control. First, the "statistical" component summarizes the data through descriptive statistics (e.g., mean, median, range or standard deviation) and determines the distribution. Secondly, the "process" component refers to the transformation of a set of inputs including material, actions, methods and operations, into desired outputs in the form of product, information, services or general results [15]. In semiconductor manufacturing, "process" refers to recipes, tools or chambers, chemicals or gases, raw materials and people. Finally the "control" components can include control charts, out of control (OOC) rules, trending rules and reaction mechanisms. The purpose of SPC is to monitor and control the process to ensure that the process operates normally and at its maximum potential. A sample SPC chart is shown in Figure 1.1. Here, the SPC chart contains a center line (or target), control limits and specification limits [16]. The control limits are calculated by the distribution of historical data in normal operation, and if the data are normally distributed, then the control limits can be calculated using the following methods:

$$UCL = \mu + 3\sigma \tag{1.1}$$

$$LCL = \mu - 3\sigma \tag{1.2}$$

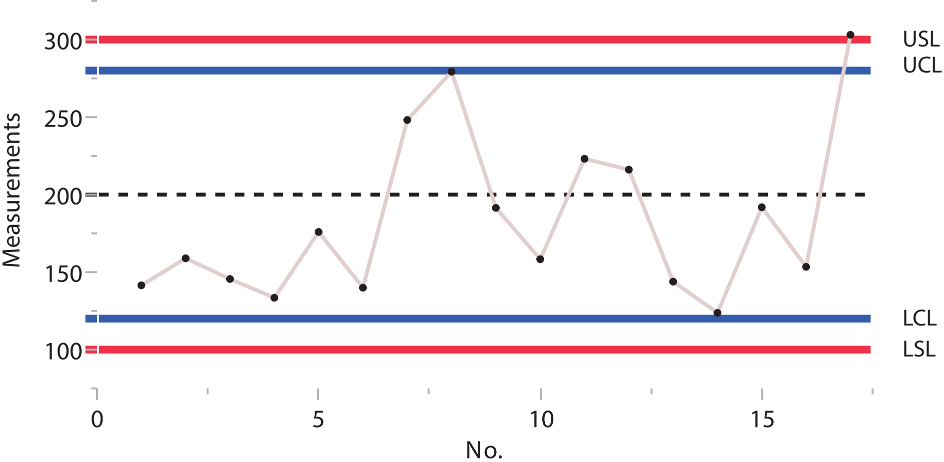


Figure 1.1. Sample SPC chart

where *UCL* is the upper control limit, *LCL* is the lower control limit, μ is the population mean and σ is standard deviation of the population. There are many types of SPC control charts and control limits can be calculated quite differently [17] for the various types of control charts. Specification limits (*USL* or *LSL*) are engineering limits, which are usually defined by the requirements of internal or external customers. Specification limits are not determined statistically.

In the past decade, FD [18,19] and R2R control systems [20] have been added into process monitoring and control systems under the PCS umbrella. A conventional process control systems (e.g., SPC) exists only as a postprocess control-based on metrology data, while FD system allows real-time control of equipment and process parameters before and during wafer processing. FD system collects and consolidates process trace data from wafer processing tools in real time. Univariate statistical analysis or multivariate statistical analysis can be used to determine control limits of the consolidated data. These established control limits can trigger fault detection and reaction mechanisms. This is also called an out of control action plan (OCAP). The benefits of FD include preventing excursions, identifying root causes of faults, matching tools and chambers' performance and optimizing tool maintenance and metrology operations. For example, electrostatic chuck (ESC) inner and outer temperature data were collected in Figure 1.2 [21], which are process trace data. The "signature" of them can be used to monitor the tool health.

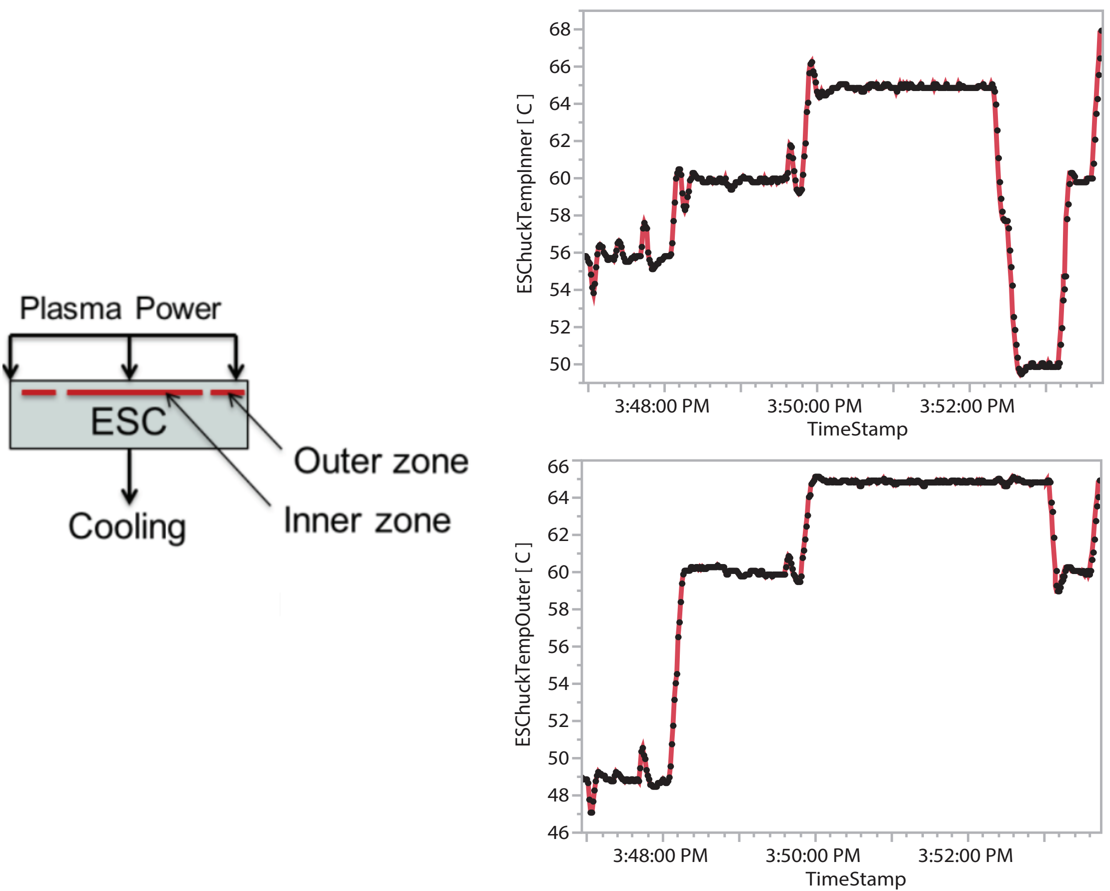


Figure 1.2. Sample FD trace data of the ESC inner and outer temperature

FD and SPC are mainly responsible for monitoring process, while R2R control plays an important role to control process and improve process capability. The process capability is often measured in a capability index C_{pk} [22] with a target value of 1.33 or 1.66 depending on customer's requirement:

$$C_{pk} = \min[\frac{USL-\mu}{3\sigma}, \frac{\mu-LSL}{3\sigma}] \tag{1.3}$$

As semiconductor features shrink in size and pitch, some critical processes are not even feasible to run without R2R controllers because the process variations from incoming steps or current steps are likely greater than the specification window. R2R controllers consist of inputs, outputs and a process model. The recipe setpoints (or inputs) are adjusted based on metrology data to achieve desired output. The controller mode can be either feed-forward (pre-metrology) or feedback (post-

metrology) and the controller types include single input and single output (SISO) and multiple inputs and multiple outputs (MIMO).

Dry etch refers to the removal of material by exposing the material to a bombardment of ions, typically a plasma of reactive gases such as the oxygen plasma, that dislodge portions of the material from the exposed surface. Unlike the isotropic etching by a wet etch, the dry etch process typically etches anisotropically. Figure 1.3 shows that a R2R controller deployed on dry etch process improved the process capability (or C_{pk}) by more than 40% [23] and similar results were obtained on the diffusion process [24], which is a way to grow a thin layer of material, for example the silicon dioxide, on silicon wafer with very high temperature.

Figure 1.4 depicts the traditional process monitoring and control system through the SPC charts system: both pre-metrology and post-metrology are fed into the SPC system, and western electric rules [17] are used to distinguish common cause variations and special cause variations. The common cause variations are natural patterns which are expected to occur during normal operations, while the special cause variations are unusual patterns or nonquantifiable variations. The SPC chart can help detect special cause variations and it then triggers the reaction mechanism when the chart is OOC or trending rules are violated. The reaction mechanisms include tool maintenance, recipe change, sampling more products and so on. The

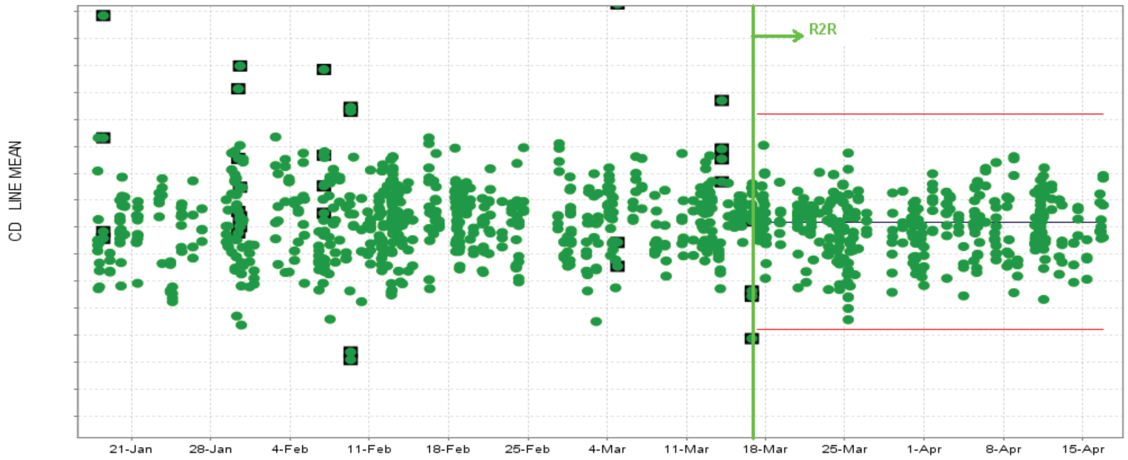


Figure 1.3. CD process capability is improved by 40% with a R2R controller

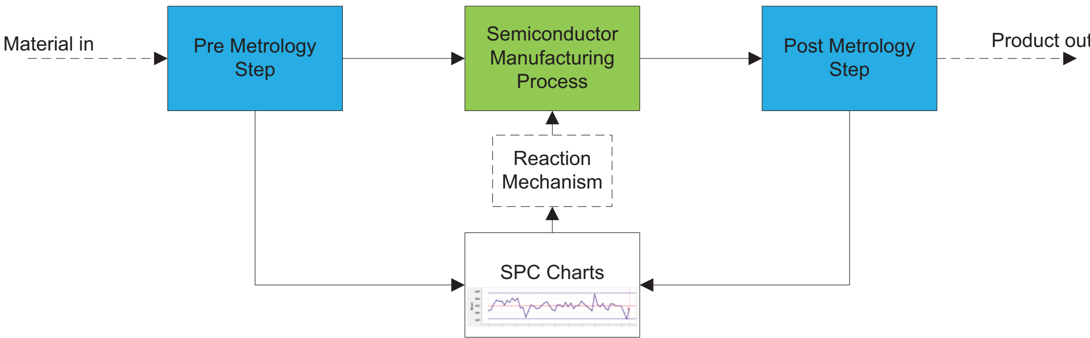


Figure 1.4. Process monitoring and control through SPC Chart

last data point in Figure 1.1 is OOS (out of spec) and in this situation, some action needs to be taken to react to such an event according to the reaction mechanism.

On top of SPC, FD and R2R controls, we propose that a virtual metrology (VM) system can be integrated into the PCS for both process monitoring and control (refer to Figure 1.5). In semiconductor manufacturing, virtual metrology refers to methods to predict properties of a wafer based on machine parameters and sensor data from the production equipment and pre-metrology, without performing the costly physical measurement of the wafer properties [25]. This new architecture integrates VM and other PCS components (SPC, FD and R2R) altogether. First, VM collects data from FD, metrology and other data source in the process, and then the predicted metrology data can be fed into the R2R controller for either feed-forward or feedback control. Finally, the predicted metrology data and their reliance index (RI) can be saved in SPC charts in either FD or SPC for process monitoring. In summary, VM interacts with every single component of the PCS system and it improves the process control capability by introducing a "predictive mechanism."

1.2 Introduction to R2R Control

In recent years, R2R control has been thoroughly adopted in semiconductor manufacturing. R2R control is defined as "a form of discrete process and machine control in which the product recipe with respect to a particular process is modified ex situ, i.e., between machine runs, so as to minimize process drift, shift, and variability" [26]. A framework of fab-wide R2R control was proposed in Figure 1.6 [1]. Currently, most of the control loops in the framework have been achieved

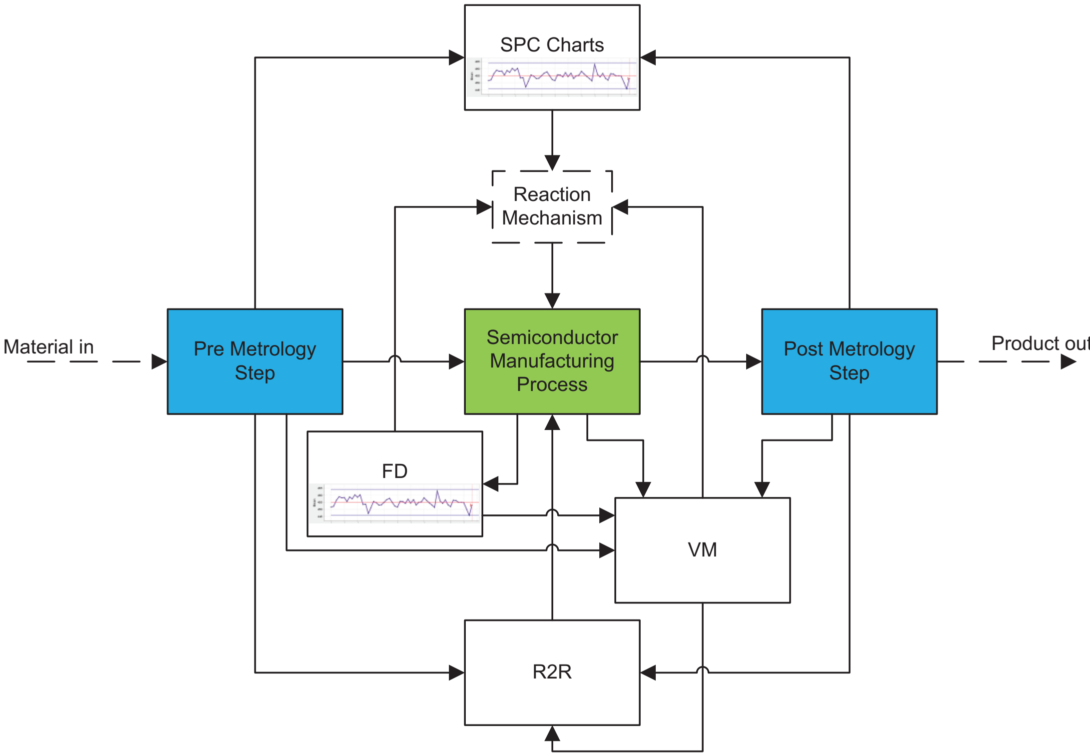


Figure 1.5. A new architecture for process monitoring and control

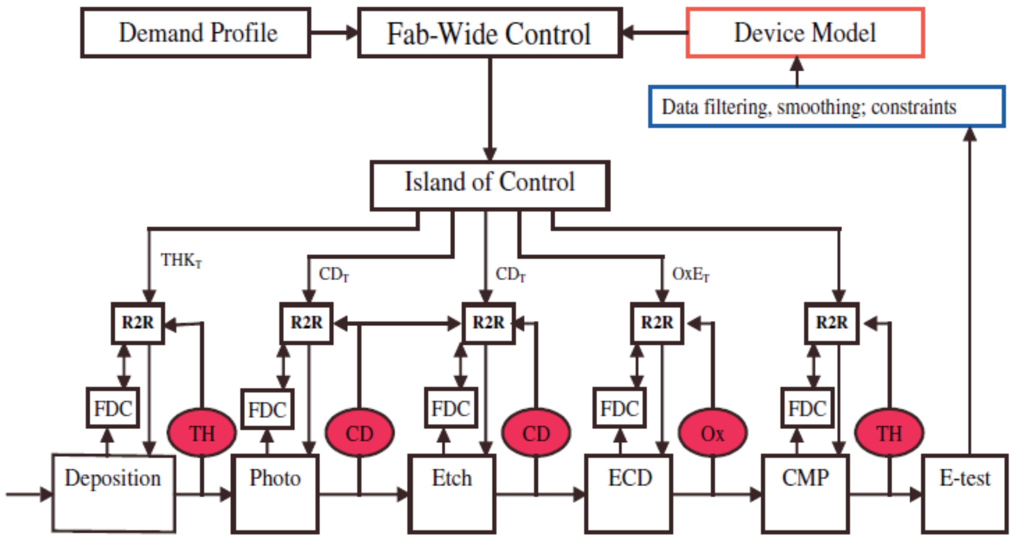


Figure 1.6. A framework of fab-wide R2R control [1].

in the production. Deposition processes, such as chemical vapor deposition (CVD) and physical vapor deposition (PVD) or diffusion, often use thickness metrology as output, and deposition time or deposition temperature as inputs (or tuning knobs). The controller mode could be feedback only, except that some PVD controllers utilize target life as feed-forward components. Photo R2R control often selects critical dimension (CD) and Registration (REG) as outputs, and it is very likely a feedback only control system, which adjusts either the CD dose or alignment REG parameters. On the other hand, etch, electrical-chemical deposition (ECD) and chemical-mechanical polishing (CMP) control systems can be more complicated. R2R control systems on these processes often involve feed-forward, feedback and multiple inputs and multiple outputs.

There are two commonly used R2R algorithms [27], exponential weighted moving average (EWMA) filter-based R2R control and model predictive control (MPC). EWMA-based R2R control earned its popularity due to its stability and easy tuning, but MPC has earned a good reputation because it can handle MIMO systems and constraints. We will have a quick introduction of these two algorithms next.

1.2.1 R2R Control Algorithms

The first EWMA-based R2R control system was developed on an epitaxial silicon deposition system [28]. A simple process model can be described in the linear regression form,

$$y_k = mu_k + b_k \tag{1.4}$$

where y_k is the output, m is the process gain (the slope of the model), u_k is the manipulated variable and b_k is the bias (intercept) of the model.

In the case that the process gain m is assumed a fixed value, where m could be obtained through a priori design of experiment (DOE) [29], the intercept state b_k can be estimated via an EWMA filter providing new metrology data are obtained:

$$\hat{b}_{k+1} = \lambda(y_m - mu_k) + (1-\lambda)\hat{b}_k \tag{1.5}$$

where the value of the weight λ, a value between 0 and 1, is selected by tuning, and the typical value of λ is between 0.2 and 0.3. A smaller λ value is often chosen

for relatively high metrology noise or model noise, while a higher λ value should be chosen otherwise. The two different model update methodologies for R2R control are shown in Figure 1.7. The graph on the left illustrates that the model is updated through estimating a new intercept b_{k+1}: the dotted line is the represented process model before the shift, with the old intercept state, and the interception with the target determines the recipe before the shift. After a tool or process shifts, a new intercept state can be estimated through the EWMA filter, $\triangle b$ represents the magnitude of the shift for the intercept or "$b_{k+1} - b_k$". The solid line represents the R2R model after the shift, and a new recipe setting can be calculated through intercepting a new R2R model of b_{k+1} with the target:

$$u_{k+1} = \frac{y_t - \hat{b}_{k+1}}{m} \tag{1.6}$$

where y_t is the control target.

An alternative option is to update process gain m_k, as seen in the graph on the right of Figure 1.7: instead of modulating the intercept state, we estimate the new process gain m_{k+1} through the EWMA filter, assuming intercept b is fixed:

$$\hat{m}_{k+1} = \lambda \frac{(y_k - b)}{u_k} + (1 - \lambda)\hat{m}_k \tag{1.7}$$

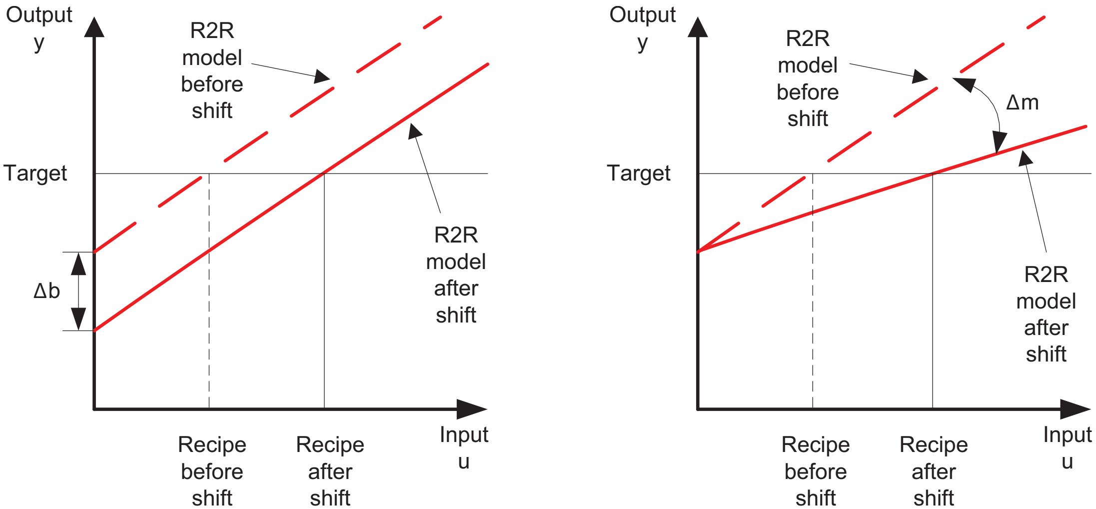


Figure 1.7. Two model update methodologies for R2R control: model update through estimating intercept (left graph); model update through estimating slope (right graph)

A new recipe setting can be calculated through intercepting a new R2R model of m_{k+1} with the target,

$$u_{k+1} = \frac{y_t - b}{\hat{m}_{k+1}} \tag{1.8}$$

Generally speaking, either estimating intercept state or estimating process gain often produces similar results, but in certain situations, estimating intercept state can outperform estimating process gain, and vice versa. This will be further discussed in the non-threaded R2R control chapter.

MPC has been well adopted in many industries, such as at petrochemical plants [30], but for semiconductor manufacturing, MPC is relatively new. A linear state space model [27,31] can be expressed as,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + F\omega_k \\ y_k &= Cx_k + Du_k + \nu_k \end{aligned} \tag{1.9}$$

where x_k is process state, A is state (or system) matrix, B is input matrix, C is output matrix and D is feed-forward matrix of the general state space representation. F is the state noise matrix. ω_k and ν_k are state noise and measurement noise, respectively.

The simple input-output model in Equation (1.4) can be easily transformed into the format of Equation (1.9) through two methods. For the first method, one can define $x_k = b_k$, then the equation below can be obtained,

$$\begin{aligned} b_{k+1} &= b_k + \omega_k \\ y_k &= b_k + mu_k + \nu_k \end{aligned} \tag{1.10}$$

In the state space form [32,33], we can simply define $A = 1$, $B = 0$, $F = 1$, $C = 1$ and $D = m$.

An alternative state space transformation of Equation (1.4) can be obtained by assumption of $u_{k+1} = u_k$ and defining state vector x_k as,

$$x_k = \begin{bmatrix} mu_k & b_k \end{bmatrix}^T \tag{1.11}$$

and in the state space form [34,35],

$$\begin{aligned} \begin{bmatrix} mu_{k+1} \\ b_{k+1} \end{bmatrix} &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} mu_k \\ b_k \end{bmatrix} + \begin{bmatrix} m \\ 0 \end{bmatrix} u_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \omega_k \\ y_k &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} mu_k \\ b_k \end{bmatrix} + \nu_k \end{aligned} \tag{1.12}$$

and comparing it with Equation (1.9), one can get $A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} m \\ 0 \end{bmatrix}$, $F = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 1 \end{bmatrix}$ and $D = 0$. The second transformation method seems to be better because feed-forward matrix, D, is zero.

The manipulated variable profile, u^N, can be calculated through minimization of the objective function [27] below,

$$\min_{u^N}\left(J = \sum_{j=0}^{\infty}(y_{k+j} - y^t)' Q (y_{k+j} - y^t) + \triangle u'_{k+j} R \triangle u_{k+j}\right) \tag{1.13}$$

subject to below constraints:

$$u_{min} \leq u_{k+j} \leq u_{max}$$
$$y_{min} \leq y_{k+j} \leq y_{max}$$
$$\triangle u_{min} \leq \triangle u_{k+j} \leq \triangle u_{max}$$

where N is the prediction horizon, Q and R are symmetric positive definite weighting matrices, and $\triangle u_i = u_i - u_{default}$, $u_{default}$ is default recommended manipulated variable. For those R2R controllers used in a semiconductor manufacturing environment, we usually set $N = 1$ and $Q >> R$. Therefore, it is simply a dead-beat controller [36] if without any constraint.

The state estimation of can be done by quadratic programming [34,37] through the objective function below:

$$\min_{\omega_k, \nu_k}\left(J = \sum_{k=-1}^{N-1} \omega'_k Q \omega_k + \sum_{k=0}^{N} \nu'_k R \nu_k\right) \tag{1.14}$$

subject to the constraints below:

$$x_0 = \bar{x}_0 + w_{-1}$$
$$x_{k+1} = A x_k + B u_i + \omega_k$$
$$y_k = C x_k + \nu_k$$
$$\omega_{min} \leq w_k \leq \omega_{max}$$

Q and R are configurable weighting matrices used as part of the optimization cost function, analogous to λ in EWMA control. The optimization function allows states to be calculated such that state noise ω_k and measurement noise ν_k are minimized

while remaining within established model constraints. The states can be estimated based on historical run data using the control model and the state estimation optimization Equation (1.14), where N is horizon length, $\bar{x}_0$ is the initial state and ω_{min} and ω_{max} are the lower and upper bounds of the state noise.

The objective function J can be transformed into the form below through algebraic manipulation [34]:

$$\begin{aligned} J &= \min_{W_k} \frac{1}{2} W_k^T H W_k + f^T W_k \\ W_{k,min} &\leq W_k \leq W_{k,max} \end{aligned} \quad (1.15)$$

where W_k is the state error vector, H is a function of Q's and R's in the moving horizon and f is a function of y's, u's and R's in the moving horizon. W_k's can be obtained after solving this objective function though quadratic programming and W_k is defined as,

$$W_k = \begin{bmatrix} \omega_{k-N} & \omega_{k-N+1} & \cdots & \omega_k \end{bmatrix}^T \quad (1.16)$$

where N is the horizon length.

Alternatively, the states can be estimated via Kalman filter [27,38,39], the linear observer is defined as

$$\hat{x}_{k+1|k} = Ax_k + Bu_k + L(y_k - C\hat{x}_{k|k-1}) \quad (1.17)$$

The observer gain L is known as Kalman gain, which can be computed with following steady state Riccati equations,

$$\begin{aligned} P &= A[P - PC^T(CPC^T + R_v)^{-1}CP]A^T + FQ_\omega F^T \\ L &= APC^T(CPC^T + R_v)^{-1} \end{aligned} \quad (1.18)$$

where Q_ω is the covariance of state noise, R_v is the covariance of measurement noise and F is state noise matrix. The disadvantages of such state estimation using Kalman filter would be that it cannot handle constraints as well as the moving horizon.

1.2.2 Threaded and Non-threaded R2R Control

Most R2R controllers are designed to function when the set of contexts, such as process chamber and device, remain fixed (threaded) from run to run, and

wafers with similar process history are assumed to have similar characteristic. Each context is segregated from other context groups; for example, a wafer was processed in the same chamber and same device. The control thread separates each of the states into a unique and single disturbance for the R2R model. In Table 1.1, we showed that the contexts of three chambers and two devices define a total of 6 threads in the threaded R2R controller and the intercept state is tracked by each defined thread.

Referring to Equation (1.4), the intercept state b_k can be assumed as the lump sum of all contexts biases or contributions. The benefit of this is that within each thread, we do not need to worry about how the individual context state changes, nor the non-linear interactions among those context bias contributions [40]. While the problem is that it often causes metrology dilution problem or insufficient data for certain threads, one can imagine what would happen if there were more than 20 chambers in the example of Table 1.1. It is well known that threaded controllers often have difficulty with low-volume products interjected into high-volume manufacturing of a typical device. The problem is exacerbated for a Fab with a high mix of products, in which case it may be expected to have only one or two lots of a given low-volume device started each week. In such circumstances, the threaded controller will produce corrective actions based on metrology information obtained several days ago when an identical thread was last run. Depending on the process, if a chamber drifts faster than metrology feedback with the same contexts, the threaded R2R control will likely fail to perform satisfactorily during the low-volume product run as a consequence of a drift that has not been captured by the metrology.

Table 1.1. Control thread definition example: contexts of three chambers and two devices

Chamber	Device	Control Thread	State
A	1	$A1$	b_{A1}
A	2	$A2$	b_{A2}
B	1	$B1$	b_{B1}
B	2	$B2$	b_{B2}
C	1	$C1$	b_{C1}
C	2	$C2$	b_{C2}

To address such problems, there is an increasing interest in designing non-threaded controllers [41–43] that share information between different control threads and loops. In such non-threaded implementation of the R2R control, the metrology information is shared between production runs for high-volume and low-volume devices. By using the information from all runs, a better control for both high-volume and low-volume devices may be expected, with the most significant benefits likely realized for infrequent threads [44].

The key difficulty in implementing non-threaded R2R control is establishing the association of the measured deviations of the device properties from reference values with the run contexts. For example, every context of the run potentially contributes to the intercept term b_k,

$$y_k = mu_k + (b_k^{Chamber(i)} + b_k^{Device(j)}) \tag{1.19}$$

where $b_k^{Chamber(i)}$ and $b_k^{Device(j)}$ are context contributions from some chamber and device respectively. A more generic form can be expressed as following,

$$b_k = \sum_i b_k^{Context_i} = C_{A,k} b_k^{Context} \tag{1.20}$$

where $b_k^{Context_i}$ is the single intercept bias contribution from every context group, $C_{A,k}$ is the context row vector and $b_k^{Context}$ is a column vector which contains the individual intercept state corresponding to each context.

Clearly, any number of combinations of individual context contributions can add up to the same value of b_k. Such nonuniqueness implies that it is not possible to obtain unique values of context intercepts for the given metrology measurements and the corresponding control inputs used during the kth run. In essence, the contribution of the individual contexts to b_k are not observable [45]. The other assumption of this method is the linear interactions among contexts state contributions, while the interactions among contexts may not be linear in certain circumstances.

It has been proved that $C_{A,k}$ is always rank deficient [41] and

$$\tau = c - 1 \tag{1.21}$$

where τ is number of unobservable state, and c is number of context categories. Let's define $C_{A,k}$ as one row of context matrix C_A for multiple runs. C_A is still unobservable [40]. However, if we denote a new state z_k, which is the linear combination of x_k:

$$z_k = C_{A,k} x_k \tag{1.22}$$

We have proved that z_k is always observable, and the non-threaded R2R recommendations of input depends on z_k not x_k. Therefore, this is the foundation of our non-threaded R2R approach, which will be discussed in detail in the non-threaded R2R control chapter.

1.3 R2R Control and Virtual Metrology

Besides the non-threaded R2R controller, VM is one of the new techniques to address the high demand of metrology operations by R2R controllers. It was proposed that VM data would be fed into the wafer to wafer (W2W) controller [2] in Figure 1.8: the metrology of every wafer is predicted by VM module, the predicted output $\hat{y}_k$ and actual output y_k can be used in the feedback loop of a W2W controller. We know that most R2R controllers in production are lot to lot (L2L) based, because measuring all wafers is very costly in terms of cycle time and metrology capacity [20]. With the help of the VM system, all wafers can be potentially controlled by a W2W controller.

Besides the W2W VM-R2R feedback applications, the VM-R2R applications below would also be very useful:

- **Feed-forward application** [46,47]. Wafer level predictions (VM data) can be fed forward to the wafer level R2R controller at downstream process steps, for example dry etch step, so that all wafers in a lot can be compensated to improve process control performance.

- **Lot level feedback application**. In most lot level feedback R2R controllers, only a small portion of the lots are sampled because of cycle time reduction and costly metrology operations. With the help of VM, 100% of the lots' metrology data becomes available, either actual metrology or predicted

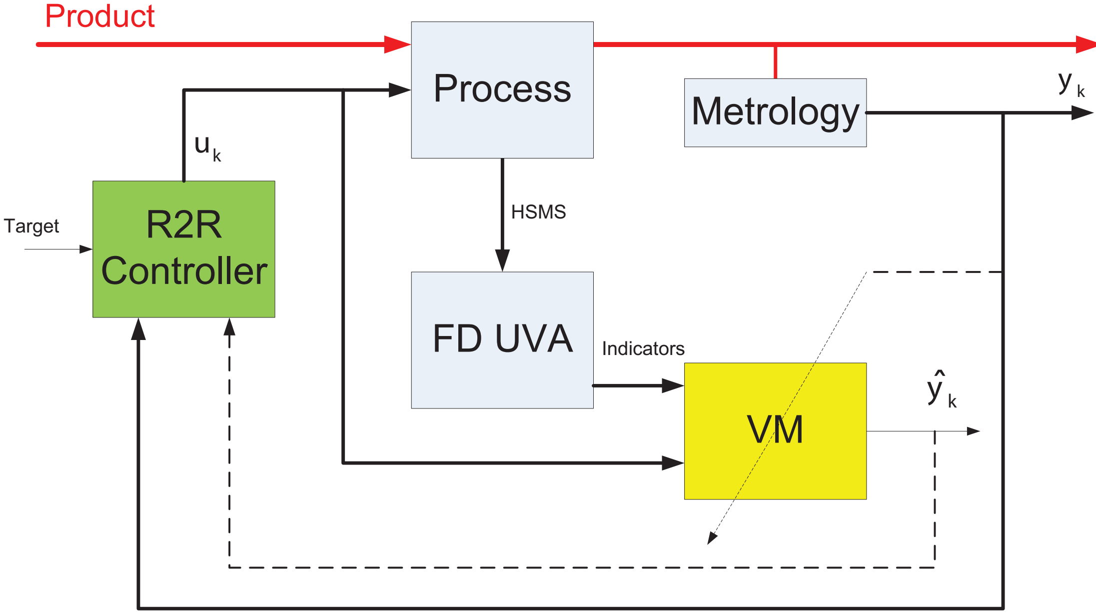


Figure 1.8. VM and W2W R2R control [2]

metrology. Such lot level VM data can be used in the feedback loops of R2R controllers for Cpk improvement.

- **R2R model update.** It's well known that the R2R control model can drift over time, often referred to as process gain m, and such plant and model mismatch problem can cause the downgrade of R2R control performance. In this research, process gain is estimated from time to time through multiphysics-based model in the VM module, while the intercept state, b_k, is still updated upon new metrology data. Therefore, a better process control can be realized.

1.4 Introduction to Virtual Metrology

VM is considered to be a system that predicts metrology data without physically measuring. The outputs of VM can be the CD or thickness of a wafer, and the inputs are typically the process trace data such as pressure, temperature, chemical or gas flows and so on. The predictions (outputs of VM) can be used in the SPC for process monitoring or, if it's accurate enough, it can be used in R2R for controlling the processes. Historically, the algorithms for VM are typically statistical methods such as regression and classification [48, 49]. Before introducing the new method

in this research, let's quickly review the statistical regression methods used for VM models.

1.4.1 Principle Component Analysis

In semiconductor manufacturing, there are a large number of measured variables (FD data or SPC data) that change over time. These variables, however, usually change in a highly correlated manner due to the underlying physical or chemical principles. These independently varying components that are indirectly observed through the measured variables are called latent variables (LV). LVs resemble state variables (e.g., intercept states in R2R control theory) because both uniquely determine the state of the system, they both are not directly measurable but they are observable through metrology data and finally, both are not uniquely defined. However, there are major differences between state variables and latent variables: the state variables are often used in dynamic systems and LVs are normally used in the steady state; also the number of state variables is usually larger than the number of measured variables, but the number of latent variables is usually less than the number of measured variables. Principle component analysis (PCA) is widely used in chemical engineering [49].

Let $x \in \Re^M$ denote a sample vector (or one wafer) of FD data with M sensors and assume that there are N wafers, which results in a data matrix $X \in \Re^{N \times M}$, and the data matrix can be decomposed into M column vectors or N row vectors as is shown in the following,

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_M \end{bmatrix} = \begin{bmatrix} x^T(1) \\ x^T(2) \\ \vdots \\ x^T(N) \end{bmatrix} \tag{1.23}$$

and the sample mean and sample standard deviation can be computed as below, and each column can be normalized by them.

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} x_{ij} \tag{1.24}$$

$$s_i = \sqrt{\frac{1}{N-1} \sum_{j=1}^{N} (x_{ij} - \bar{x}_i)^2} \tag{1.25}$$

where x_{ij} is the ij^{th} element of matrix X.

Principal component analysis (PCA) extracts the direction of the largest variance in the m dimensional measurement space. The data decomposition can be expressed as follows,

$$X = t_1 p_1^T + t_2 p_2^T + \cdots + t_k p_k^T + \tilde{X} = TP^T + \tilde{T}\tilde{P}^T \tag{1.26}$$

where t_1 and p_1 are the score and the loading of the first principal component, t_2 and p_2 are the score and the loading of the second principal component and so on. $\tilde{X} = \tilde{T}\tilde{P}^T$ is the residual matrix, $T = \begin{bmatrix} t_1 & t_2 & \cdots & t_k \end{bmatrix}$ and $P = \begin{bmatrix} p_1 & p_2 & \cdots & p_k \end{bmatrix}$.

Figure 1.9 [3] is used to illustrate what score t and loading p mean exactly. The principle component is the best fit of all 6 data points shown in Figure 1.9 B. Only one principle component is shown for ease of visualization. The loading has two vectors, p_1 and p_2, which are the direction of cosines. The score vector in this case has six components, and each component corresponds to the projection of the data point on the principal component line.

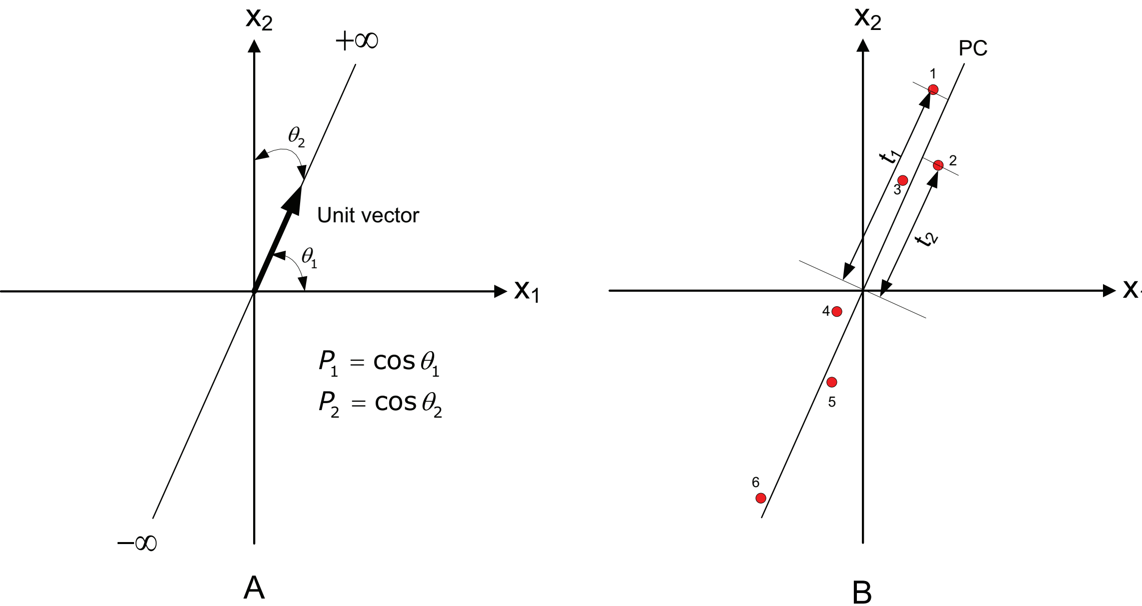


Figure 1.9. A principle component in the case of two variables: A. The loading for the principle component, also the direction of principle component B. The scores of the principle component, which are the projections of the sample points (1-6) on the principle component direction [3]

Perhaps one can notice that the principle component loadings $p = \cos(\theta_1)^2 + \cos(\theta_2)^2 = \cos(\theta_1)^2 + \sin(\theta_1)^2 = 1$, and this property can be extended to more dimensions, for arbitrary direction $p \in \Re^M$ with $||p|| = p^T p = 1$. The projection of X onto this direction line is,

$$t = Xp \tag{1.27}$$

The objective of PCA is to maximize the variance along the loadings, which can be written as,

$$\max t^T t = (Xp)^T Xp = p^T X^T Xp \tag{1.28}$$

and subject to

$$p^T p = 1 \tag{1.29}$$

Applying the Lagrange multiplier, the above objective function can be rewritten as,

$$J = p^T X^T Xp + \lambda(1 - p^T p) \tag{1.30}$$

and the solution of maximizing of this objective function is,

$$X^T Xp = \lambda p \tag{1.31}$$

$$\lambda = p^T X^T Xp \tag{1.32}$$

Both nonlinear iterative partial least squares (NIPALS) [50] and singular value decomposition (SVD) [49] algorithms can be used to decompose the matrix X into P and T. The procedure of NIPALS is as follows:

1. Scale X to zero mean and unit variance and set $X_i = X$

2. Choose a starting point of t_i as some columns of X_i, and iterate below relations defined in the following equation until t_i converges or the maximum number of iterations is reached.

$$p_i = X_i^T t_i / ||X_i^T t_i|| \tag{1.33}$$

$$t_i = X_i p_i \tag{1.34}$$

3. Compute residue

$$X_{i+1} = X_i - t_i p_i^T \tag{1.35}$$

4. Iterate through all factors by setting $i = i + 1$

There are some important properties of PCA, where PCA loadings and scores are orthogonal to each other:

$$\begin{aligned} p_i^T p_j &= 0 \\ t_i^T t_j &= 0 \end{aligned} \tag{1.36}$$

for $i \neq j$.

1.4.2 Partial Least Squares

Partial least squares (PLS) is a widely used algorithm for predictions, especially for VM [51,52]. It would be useful to understand ordinary least squares (OLS) [53] first before we disucss PLS. For n wafers FD input data X and metrology output data Y, the data matrices can be described as,

$$X = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \end{bmatrix}^T \in \Re^{N \times M} \tag{1.37}$$

$$Y = \begin{bmatrix} y(1) & y(2) & \cdots & y(N) \end{bmatrix}^T \in \Re^{N \times M} \tag{1.38}$$

A linear process model is assumed, then

$$\begin{aligned} Y &= X\theta + V \\ y(k) &= \theta^T x(k) + v(k) \end{aligned} \tag{1.39}$$

and the OLS solution of the process gain is,

$$\theta_{ls} = (X^T X)^{-1} X^T Y \tag{1.40}$$

While in many semiconductor processes, lots of process trace (or FD data) data items can be highly correlated or collinear. The collinearity may come from underlying physical or chemical correlations due to material or energy balances and restricted variability among the process variables required by safety or tuning knob constraints. In any of these cases, the OLS solution is ill-conditioned due to high collinearity among data items. Fortunately, the PLS algorithm could be the solution by virtue of the fact that the principle components are not correlated and

the regression coefficients are not correlated either, as we discussed in the PCA section earlier.

Similar to PCA, PLS decomposes input X and output Y data matrices into principal components,

$$\begin{aligned} X &= tp^T + E \\ Y &= uq^T + F \end{aligned} \tag{1.41}$$

where t and u are scores, p and q are scores, and E and F are residuals. $t = Xw$ and $u = Yq$ and the loadings w and q are subject to the below constraints,

$$\begin{aligned} ||w|| &= 1 \\ ||q|| &= 1 \end{aligned} \tag{1.42}$$

The difference between PLS and PCA is that PLS not only minimizes the residuals, E and F, but also tries to maximize the correlation between the scores [54–56], u and t. This objective can be expressed as,

$$J = \max(t^T u) \tag{1.43}$$

Applying Lagrange multipliers, the objective function J can be rewritten as,

$$J = \max\{w^T X^T Y q + \frac{1}{2}\lambda_w(1 - w^T w) + \frac{1}{2}\lambda_q(1 - q^T q)\} \tag{1.44}$$

Taking derivatives with respect to w and q,

$$\begin{aligned} \frac{\partial J}{\partial w} &= X^T Y q - \lambda_w w = 0 \\ \frac{\partial J}{\partial q} &= Y^T X w - \lambda_q q = 0 \end{aligned} \tag{1.45}$$

The solution of the PLS objective function leads to eigenvectors and eigenvalues as below,

$$\begin{aligned} X^T Y Y^T X w &= \lambda_w \lambda_q w \\ Y^T X X^T Y q &= \lambda_q \lambda_w q \end{aligned} \tag{1.46}$$

where the loadings w and q are the eigenvectors of $X^T Y Y^T X$ and $Y^T X X^T Y$, respectively.

It is easy to show that λ_q and λ_w have to be the largest eigenvalues associated with the eigenvectors w and q, and

$$\lambda_q = \lambda_w \tag{1.47}$$

The data input and output matrices, scores and loading, and their dimensions are shown in Figure 1.10 [48] and the PLS algorithm with NIPALS [4] is listed below:

1. Outer regression: get a starting vector u_i from some column of Y_i (if Y is single column, then $u_i = Y$), and iterate the following equations until t_i converges or a maximum number of iterations is reached.

$$w_i = X_i^T u_i / ||X_i^T u_i|| \tag{1.48}$$

$$t_i = X_i w_i \tag{1.49}$$

$$q_i = Y_i^T t_i / ||Y_i^T t_i|| \tag{1.50}$$

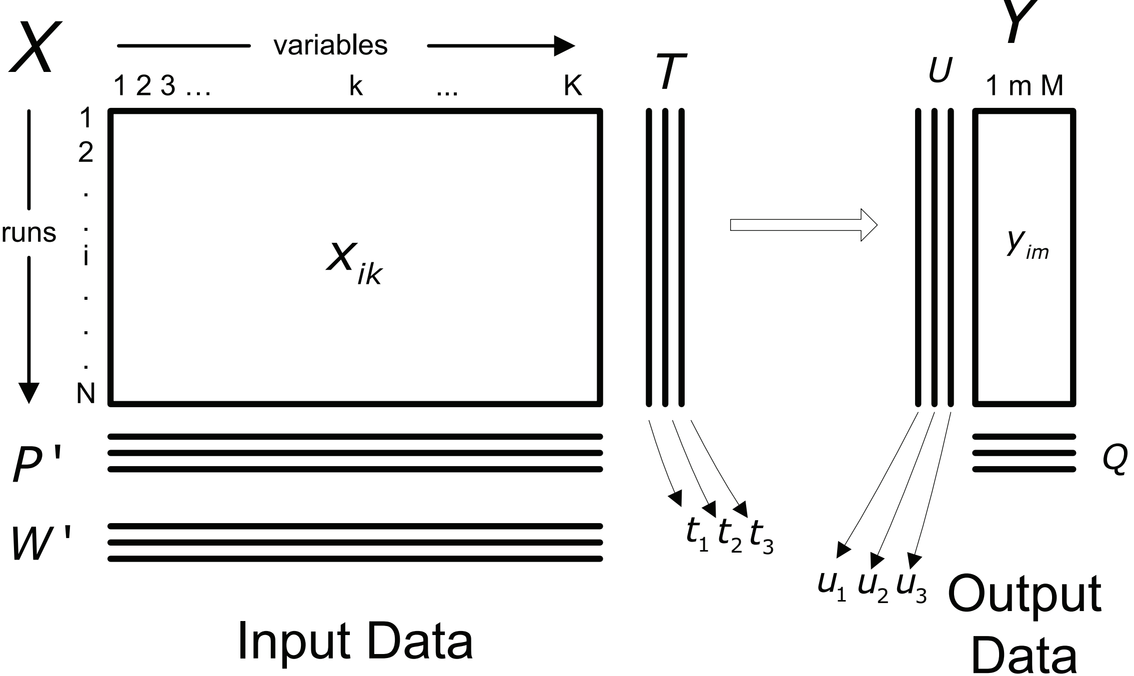


Figure 1.10. Partial least square regression [4]

$$u_i = Y_i q_i \tag{1.51}$$

2. Inner regression: calculate the inner regression coefficient,

$$b_i = u_i^T t_i / t_i^T t_i \tag{1.52}$$

3. Residual deflation: remove the present component from X_i and Y_i and use deflated matrices as the next component X_{i+1} and Y_{i+1}.

$$p_i = X_i^T t_i / t_i^T t_i \tag{1.53}$$

$$X_{i+1} = X_i - t_i p_i^T \tag{1.54}$$

$$Y_{i+1} = Y_i - \hat{u}_i q_i^T = Y_i - b_i t_i q_i^T \tag{1.55}$$

also noting that the inner model estimate $\hat{u}_i$ replaces the Y scores u_i to give the prediction of Y from X.

4. Set $i = i + 1$ and return to the second step until $i = i_{max}$ or there is no more significant information in X about Y.

1.4.3 Neural Networks

Since PLS deals with linear models only, neural networks (NN) would be very useful for nonlinear VM models. Back-propagation neural network (BPNN) has been used to predict CVD thickness [57–59], and other neural network methods, including piecewise linear neural networks (PLNN), fuzzy neural networks (FNN), simple recurrent neural networks (SRNN) and radial basis function neural networks (RBFN) have been also tested for VM modeling [5,60,61].

The following is a quick introduction to BPNN [62–64]. A typical neural network consists of an input layer X, a hidden layer Z and an output layer Y, as shown in Figure 1.11, and each layer has multiple components or units. For instance, there are four inputs, two hidden layer units and three outputs in the case of Figure 1.11.

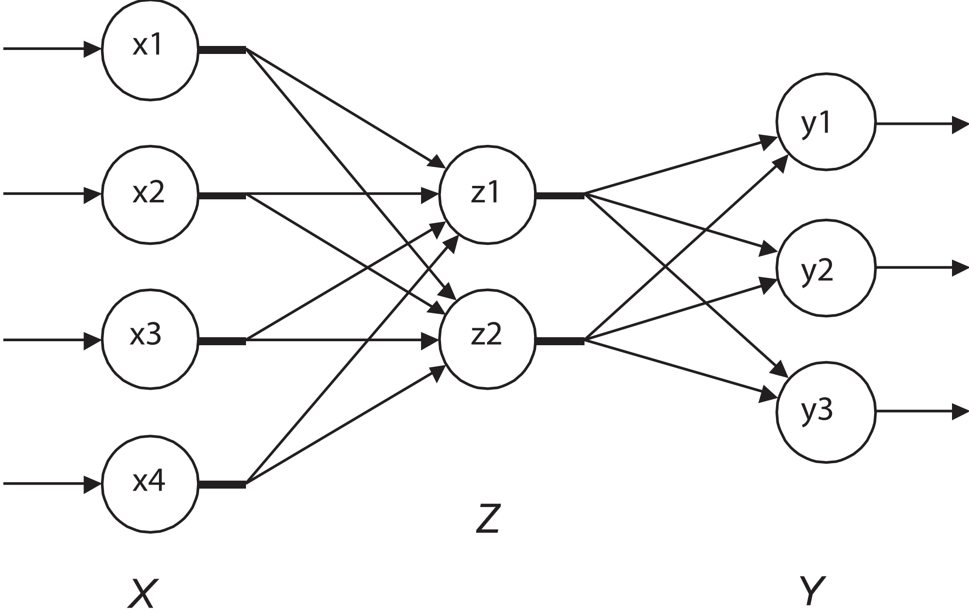


Figure 1.11. Back-propagation neural network architecture

The feed-forward process involves transferring an input to input layer neurons which passes the input values into the hidden layer. Each of the hidden layer nodes computes a weighted sum of its inputs, passes the sum through its activation function and the output of activation function presents the result to the output layer.

The neuron uses the following transfer function,

$$p_i = \sum_{i=1}^{n} x_i w_i \tag{1.56}$$

where w_i is the weight of an input, and n is the number of inputs.

The output of a hidden unit $z_i = 1$ when $p_i \geq \theta$, while $z_i = 0$ if $p_i < \theta$. θ is called the threshold of activation function, again for mathematical convenience, we can set

$$\sum_{i=1}^{n} x_i w_i - \theta = 0 \tag{1.57}$$

The activation function is often chosen to be the Sigmoid function by virtue of the fact that the derivative of the output with respect to input is a function of output only.

$$\beta = \frac{1}{1 + e^{-\alpha}} \tag{1.58}$$

$$\frac{d\beta}{d\alpha} = \beta(1-\beta) \tag{1.59}$$

The BPNN algorithm is to adjust the weights to minimize the difference, mean square error (MSE), between desired output (or target) and actual output, and it uses supervised learning, meaning that it uses training data for which both inputs and desired outputs are known. The network weights will be set or frozen after the network has been trained.

For the simplest BPNN in Figure 1.12, for demonstration purposes only, it has only one input, one hidden unit and one output. It is easy to show that,

$$\begin{aligned} p_1 &= w_1 x \\ p_2 &= w_2 z \end{aligned} \tag{1.60}$$

where $z = sigmoid(p_1)$ and $y = sigmoid(p_2)$.

The performance function J is introduced to minimize the distance between the output y and the desired output d,

$$J = -\frac{1}{2}(d-y)^2 \tag{1.61}$$

Note that the constant coefficient, $-\frac{1}{2}$, is picked just for mathematical convenience.

The name of back-propagation can be explained by the fact that w_2 is always updated first after obtaining new output data y (or J), and then w_1 can be updated with new w_2 through the following steps:

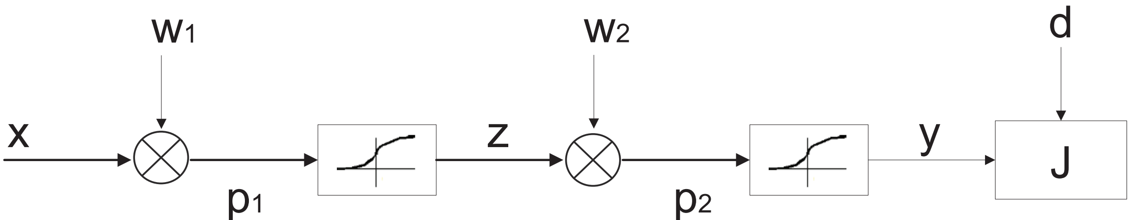


Figure 1.12. The simplest neural network: x is the input, z is the hidden layer, y is the output, d is the desired output, w_1 and w_2 are the weighting factors and J is the objective function or performance function.

- Update w_2 first:

$$w_2(k+1) = w_2(k) + \Delta w_2(k) = w_2(k) + \gamma \frac{\partial J}{\partial w_2} \tag{1.62}$$

where γ is the learning weight, which is a user defined value.

$$\frac{\partial J}{\partial w_2} = (d-y)y(1-y)z \tag{1.63}$$

- Update w_1 later with a new w_2:

$$w_1(k+1) = w_1(k) + \Delta w_1(k) = w_1(k) + \gamma \frac{\partial J}{\partial w_1} \tag{1.64}$$

$$\frac{\partial J}{\partial w_1} = x(d-y)y(1-y)w_2(1-z)z \tag{1.65}$$

Above is a quick discussion for feed-forward neural network using Sigmoid functions, trained by the error backpropagation algorithm. The basic idea of BPNN is gradient descent and BPNN limitations are discussed in the literature [65].

CHAPTER 2

SCOPE OF THIS WORK

In last chapter, we reviewed all components in the process control system of semiconductor manufacturing and their interactions. In this research, we focus on two of them, non-threaded R2R control and virtual metrology, both of which are relatively new and have drawn a lot of attention from process control engineers and researchers in recent years.

2.1 Motivations and Objectives

Metrology data are very costly in semiconductor manufacturing in terms of cycle time and metrology tool cost. In a high mix production environment, the traditional threaded R2R controller likely fails to function due to metrology dilution problems as we discussed in the introductory chapter. The motivation of this research is to develop new non-threaded R2R control algorithms through sharing information among different control threads in a high mix production without imposing a metrology penalty. Two fundamental problems of non-threaded R2R controllers, observability and computational cost, have prevented non-threaded R2R control from being widely deployed [66]. First, a reliable R2R solution becomes vital, because excursions of R2R control can be very costly in semiconductor manufacturing. Much past research attempted to address this problem through either model reduction or by imposing additional constraints on the non-threaded R2R model; in essence, the system needs to be observable. In addition, these methods, to some extent, increase the computational cost and make non-threaded control impractical in the production environment. The other motivation of this research is to solve these practical problems which would enable factory-wide deployment of non-threaded R2R control systems.

On the other hand, in the VM research, we use physical and chemical reaction models in building virtual metrology systems. This approach is substantially different from current approaches of purely empirical regression modeling, which identifies correlations based exclusively on process trace data or FD data. Most of the published VM papers were built using PLS, Kalman filter or Neuron Network methods. Lately, the PLS algorithm seems to be the most dominant approach in building VM systems. The problem of these statistical methods is that predicted metrology is often not accurate enough to be used by the R2R controller at critical process steps, so we want to incorporate multiphysics knowledge into VM modeling to improve the quality of prediction. One of our motivations of this research is to feed virtual metrology data into R2R controllers to improve process capabilities at critical process steps. Furthermore, we want to benchmark physical and chemical reaction modeling with statistical modeling, such as PLS or Kalman filtering, and identify the pros and cons of the two methods. Although VM has been proposed to be used in W2W R2R controllers, with the current metrology and manufacturing execution system (MES) scheme in semiconductor manufacturing, W2W control is still impossible, except for some tool types with onboard metrology systems. Currently, MES with "semiconductor equipment and materials international" standard (or SEMI standard) does not support W2W R2R control either. The predictions in our research projects are etch rate of diluted HF solutions, thickness profile of a diffusion furnace and etch rate in an O_2 plasma resist descum process. Although we cannot achieve W2W R2R control in this research, we still can reduce the etch rate qualification frequency on test wafers and provide high-quality predicted etch rate data to the batch-to-batch (B2B) R2R control system, which improves process capability and process monitoring on production wafers. In the diffusion VM research, the thickness of every wafer in a batch is predicted, where it is potentially used in a wafer level feed-forward controller in dry etch (DE) area to reduce wafer level variations. Finally, the etch rate monitoring of every product wafer is proposed in the O_2 plasma resist descum project. Although VM prediction cannot be used in a R2R controller in the resist descum case, it would be very useful to prevent process or equipment excursions

through wafer level etch rate monitoring.

2.2 Overview of This Dissertation

2.2.1 Hybrid Non-Threaded R2R Control

In Chapter 3, we begin with a problem statement of non-threaded R2R control, in which there is a need for non-threaded R2R to solve metrology dilution problems. However, we encounter three practical problems associated with non-threaded R2R controllers. First, the unobservable states in the control system can cause bias in the state estimation of a non-threaded R2R control system. The second challenge is that the matrix size changes when new tools and new devices are added into production, so the R2R model dimension changes also. Finally, the computational cost of non-threaded R2R is very high; for instance, it takes a lot of time to accomplish the states estimation.

After we reviewed several approaches to solve the observability problem in the literature, we propose a hybrid non-threaded R2R methodology to solve this observability problem by downgrading the controller mode from non-threaded R2R control to threaded R2R control. The business rules define the downgrade criteria, which is used to prevent biased non-threaded R2R recommended settings from downloading into production tools. We present two state space model representations, and explain how to handle matrix size changes when new context items are added. Limiting number of states and horizon length is discussed for long state estimation execution, and we also propose a load balance design, parallel computing, to speed up the computation of the next recommendation settings for all context combinations. In the demonstration section, two real non-threaded implementations and their results in high volume production Fab are presented.

Our new contributions include the following:

- We propose a novel approach to the design of non-threaded controllers without imposing additional constraints or complexity to ensure observability.
- This solution has been implemented and successfully deployed at the real production environment, at IM Flash, and has demonstrated such advantages as high reliability and ease for maintenance.

- A new frame work of auto-tuning non-threaded R2R control is also proposed.

2.2.2 Etch Rate Prediction of Diluted HF Solution

In Chapter 4, we demonstrate that our new methods, incorporating physics and chemical reaction model into VM, can be used to improve prediction of the etch rate of silicon oxide in a diluted HF solution. Compared with traditional linear regression models, this is a better method to select key process variables and it is a better method to build meaningful process indicators. Multiphysics models also require less training data than traditional approaches. We demonstrate that the prediction results are better than those that can be obtained using traditional regression methods. Since feeding VM data into R2R controllers is one of the important ways to materialize VM's benefits, in this research, we have also integrated virtual metrology and R2R control in the local oxidation of silicon (LOCOS) process module in a real production environment. The benefits we achieved include excursion prevention, process capability improvement, yield and cycle time improvement and cost reduction.

Our new contributions and conclusions to this virtual modeling work include:

- **Demonstrated the usage of multiphysics models to select process variables for VM models.** For example, the temperature factor can be removed or included for the VM model and we justified why it can be removed in the model, using process or chemical reaction knowledge.

- **Built meaningful process indicators.** We used the ratio of two chemical flows as prediction indicators in this work. Traditionally, only mean and standard deviation of process trace data are used in the VM model.

- **Accounted for the incoming batch variations from chemicals (e.g., HF 49% concentrations or gases).** Without taking account of those incoming materials, VM prediction accuracy would be compromised or downgraded.

- **Created a multiphysical model requiring less training data compared with statistical model.** For example, 500 wafers data was used for training VM

model [67], but the multiphysics model we developed only needs as few as 3 or 4 data samples to train the VM model.

2.2.3 Thickness Profile Prediction of Diffusion Furnace

In Chapter 5, we first introduce the diffusion furnace process and equipment, and then we elaborate on the generic R2R controller of the diffusion furnace. Although R2R controls flatten the thickness profile to some extent, the thickness profile cannot be completely removed. Next, we incorporate physics insights, equipment knowledge and design of experiment to develop a new multiphysic model, which consists of five Gaussian curves and one intercept term, to predict the thickness profile of a furnace. The model parameters can be updated via the latest metrology data. On the other hand, we encounter some challenges on building a reliable furnace VM model, such as queue time effect of metrology tools without nitrogen purge, which impacts the accuracy of actual thickness measurements. Finally, excellent results are obtained using such multiphysical models after overcoming those challenges. In Section 5.7, we propose to use VM prediction data at downstream process steps, like the dry etch step or the ion implant step.

Our new contributions and conclusions to this virtual metrology project include:

- Described the physics insights of R2R control and the VM system of the diffusion furnace.
- Transformed the parameter estimation problem to a state estimation in the state space representation of the VM model.
- Converted a complicated nonlinear system to a linear system via reasonable assumptions.
- Identified VM project difficulties such as queue time related problems.
- Extended diffusion VM usage to other processes, including dry etch and ion implant.

2.2.4 Etch Rate of Resist Descum

In Chapter 6, we introduce the plasma dry etch processes and chemical reactions of O_2 descum processes. Several etch rate models in the literature are also reviewed. In the next section, key model parameters are selected using the background knowledge of process and chemical reactions. A PLS regression model is tested, while the "Zonal" data analysis is proposed to improve the prediction quality on top of the traditional PLS regression using in-line data. Model update is another important aspect of constructing a reliable VM system. Three model update methods are also simulated and compared. Furthermore, along with a new process indicator, gas ratio, is created by using extensive chemical reaction knowledge, the multiphysics-based model is constructed and the prediction result is improved. Finally, some challenges with unknown incoming variations are discussed and future work is proposed.

The new contributions of this work include the following:

- Proposed a new "Zonal" data analysis, which is a potential method to obtain accurate process gains without a costly DOE.
- Created new process indicators using deep chemical reactions knowledge.
- Simulated and compared three different model update methods of the PLS regression method.
- Constructed a more complicated multiphysics-based VM model compared with etch rate prediction in the wet etch project.

CHAPTER 3

HYBRID NON-THREADED RUN-TO-RUN CONTROL

3.1 Abstract

Although a non-threaded Run-to-Run (R2R) controller has many advantages, such as sharing metrology information among different control threads and improved low-volume products process control performance, model-based non-threaded R2R controllers often have practical issues such as unobservability and high computational costs of state estimation. In this paper, we propose solutions to these practical issues. Such non-threaded R2R control solutions have been implemented at the high volume production fabrication plant, IM Flash. Both process capability and OOC (out of control) events across all products, including both high volume and low volume products, were significantly improved by such non-threaded R2R control implementation.

3.2 Introduction

Semiconductor devices are produced at fabrication plants (Fabs) in a sequence of many hundreds of batch processing steps, many of which require nanoscale precision. A batch consists of one or more lots of wafers, each lot typically containing 25 wafers. Batches are serially processed by such steps as lithography, dry and wet etch, chemical vapor deposition (CVD) and diffusion. Fabrication steps impact critical dimensions, electrical and – more generally – material properties, feature alignments, interconnects and other properties that often cannot be assessed in real time. Instead, they are measured during separate metrology steps conducted after critical fabrication steps are completed. Metrology characterizes the deviations of the measured properties from the target. During the subsequent runs, the

processing conditions (known as a *recipe*) are adjusted to reduce deviations from the target and maintain each fabrication step in control, thus ensuring high product yield. The function of the Run-to-Run (R2R) controllers is to make the needed recipe adjustments so that the next batch hits the target. Depending on the process and equipment, these adjustments may be implemented for the whole batch, one lot, or even an individual wafer. To calculate the needed correction for the current run k, a simple linear R2R model is often used to describe the relationship between the metrology output y_k and the manipulated variable u_k:

$$y_k = mu_k + gf_k + b_k \tag{3.1}$$

where m is the process gain (the slope of the model); g and f_k are the feed-forward gain and the disturbance introduced during proceeding fabrication steps [68], which together describe the influence of processing history on the current processing step; and b_k is the bias (intercept) of the model.

The block diagram depicted in Figure 3.1 summarizes the operation of the R2R controller. After the completion of kth run of a given fabrication step, the metrology measurements, y_m, of critical dimensions, film thickness, or other parameters are acquired and the model (3.1) is updated. Depending on the process, either the slope m, or the intercept b_k is adjusted. For example, in dry etch (DE) processes, the etch rate is often relatively stable and the process variations are adequately captured

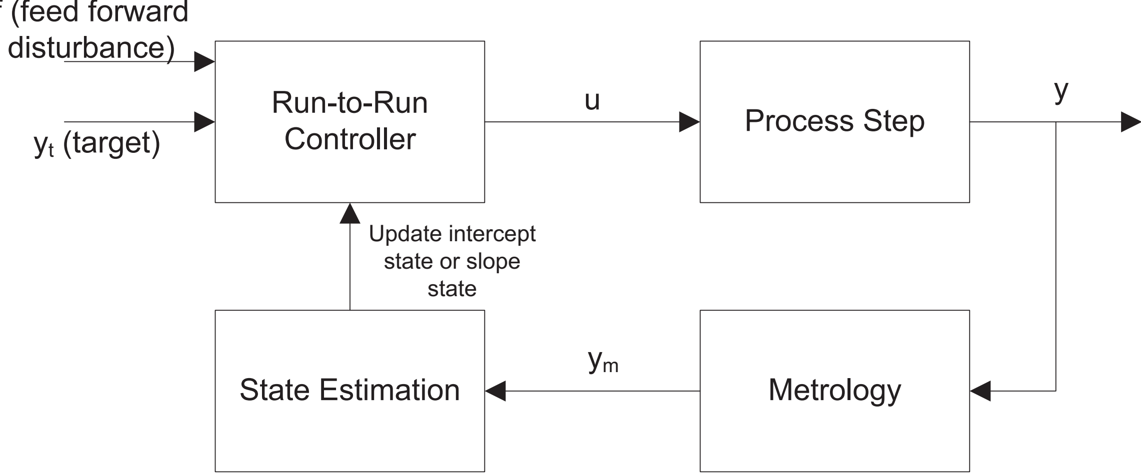


Figure 3.1. Run-to-Run controller.

by a changing bias, with the etch rate m being obtained through the design of experiment (DOE) [17]. To update the bias, the following exponentially-weighted moving average (EWMA) [28,69] filter is often used:

$$b_{k+1} = \lambda(y_m - mu_k - gf_k) + (1-\lambda)b_k \tag{3.2}$$

where the value of the weight λ (between 0 and 1) is selected by tuning.

For other processes, it is more appropriate to update the process gain. A representative example is a CVD process. Since the deposition occurs not only onto the surface of the wafers but also onto the walls of the CVD chamber [70], the rate of the chemical vapor deposition changes with time, which causes the process gain to change. Using the EWMA filter, the updated gain can then be obtained as:

$$m_{k+1} = \lambda \frac{y_m - b - gf_k}{u_k} + (1-\lambda)m_k \tag{3.3}$$

where the values of λ in (3.2) and (3.3) are generally different. The correct choice between updating b or m (both of which are often referred to as *states*) becomes particularly important, when feed-forward disturbances are present and their values are known from metrology measurements conducted prior to the current fabrication step. On the other hand, when $gf_k = 0$ or unknown, then either the intercept or the slope adaptation is often equally acceptable.

After the model is updated based on the available metrology data ("State Estimation" block in Figure 3.1), the R2R controller uses the model predictions to adjust the manipulated variables in order to achieve the target metrology values y_t during the next run. For example, assuming that the metrology data, y_k, were used to adjust the intercept of the model with the gain held constant, the deadbeat R2R controller will select the following value of the manipulated variable during the next run [1,27,71]:

$$u_{k+1} = \frac{y_t - gf_{k+1} - b_{k+1}}{m} \tag{3.4}$$

Alternatively, if the metrology was used to estimate a new value of the process gain, then

$$u_{k+1} = \frac{y_t - gf_{k+1} - b}{m_{k+1}} \tag{3.5}$$

There are many factors that lead to deviation from the metrology targets, including the sequence of processing steps that the wafer has been subjected to, the

specific equipment (known as *tools*) used in the processing sequence, and the type of the product that is being manufactured. Collectively, these factors are known as *contexts*. The sequence of processing steps and tools used to perform them are known as a *thread* associated with a given wafer. R2R control is most effective when the corrective actions are based on the metrology obtained, when all contexts of runs k and $k+1$ are matched, or *threaded*. Threaded controllers are R2R controllers designed to function when all contexts are matched.

In certain Fabs, threaded R2R control strategy is problematic. This includes Fabs with a "high mix" of products, or when low-volume products are produced as an infrequent interjection into a high-volume manufacturing of a typical product [72]. In such situations, there may be only one or two lots of a given low-volume product started each week. Consequently, a threaded controller has to produce corrective actions based on "old" metrology data obtained several days ago when an identical thread was last run. If a tool or a process drifts faster than metrology feedback, the threaded R2R control will likely fail to perform satisfactorily during the low-volume product run. Even in high-volume manufacturing, if the contexts are defined narrowly (e.g., down to the slot position in a diffusion furnace, etc.), a large number of control threads will be present and long metrology delays may occur between threaded runs.

The limitations of the threaded control may be addressed by designing and deploying controllers that share information among different control threads [73]. Such non-threaded R2R controllers use all available metrology measurements to take corrective actions, irrespective of matched or mismatched contexts. By sharing the metrology data between high-volume and low-volume products, a better control may be achieved, with the most significant improvements likely realized during infrequent threads.

Many algorithms implementing the basic idea of non-threaded control have been proposed. The key element in this strategy is the model update based on non-threaded information. This problem, often referred to as the state estimation of m and/or b, was approached in the past by using Kalman filtering [40,41,66,74], the recursive least square (RLS) [75,76], the best linear unbiased estimation (BLUE) [40]

and the dynamic analysis of variance (ANOVA) [77] to estimate non-threaded R2R states. The previously proposed methods also differ in the formalism used to share non-threaded information, which includes such approaches as the just-in-time adaptive disturbance estimation (JADE) algorithm [42] and a random walk model [78,79].

Despite significant efforts invested into the development of non-threaded R2R control technology, it is still common to see spurious results produced by non-threaded controllers that are addressed in the production environment, by limiting the maximum allowed change in the manipulated variable between consecutive runs and other *ad hoc* means. Such control excursions can often be traced to un-observability of a model used in the state estimations, changing model size as new contexts are added to describe a thread and the high computational cost of solving the state estimation problem in real time.

In this paper, we describe a hybrid approach that uses a non-threaded controller under normal circumstances, but reverts to the threaded algorithm mode when certain "business rules" are violated. This approach is the first hybrid control system design implemented in the high-volume production of NAND flash memory products. It showed a significant improvement compared to the standard threaded control and was more reliable than other non-threaded control designs.

This chapter is organized as described here. The next section formulates the non-threaded control problem and highlights practical issues with implementing non-threaded controllers in the production environment. Section 3.4 summarizes several algorithms proposed in the literature and discusses their limitations. The proposed hybrid method is described in Section 3.5 and its application in the high-volume production environment is demonstrated in Section 3.6. The chapter concludes with a discussion of the proposed approach and directions for future improvements.

3.3 Problem Statement

A typical problem in high production mix environments is that often there are many control threads for certain threaded R2R controllers, which causes the

dilution of metrology data and downgrades the process control performance. For example, Figure 3.2 [42] demonstrates this problem, assuming both photo tools and reticles are drifting over time, and every photo tool and reticle combination has to be treated as a separate thread, and $j \times l$ threads are required, where $j(=2)$ is number of photo tools and $l(=2)$ is the number of reticles. The drawback of such a control scheme is metrology dilution. A 100% of lots sampling rate for a device results in an effective sampling rate of only 25%, because there are $j \times l = 4$ threads instead of 2 ($j = 2$) threads, one for each tool. In this case, some of the feedback loops may operate with very long metrology delays, which can cause failure in process control. Therefore, a non-threaded R2R controller is demanded to solve such metrology dilution problems by sharing information among various threads.

Consider a simple linear model with feed-forward disturbance (3.1) and the single intercept state b_k can be separated into relevant contexts based errors as shown in (3.6):

$$y_k = mu_k + gf_k + \sum_i b_k^{Context_i} \tag{3.6}$$

and in matrix form,

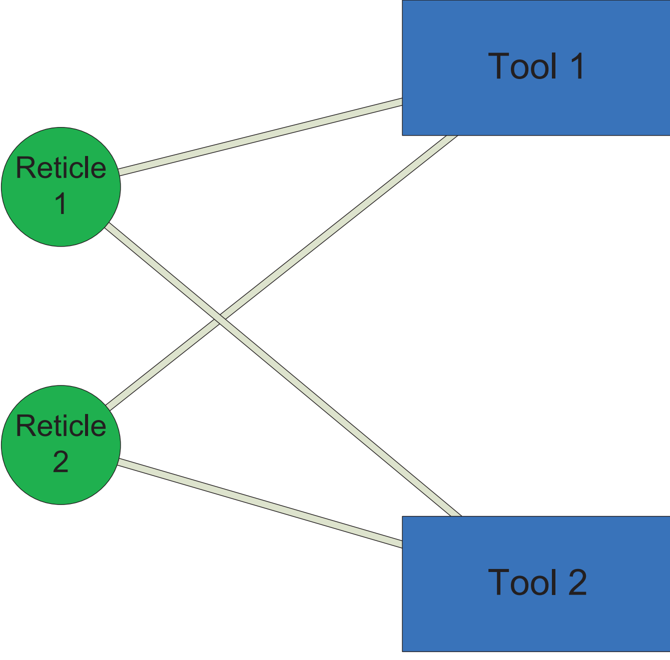


Figure 3.2. Metrology data are diluted due to increased control threads.

$$b_k = \sum_i b_k^{Context_i} = C_{A,k} b_k^{Context} \tag{3.7}$$

where $b_k^{Context_i}$ is the single intercept bias contribution from every context group, $C_{A,k}$ is the context row vector and $b_k^{Context}$ is a column vector, which contains the individual intercept state corresponding to each context.

Using the classic photo example, context matrix $C_{A,k}$ has four columns when there are two tools and two reticles,

$$C_{A,k} = \begin{bmatrix} 1 & 0 & | & 1 & 0 \end{bmatrix} \tag{3.8}$$

where columns 1, 2, 3 and 4 represent Tool 1, Tool 2, Reticle 1 and Reticle 2, respectively. Tool 1 and Reticle 1 were used in the example of (3.8). On the other hand, $b_k^{Context}$ has four elements, one for each context item.

In the more generic form of the photo R2R control example,

$$\sum_i b_k^{Context_i} = b_k^{Tool(j)} + b_k^{Reticle(n)} \tag{3.9}$$

where photo tool and reticle are the related contexts which contribute to the errors of the intercept state. An alternative threaded R2R solution using the same context definitions will result in $j \times n$ control threads, considering j photo tools and n reticles, which often causes insufficient metrology data for some of the feedback loops.

For multiple runs, the context matrix C_A has a dimension of $r \times s$, r is the number of runs and s is the number of contexts. $C_{A,k}$ is one of the rows of C_A.

$$C_A = \begin{bmatrix} C_{A,r} \\ \vdots \\ C_{A,k} \\ \vdots \\ C_{A,2} \\ C_{A,1} \end{bmatrix} \tag{3.10}$$

This assumes one can estimate all intercept contributions of all related contexts. The deadbeat control law is used by a non-threaded R2R controller to calculate the recommended recipe setting of the next lot:

$$u_{k+1} = \frac{y_t - g f_k - C_{A,k+1} \hat{b}_{k+1}^{Context}}{m} \tag{3.11}$$

One of the state space models that can be used to describe such a non-threaded R2R controller is,

$$\begin{aligned} x_{k+1} &= x_k + \omega_k \\ y_k &= C_k x_k + D u_k + \nu_k \end{aligned} \tag{3.12}$$

where $x_k = b_k^{Context}$ is the state vector, y_k is the output, D (= m) is the process gain, u_k is the input, ω_k is the state noise and ν_k is the measurement noise. C_k (= $C_{A,k}$) is the output matrix.

The first problem here is how to estimate the contexts-based states $b_k^{Context}$. The detailed state space design and state estimation methodology of such non-threaded controllers will be outlined in Section 3.5. At this time, we also want to point out three practical issues associated with such non-threaded control systems: the first one is that controller is unobservable because context matrix is always rank deficient [40,41]. The second issue is that state space model sizes are continuously changing (or growing) [41, 66] when new contexts are added, such as new tools or new devices. Finally, the state estimation often involves high computational cost [66]. For example, the state estimation can take a long time.

3.3.1 Bias in State Estimation

The difficulty in implementing a model-based non-threaded R2R control is in establishing the association of the measured deviations of the part properties to reference values with the run contexts. Clearly, any number of combinations of individual context contributions can add up to the same value of b_k. Such nonuniqueness implies that it is impossible to obtain unique values of context intercepts for the given metrology measurements and the corresponding control inputs used during the kth run. In essence, the contributions of individual contexts are not observable. Zheng et al. [80] pointed out that the tool-based approach is unstable when the plant is nonstationary. The product-based control will be stable, but its performance will be inferior to single product control when the drift is significant. Hanish [41] pointed out that the number of unobservable states is one less the number of context categories; Wang [40] proved that the context matrix C_A is always rank deficient, so there is no guarantee that unbiased estimates of each

individual intercept state can be obtained. Three common scenarios can introduce biased state estimations in non-threaded R2R control:

- Addition of new contexts by adding new tools (chambers) or new devices and so on.
- Tool maintenance events
- Not having historical data in the moving horizon

All these scenarios can cause biased individual state estimation of a non-threaded R2R controller. The overall intercept state can also be biased because it is only linear combinations of individual context states (3.7). Consequently, the controller recommended recipe settings calculated by (3.11) can be biased too. Such bad R2R recipe adjustments can increase the risk of excursion for products. The cost of such excursion in a production Fab can be very high.

3.3.2 The Change of Matrices Sizes

Another challenge of a model-based non-threaded R2R control is that the matrix size of $C_{A,k}$ and $b_k^{Context}$ will be changed when new tools are added or new devices are introduced at the production Fab. Using the classic photo example again, when a third reticle is added, context matrix $C_{A,k}$ changes its dimension. Compared with (3.8), it now has five columns as shown below,

$$C_{A,k} = \begin{bmatrix} 0 & 1 & | & 0 & 0 & 1 \end{bmatrix} \tag{3.13}$$

Tool 2 and Reticle 3 were used in example of (3.13). At the same time $b_k^{Context}$ also changes from a four element vector to a five element vector. Such matrix size changes must be considered in the non-threaded R2R design, or the non-threaded R2R control will be broken, whenever a new context item is added. Harirchi et al. [66] proposed modifying the state vector and associate matrices so that the prediction of Kalman filter can continue. In this way, extra steps of the Kalman filter procedure are added, and the computational complexity or cost is significantly increased. In Section 3.5 of this dissertation, we will propose how to address this issue without increasing the computational complexity.

3.3.3 Long Execution Time for State Estimation

The state estimation execution time of a model-based non-threaded R2R can last a relatively long time compared with threaded R2R controls, because of the intensive calculations introduced by the large dimension of contexts matrix and long horizon length. Harirchi et al. [66] analyzed the computational complexity in the case of Kalman Filtering-based non-threaded R2R control.

The state estimation execution duration can take as long as five minutes or more based on our experiments, which will be shown in Section 3.5. In addition, we still have to compute the non-threaded R2R recommended settings for each unique context combination at metrology update time, which can take a lot of time too. These intensive calculations can potentially cause unintended server or database problems. In Section 3.5 of this dissertation, we will propose how to minimize the impact of long execution time of state estimation and how to mitigate the risk of this kind of server event.

3.4 Background

After reviewing non-threaded R2R control literature, we discovered that most of them identify practical issues, which are either the un-observability of the control system or the change of model dimension due to the addition of new context items. Such practical issues are not addressed explicitly in any non-threaded R2R control literature, which can possibly lead to negative impact in real implementation.

3.4.1 Kalman Filtering

When process gain is equal to 1, the state space model of (3.12) becomes:

$$\begin{aligned} x_{k+1} &= x_k + \omega_k \\ y_k &= C_k x_k + u_k + \nu_k \end{aligned} \tag{3.14}$$

For multiple runs in the moving horizon, one can define a new state space system with h measurements in the horizon,

$$\begin{aligned} x_{k+1} &= x_k + \omega_k \\ y_k^a &= \tilde{C}_k x_k + u_k^a + \nu_k^a \end{aligned} \tag{3.15}$$

where y_k^a is the augmented outputs with $\dim(y_k^a) = h \times 1$, u^a is the augmented inputs with $\dim(u_k^a) = h \times 1$, $\tilde{C}_k$ is the contexts matrix and each row of $\tilde{C}_k$ represents the

context of a single run. Also $\tilde{C}_k$ is with α order rank deficient compared with the number of states in x_k.

A reduced order system [41,81] can be obtained via state space transformation, such that the system becomes observable for the transformed system. If we define,

$$\bar{x}_k = T x_k \tag{3.16}$$

then

$$\begin{aligned} y_k^a &= \tilde{C}_k x_k + u_k^a \\ y_k^a &= \tilde{C}_k T^{-1} T x_k + u_k^a \\ y_k^a &= \bar{C}_k \bar{x}_k + u_k^a \end{aligned} \tag{3.17}$$

A transformation matrix T can be found, so that the last α columns of $\bar{C}_k$ (or $\tilde{C}_k T^{-1}$) are zero, so the last α states elements of $\bar{x}_k$ have no effect on y_k^a and only the first few states in $\bar{x}_k$ are to be estimated. Now one can solve the state estimation problem via the Kalman filter [27,82] without any observability problems.

However, the Kalman filter method is not practical for the production environment when the number of states changes [41]. For example, lithography gets a new state with every new reticle. Harirchi et al. [66] also investigated the methods of solving problems introduced by new context items and evaluated the computational cost of the Kalman filter, which are the two practical problems of Kalman filtering based non-threaded R2R implementation.

3.4.2 Jade Algorithm

Firth et al. [42] proposed JADE algorithm, for given context matrix C_A , the total observed bias can be estimated by (3.18),

$$C_A \hat{b}^{Context} = b \tag{3.18}$$

where $\hat{b^{Context}}$ is an $s \times 1$ vector of context-based state estimates, the matrix C_A dimension is $r \times s$ and b is an $r \times 1$ vector; r is the number of runs and s is the number of contexts.

JADE reconstructs context matrix in the sense of EWMA and finds context-based states using the least square solution:

$$\hat{b}^{Context} = (C_A^T C_A)^{-1} C_A^T b \tag{3.19}$$

Using the classic two tools and two reticles photo example of (3.8) when Tool 1 and Reticle 1 were the run context, the unique solution of $\hat{b}^{Context}$ in (3.19) is not guaranteed for a single run, because the number of observations is less than the number of variables being estimated. On the other hand, context matrix C_A is rank deficient, because the separate context items are confounded with one another [40]. One of the extreme examples is that C_A becomes (3.20) for multiple runs, where Tool 2 and Reticle 1 were used for all four runs:

$$C_A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{3.20}$$

The unique solution of $\widehat{b^{Context}}$ in (3.19) is not guaranteed because the rank of C_A is less than the number of unknowns. In order to solve the rank deficient context matrix, JADE modified equation (3.19) into the recursive update of contexts bias contributions (3.21).

$$\begin{bmatrix} C_A \\ I \end{bmatrix} \hat{b}_{k+1}^{Context} = \begin{bmatrix} b \\ \hat{b}_k^{Context} \end{bmatrix} \tag{3.21}$$

where I is an $s \times s$ identity matrix and s is the number of context items, $\widehat{b_k^{Context}}$ is the estimation of context-based state contributions. C_A is truncated to a specified number of rows, q , which is chosen by the user. In this way, only the biases for the context items used in these latest runs are updated. Biases for context items not used remain unchanged. By introducing a weighting matrix Q to assign preference to the current measurement versus past measurement for state contribution estimation, an objective function can be written as (3.23). For simplicity in the notation, define $\Omega = \begin{bmatrix} C_A \\ I \end{bmatrix}$ and $\tilde{b} = \begin{bmatrix} b \\ \hat{b}_k^{Context} \end{bmatrix}$.

$$J(\hat{b}_{k+1}^{Context}) = \frac{1}{2}(\tilde{b} - \Omega \hat{b}_{k+1}^{Context})^T Q (\tilde{b} - \Omega \hat{b}_{k+1}^{Context}) \tag{3.22}$$

Q composes the following submatrices,

$$Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix} \tag{3.23}$$

where Q_2 and Q_3 are matrices of zeros , Q_1 is $q \times q$ matrix and Q_4 is $s \times s$ matrix,

$$Q_1 = diag\begin{bmatrix} \lambda & \lambda^2 & \cdots & \lambda^q \end{bmatrix} \tag{3.24}$$

$$Q_4 = diag\begin{bmatrix} \alpha_1(1-\lambda) & \alpha_2(1-\lambda) & \cdots & \alpha_s(1-\lambda) \end{bmatrix} \tag{3.25}$$

where λ is a tuning parameter to weight current measurement to past state estimates and α_i represents relative weighting between the context-based state estimation.

The objective function (3.23) can be solved using the least square solution procedure; the context-based bias estimation can be obtained,

$$\hat{b}_{k+1}^{Context} = (\Omega^T Q \Omega)^{-1} \Omega^T Q \tilde{b} \tag{3.26}$$

The limitation of JADE is that it needs qualification runs to obtain unique context-based state estimates (or initial state). Without the correct initial states from those qualification runs, the context-based state estimation can be biased.

3.4.3 EWMA Method

Recently, Prado and Feng [83] explored EWMA method again for non-threaded R2R state estimation in a real production environment:

$$y_k = mu_k + b_k^{device} + b_k^{tool} \tag{3.27}$$

$$\hat{b}_{k+1}^{device} = \hat{b}_k^{device} + \lambda_{device}(y_k - mu_k - \hat{b}_k^{device} - \hat{b}_k^{tool}) \tag{3.28}$$

$$\hat{b}_{k+1}^{tool} = \hat{b}_k^{tool} + \lambda_{tool}(y_k - mu_k - \hat{b}_k^{device} - \hat{b}_k^{tool}) \tag{3.29}$$

where λ_{device} is device gain and λ_{tool} is tool gain. $\hat{b}_k^{device}$ and $\hat{b}_k^{tool}$ are device state and tool state, respectively. Prado and Feng used deadband and filter to remove noise before estimating the states. Although the evaluations of Hanish [41] showed that Kalman filter outperforms the EWMA method, he pointed out that the Kalman filter algorithm has a practical problem, which is the changing size of matrices in the production environment. Therefore, the EWMA algorithm was recommended

in the real production environment. The algorithm is based on the EWMA method, which makes it very easy to replicate and understand, but changes to the gain of each context (e.g., λ_{device} or λ_{tool}) are a limitation. It still needs to find a way to update them automatically. Stuber [84] used this method for a DE CD control, and he used an optimizer in Excel and historical data to estimate the initial states for $\hat{b}_o^{device}$ and $\hat{b}_o^{tool}$ which helped to achieve optimal performance faster. Stuber also studied the limitation of the EWMA method post maintenance events. He showed that the shift in a tool state post maintenance events can bump device state estimates and vice versa.

3.4.4 Model Regularization

Patel [43] proposed the model regularization for high product mix control, which enforced a unique solution by adding constraints to ensure that the bias (states) estimates stay centered on 0 for $n-1$ out of n contexts. Only one context state is allowed to track arbitrary drift without any constraint, and the rest of the context states are constrained to be centered around zero.

Let $\tilde{y}_k = y_k - mu_k$ and if one has a set of observations, then these observations can be stacked to get the following:

$$\tilde{Y}_{r-1} = C_{A,r-1} b_{r-1}^{Context} + W_{r-1} \tag{3.30}$$

with

$$Y_{r-1}^{\sim} = \begin{bmatrix} \tilde{y}_0 \\ \tilde{y}_1 \\ \vdots \\ y_{r-1}^{\sim} \end{bmatrix} \tag{3.31}$$

$$W_{r-1} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{r-1} \end{bmatrix} \tag{3.32}$$

where $\tilde{Y}_{r-1}$ is the vector of $\tilde{y}_k$, $C_{A,r-1}$ is context matrix, $b_{r-1}^{Context}$ is context bias vector and W_{r-1} is noise terms.

Patel added another set of equations (or constraints) on top of (3.30):

$$\begin{bmatrix} \tilde{Y_{r-1}} \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} C_{A,r-1} \\ \cdots \\ \eta \end{bmatrix} b_{r-1}^{Context} + \begin{bmatrix} W_{r-1} \\ \cdots \\ \omega \end{bmatrix} \tag{3.33}$$

where η is a matrix which helps enforce the constraints. As an example, consider a system with three contexts groups, with $S_1 = 3$ (meaning three components in context group 1), $S_2 = 2$ and $S_3 = 3$.

$$\eta = \begin{bmatrix} 0 & 0 & 0 & \vdots & 1 & 1 & \vdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \vdots & 0 & 0 & \vdots & 1 & 1 & 1 \end{bmatrix} \tag{3.34}$$

with the above designed constraints, the context biases (or states) can be obtained uniquely. Thereby, Patel presented a solution to the observability problem of non-threaded R2R control.

3.5 Methods

In this section, we first introduce two non-threaded state space models and then we propose some new methods to solve those practical problems related to model-based non-threaded R2R controllers.

3.5.1 Two Model-Based Non-threaded R2R State Space Representations

A single input and single output (SISO) system can be described in equation (3.1), if we define state vector x_k:

$$x_k = \begin{bmatrix} mu_k & b_k \end{bmatrix}^T \tag{3.35}$$

then a state space model [30, 34, 85] can be used to represent such a system,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + F\omega_k \\ y_k &= Cx_k + Gf_k + \nu_k \end{aligned} \tag{3.36}$$

where x_k is the state vector, ω_k is the state noise and ν_k is the measurement noise. A is the state matrix, B is the input matrix, C is the output matrix and G is the feed-forward matrix.

A modified version state space model is designed for non-threaded Run-to-Run control as following:

$$
\begin{aligned}
x_{k+1} &= Ax_k + Bu_k + F\omega_k \\
y_k &= C_k x_k + Gf_k + \nu_k
\end{aligned}
\tag{3.37}
$$

where the output matrix C_k is no longer a constant matrix, and it depends on the run contexts.

Compared with the earlier definition of $C_{A,k}$, C_k is defined as

$$C_k = \begin{bmatrix} 1 & C_{A,k} \end{bmatrix} \tag{3.38}$$

Using the two tools and two reticles example, the intercept state can be separated into different contexts contributions, refer to (3.7). One can have,

$$b_k = x_k^{Tool_i} + x_k^{Reticle_j} \tag{3.39}$$

where $Tool_i$ can be Tool 1 or Tool 2, and $Reticle_j$ can be Reticle 1 or Reticle 2. When the contexts of the current run are Tool 1 and Reticle 1, the output matrix C_k can be obtained,

$$C_k = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{3.40}$$

and the corresponding state vector x_k is,

$$x_k = \begin{bmatrix} mu_k & x_k^{T_1} & x_k^{T_2} & x_k^{R_1} & x_k^{R_2} \end{bmatrix}^T \tag{3.41}$$

Similarly, if contexts of the next run are Tool 2 and Reticle 2, then the output matrix will become,

$$C_k = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.42}$$

One can immediately notice that the output matrix C_k depends on contexts. However, matrices A , B , F and G do not depend on the contexts and they stay as a constant matrix like a standard state space model (3.36). After evaluating the observability of the system (3.37) by computing the observability matrix [81], we concluded that the system (3.37) is not observable.

For an alternative state space model (3.12), let's denote a new state z_k, which is the linear combination of x_k:

$$z_k = C_{A,k} x_k \tag{3.43}$$

$$\begin{aligned} z_{k+1} &= z_k + \omega_k \\ y_k &= z_k + D u_k + \nu_k \end{aligned} \tag{3.44}$$

where D is the process gain. One can prove that system (3.44) is now observable. In fact, we only care about the linear combination of the states, z_k, instead of individual states x_k for non-threaded R2R control, because non-threaded recommendations for inputs of the next run depends on z_k, not x_k:

$$u_{k+1} = \frac{y^t - z_k}{m} \tag{3.45}$$

where y^t is the output target.

3.5.2 The Hybrid Non-threaded R2R Control Methodology

During the non-threaded R2R controller testing phase, we discovered that sometimes the biased z_k's were calculated, because of the issues discussed in Section 3.3. To solve this problem, a new architecture is proposed in Figure 3.3. Both threaded and non-threaded modules are executed at the same time; threaded recommended settings serve as the reference. When non-threaded recommended settings significantly deviate from the reference, according to the business rules defined in the controller, the control mode will be downgraded from non-threaded control to threaded control. Such hybrid non-threaded design is more reliable, because it can prevent biased non-threaded recommended settings from being sent to the manufacturing execution system (MES).

Using the two tools and two reticles example, both threaded and non-threaded estimators provide correct state estimation of z_k's. The threaded estimator does this by using a separate estimator for each thread. For the example with states x_k^{R1}, x_k^{R2}, x_k^{T1}, and x_k^{T2}, the threaded estimation of $z_k^1 = x_k^{R1} + x_k^{T1}$, $z_k^2 = x_k^{R1} + x_k^{T2}$, $z_k^3 = x_k^{R2} + x_k^{T1}$, and $z_k^4 = x_k^{R2} + x_k^{T2}$ is obtained using four separate threaded estimators. For each

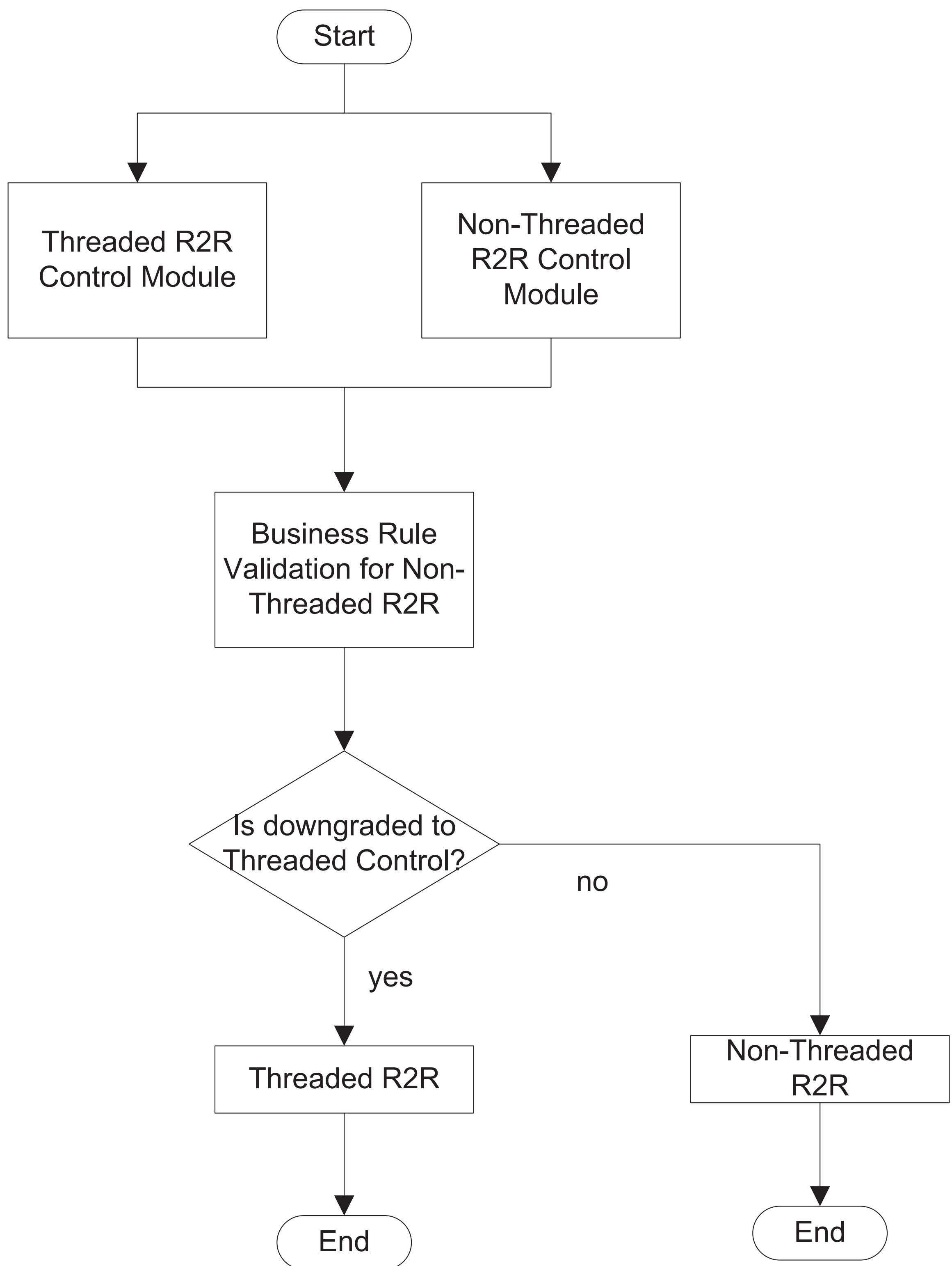


Figure 3.3. Non-threaded R2R control architecture

context defined in threaded R2R control module, the state estimator is described as below:

$$z_{k+1} = \lambda(y_k - mu_k) + (1-\lambda)z_k \quad (3.46)$$

Such full factorial threaded design often results in the risk of metrology dilution. In real implementation, the variability introduced by reticles is ignored, and only two threads are considered, z_k^{T1} and z_k^{T2}, corresponding to Tool 1 and Tool 2. In such implementation, the metrology is less diluted, or the estimate is updated more often, but at the cost of reduced performance, caused by ignoring the contribution of the selected reticle to the measured deviations from the target obtained at the metrology step.

On the other hand, a single non-threaded estimator needs to estimate x_k's. Although we are only interested in z_k's, and only z_k's are guaranteed to converge to correct values. For convenience, z_k's are updated by updating the contributing states, though there is no guarantee or expectation that the updated x_k's give the correct estimate of individual states. For example, after z_k^1 is found, arbitrarily the values of x_k^{T1} and x_k^{R1} are updated to be equal to $x_{k-1}^{T1} + \omega_k^{T1}$ and $x_{k-1}^{R1} + \omega_k^{R1}$, and $\omega_k^{T1} = \omega_k^{R1}$. ω_k's are the state noise. This update of states is done purely for bookkeeping purposes. Such a method provides the estimate of the linear combination of biases of individual contexts that contribute to the overall bias measured by metrology. In our terminology, we call these linear combinations, z_k's, while individual context biases are denoted as states, x_k. No attempt to estimate correct values of x_k's is made. States are estimated so that only their linear combinations are correct.

For non-threaded R2R control, state vector can be estimated by quadratic programming [34, 37] through the objective function below:

$$\min_{\omega_k, \nu_k}\left(J = \sum_{k=-1}^{N-1} \omega_k' Q \omega_k + \sum_{k=0}^{N} \nu_k' R \nu_k\right) \quad (3.47)$$

subject to following constraints:

$$
\begin{aligned}
& x_0 = \bar{x}_0 + w_{-1} \\
& x_{k+1} = A_{k+1}x_k + B_{k+1}u_k + \omega_k \\
& y_k = C_k x_k + \nu_k \\
& \omega_{min} \le w_k \le \omega_{max}
\end{aligned} \tag{3.48}
$$

Q and R are configurable weighting matrices used as part of the optimization cost function, analogous to λ in EWMA control. The optimization function allows states to be calculated in order that state noise ω_k and measurement noise ν_k are minimized, while remaining within established model constraints. The states can be estimated based on historical run data using the control model and the state estimation optimization equations (3.47) (3.48), where N is horizon length, $\bar{x}_0$ is the initial state and ω_{min} and ω_{max} are the lower and upper bounds of the states.

The objective function J can be transformed into the form below through algebraic manipulation [34]:

$$
\begin{aligned}
J = \min_{W_k} \frac{1}{2} W_k^T H W_k + f^T W_k \\
W_{k,min} \le W_k \le W_{k,max}
\end{aligned} \tag{3.49}
$$

where W_k is the state error vector, H is a function of Q's, R's and C's in the moving horizon and f is a function of y's, u's, R's and C's in the moving horizon. W_k's can be obtained after solving this objective function though quadratic programming and W_k is defined as,

$$
W_k = \begin{bmatrix} \omega_{k-h} & \omega_{k-h+1} & \cdots & \omega_k \end{bmatrix}^T \tag{3.50}
$$

where h is the horizon length. Note, that not all states are updated after each metrology, and only the states used in C_k are updated. The estimates of other states are kept at the prior values for the last run.

The linear combinations of states z_k's is observable. The consequence of correctly estimated z_k's is the correct estimation of the difference in states corresponding to the same context. For example, with the correctly estimated z_k^1 and z_k^2, we get the correct value of $x_k^{T1} - x_k^{T2}$ (=$z_k^1 - z_k^2$), and so on. Therefore, the correctly estimated state differences are the consequence of correctly estimated z_k's, not the other way around, and this makes the non-threaded control effective.

Compared with non-threaded state estimation, the threaded estimator is tuned more aggressively to provide rapid conversion. A relatively high λ value is used in our practice in some special cases, either after tool maintenance or at the starting point of a new thread, we increase the λ equal to one. The conversion of the non-threaded estimator is slower than the threaded R2R control, but the update of x_k's is obtained after every metrology run, not just the threaded run.

The business rules define downgrade criteria [44], which can include but are not limited to the following:

- The number of valid records in the moving horizon for a given context is less than certain threshold.
- Adjustment directions between threaded R2R control and non-threaded R2R control are opposite, and the opposite is not only justified by the sign, but also by the magnitude.
- The maximum absolute recommended settings difference between threaded R2Rs and non-threaded R2Rs is greater than certain tolerance.

Figure 3.4 showed the evaluation results of state estimation for both threaded and non-threaded R2R modules: the values of z_k's in the non-threaded module do not converge to the threaded module z_k^{T1} (Tool 1 context) or z_k^{T2} (Tool 2 context) before run 40. In this case, the controller mode is downgraded from non-threaded R2R control to threaded R2R control before run 40, and it then recovers to non-threaded R2R control after run 40.

3.5.3 Handling of Varying Matrices Sizes

The dimension output matrix (3.38) can be changed when adding new reticles, which has been discussed [41,66]. We propose that dummy contexts can be added to solve this issue. For example, four reticles (or more) can be configured in the model, although only two of them are currently used in production. The other two are so called *dummy* contexts reserved for future use. In this case, equations (3.40)

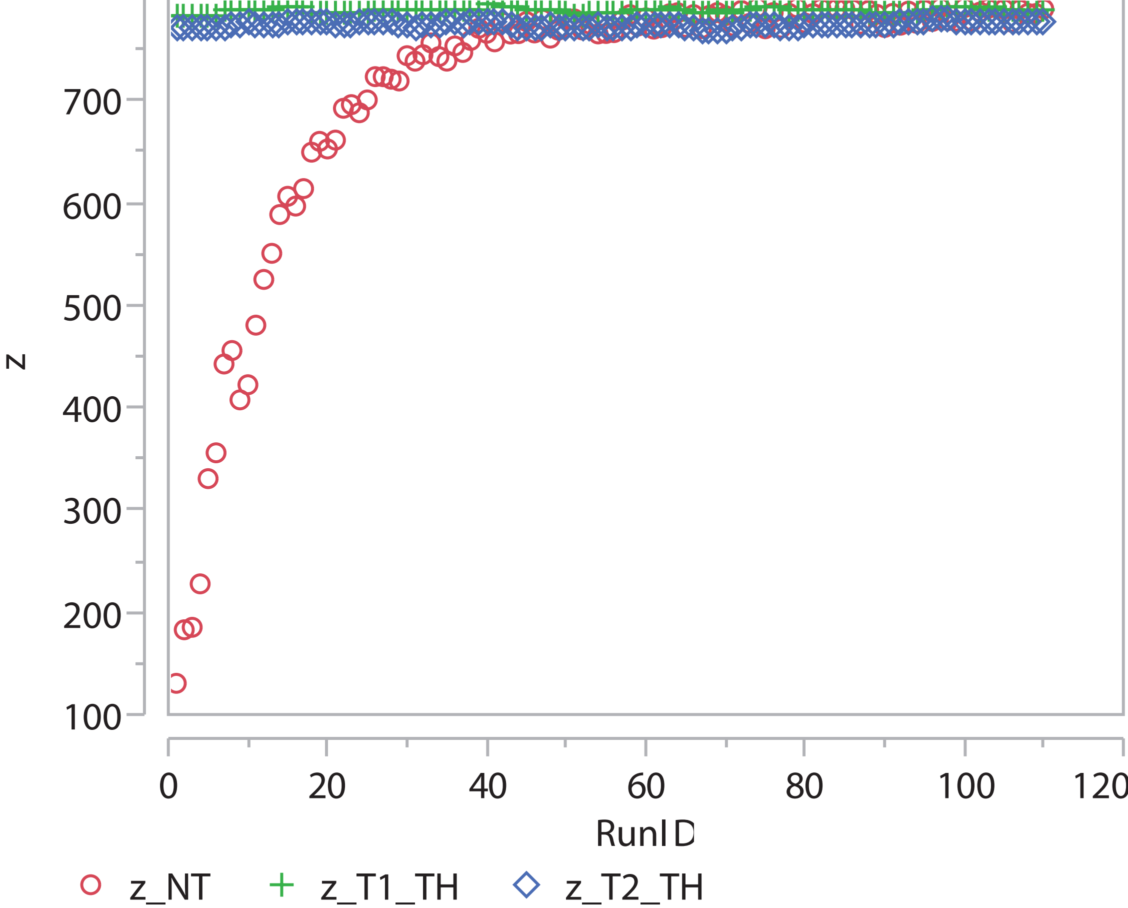


Figure 3.4. Evaluation of state estimation for threaded and non-threaded R2R control

and (3.41) become (3.51) and (3.52); Contexts of current run are Tool 1 and Reticle 1,

$$C_k = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.51}$$

and the corresponding state vector x_k is,

$$x_k = \begin{bmatrix} mu_k & x_k^{T_1} & x_k^{T_2} & x_k^{R_1} & x_k^{R_2} & x_k^{D_1} & x_k^{D_2} \end{bmatrix} \tag{3.52}$$

The last two elements, $x_k^{D_1}$ and $x_k^{D_2}$ in (3.52) are called dummy context states. Similarly, this methodology can be extended to other context groups, for example the tools context. However, there must be a maximum limit for the number of dummy contexts one can reserve (e.g., max number of states is equal to 100). If the state vector has too many elements, more than 100 for example, then the state

estimation execution could be very long in terms of time, which we will discuss in the next section.

3.5.4 Handling of Long Execution Strategy

The state estimation of non-threaded R2R can be very expensive in terms of execution time. The execution time of state estimation is correlated with both number of states and the horizon length, which is shown in Table 3.1. In mass deployment of non-threaded R2R controls, the cost of state estimation has to be considered to prevent R2R server problems and database blocking issues.

Though a larger number of states would normally require a longer horizon to ensure that all linear combinations z_k's are updated, note the following limitations. First, an increase in the horizon leads to higher computational demand, which may prevent computation of the control updates with the required frequency. Second, a longer horizon may result in slower convergence. The longer the horizon, the more average of the historical data (or less forgetting the past or historical data), and therefore it results in less aggressive tuning.

In our non-threaded controller design, we recommended that horizon length be less than or equal to 30 records, and the number of states be less than or equal to 100, due to our system constraints. These numbers can be chosen differently, according to the specific server capacities.

Besides the state estimation, computing non-threaded recommended settings of every context combination can be time consuming, too. A load balancing in

Table 3.1. Execution time of non-threaded state estimation

No. of States	Horizon Length	Execution Time (Seconds)
106	18	74
106	30	220
106	50	240 (or more)
106	100	240 (or more)
94	18	32
94	30	93
94	40	240 (or more)

strategy design is outlined in Figure 3.5. For a 3-chamber DE tool, and assuming that there are valid post-metrology data for every chamber, a web service can be sent for each chamber with valid metrology. Since there are multiple servers in the pool, three different events can be processed by three different servers in parallel. Therefore, a better load balancing of long execution strategy can be achieved.

3.6 Demonstration

3.6.1 Non-threaded Run-to-Run: CMP Mismatch Handling

A CMP mismatch handling system is one of the non-threaded R2R controllers built at IM Flash, referring to Figure 3.6, which controls the recess of trench oxide to form Faraday Blades between (floating gate) FG structures and increasing surface area for gate coupling [86]. The graph on the left is before the DE step and the graph on the right is after the DE step.

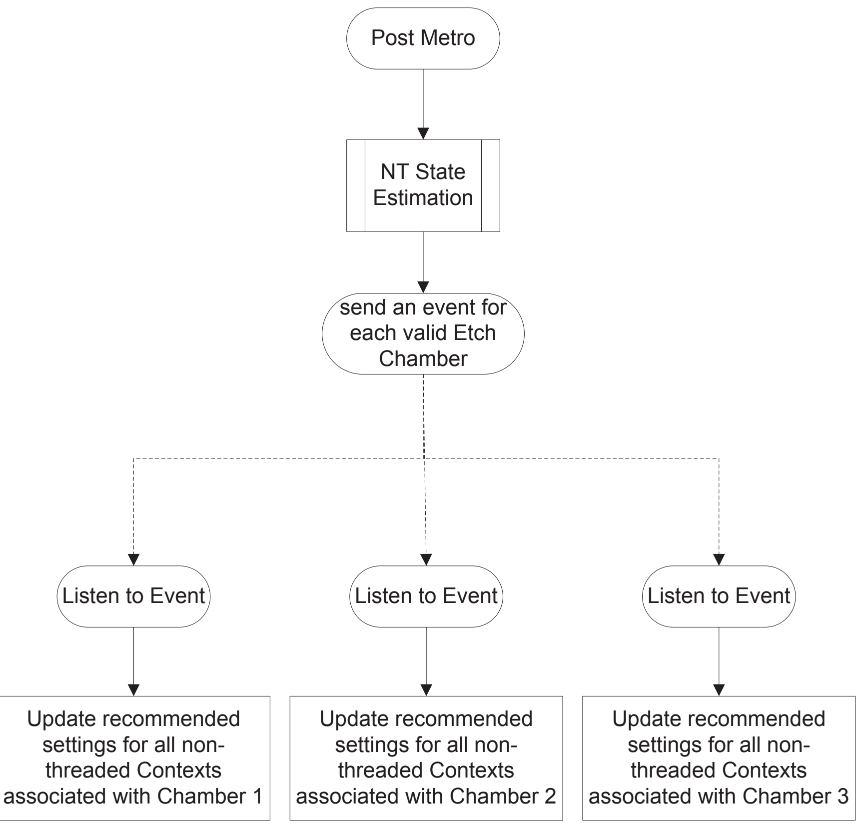


Figure 3.5. Load balancing among different servers

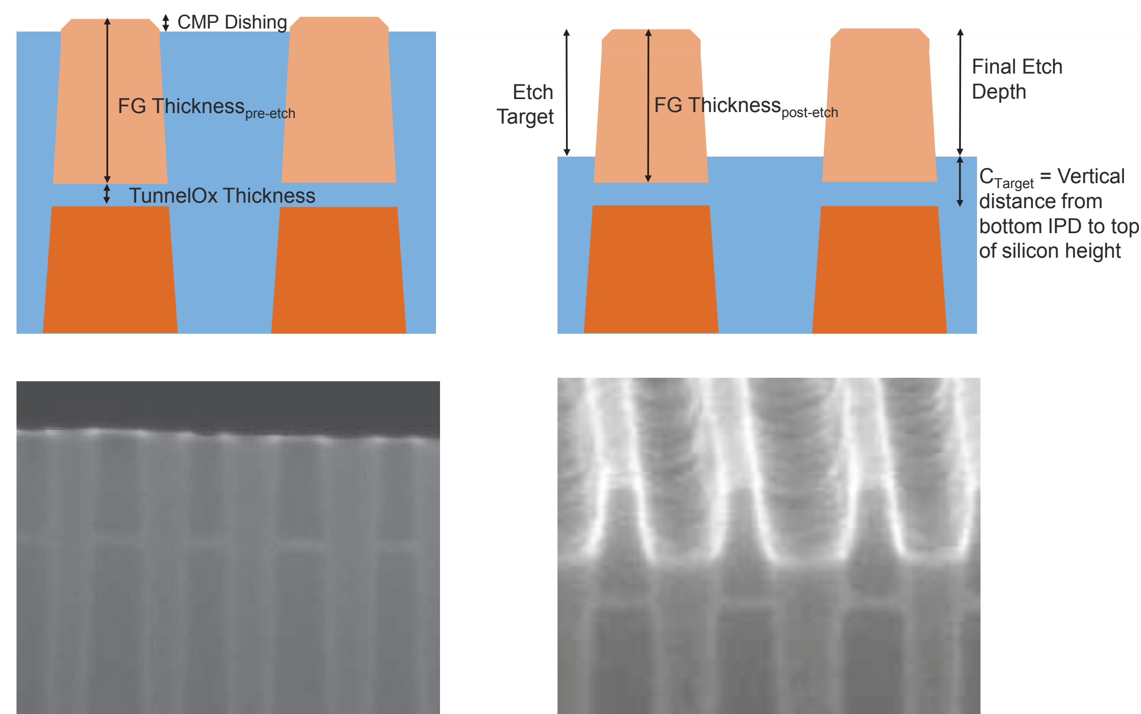


Figure 3.6. Recess oxide before and post dry etch step.

At one time, chemical mechanical planarization (CMP) on-board metrology mismatch was a major problem for the process control in this module. Such mismatch problems were fixed through metrology tool calibration and applying metrology calibration offsets. However, the CMP tool to tool mismatch issue post DE was then discovered, which is shown in Figure 3.7: polysilicon thickness measurements after the DE step have a strong dependency and separation on CMP tools. For example, the wafers processed by CMP Tool 1 tend to have lower poly thickness.

A non-threaded R2R controller was built to compensate for the CMP tool mismatch problem post DE process. The control schematic is shown in Figure 3.8. An alternative solution is to create four different processes in the threaded controller, one process for each CMP tool, but process control performance can be downgraded in this way due to metrology data dilution. The dilution of metrology data occurs because that metrology data is divided into four different CMP tool threads. In the non-threaded design, the intercept state is separated into three different context groups. They are DE chamber states, Device states and CMP tool

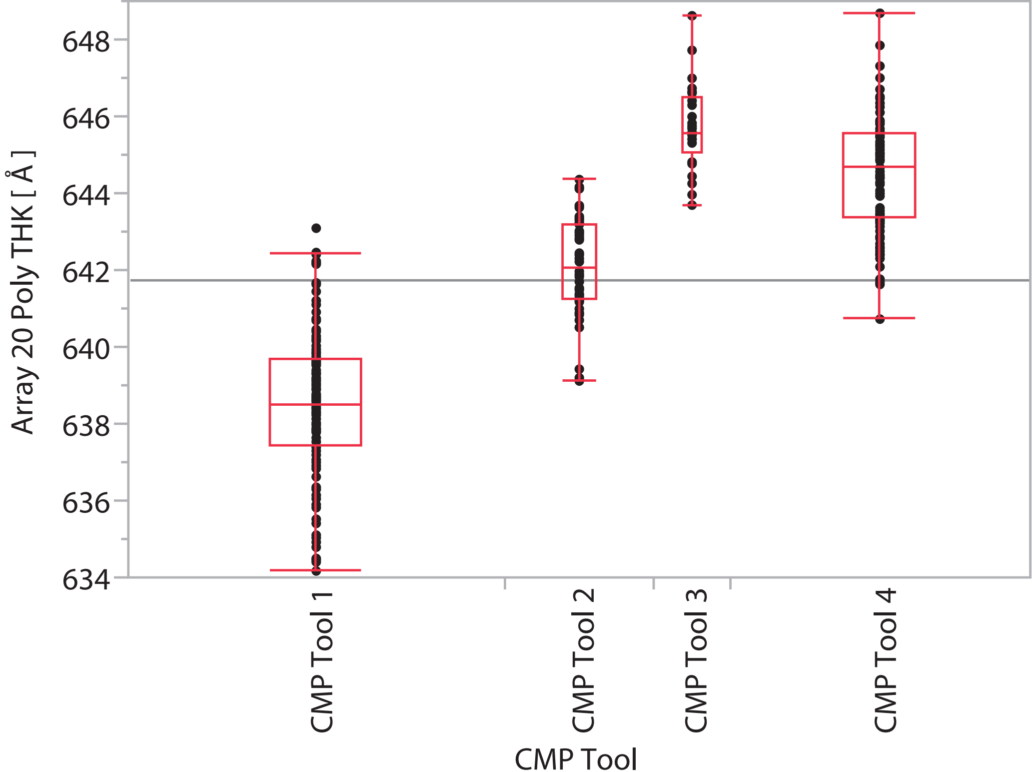


Figure 3.7. CMP tool to tool mismatch post dry etch step.

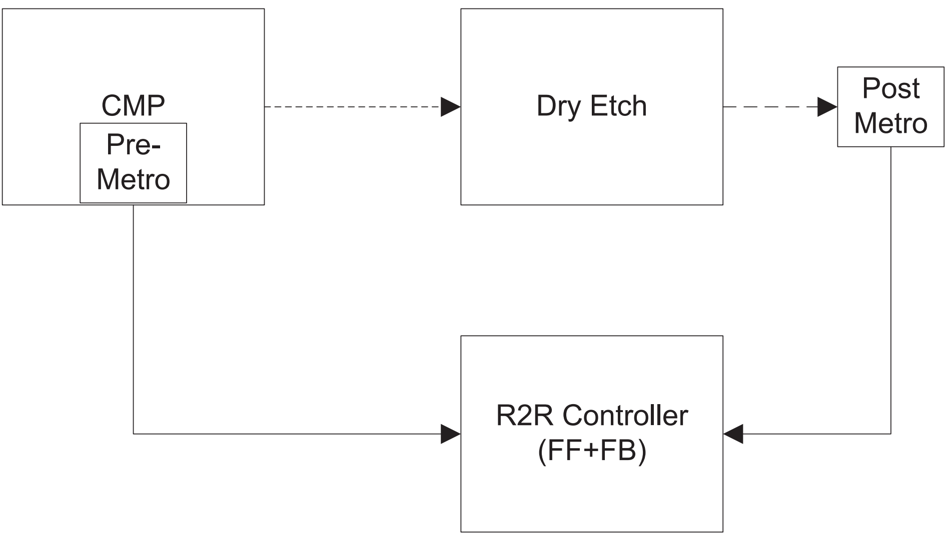


Figure 3.8. Wafer level multiple inputs multiple outputs feed-forward and feedback non-threaded dry etch R2R control schematic.

states which is shown in the following,

$$b_k = x_k^{Chamber_i} + x_k^{Device_j} + x_k^{CMPTool_k} \tag{3.53}$$

The advantage of non-threaded R2R control is that all metrology data are fully utilized by state estimation. Parallel testing between threaded and non-threaded controllers showed that the *Cpk* of non-threaded is 8% better than that of threaded R2R control, shown in Figure 3.9: one of the DE chambers was released for non-threaded R2R control, while the rest of the chambers remained as threaded R2R control. After the non-threaded R2R control was released for all chambers and all devices, we discovered that out of control (OOC) events in statistical process control (SPC) chart were reduced significantly across all products, as is shown in Figure 3.10.

3.6.2 Non-threaded Run-to-Run: Photo Tool Mismatch

Photo tool to tool mismatch issues have often been observed in DE R2R control. Figure 3.11 shows that one photo tool's state is higher than the other one's in

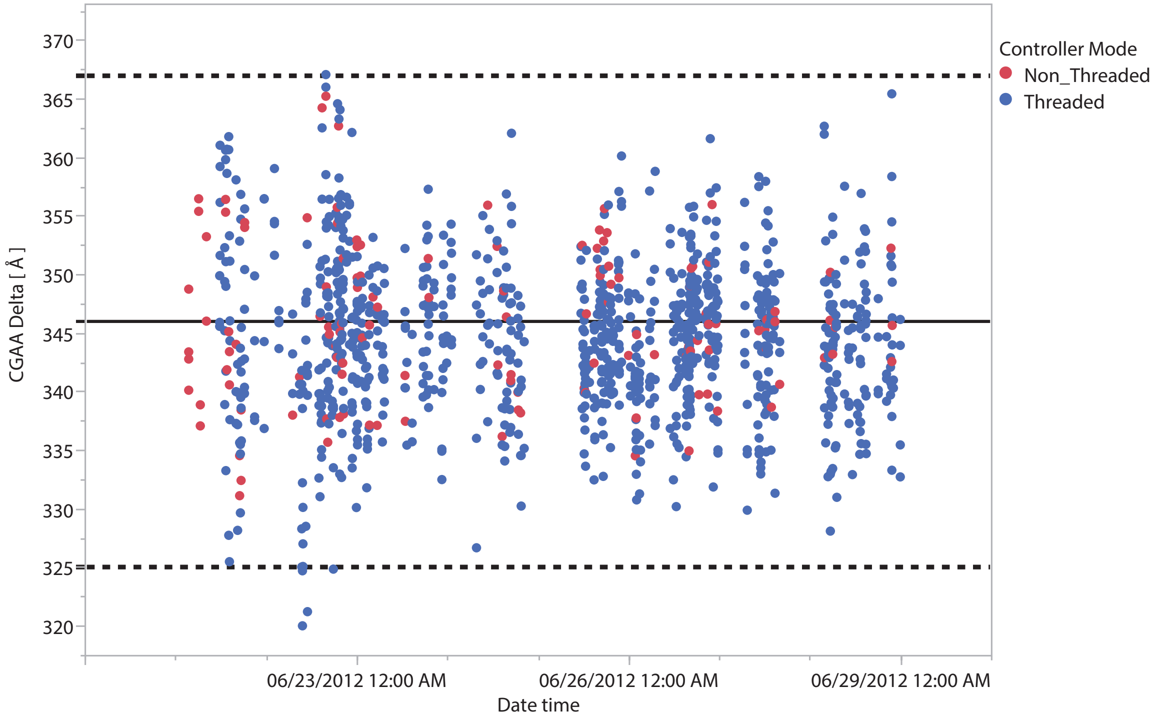


Figure 3.9. One chamber (red) was deployed with non-threaded R2R control and the other three chambers (blue) remained threaded R2R control.

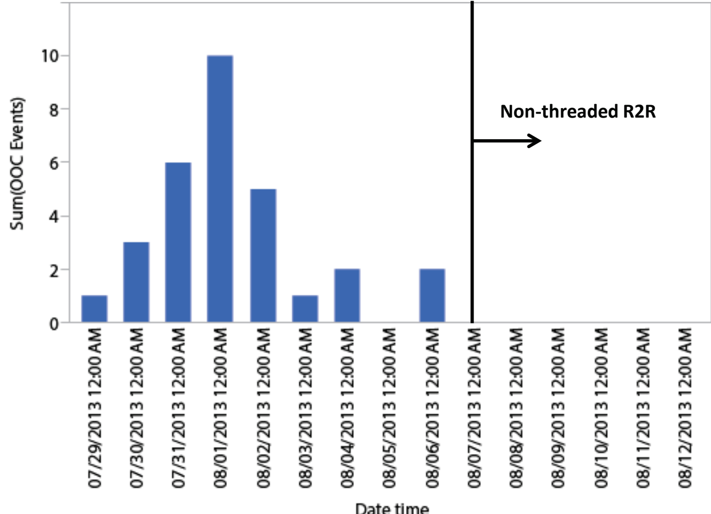


Figure 3.10. Out of control events were reduced after non-threaded Run-to-Run deployment.

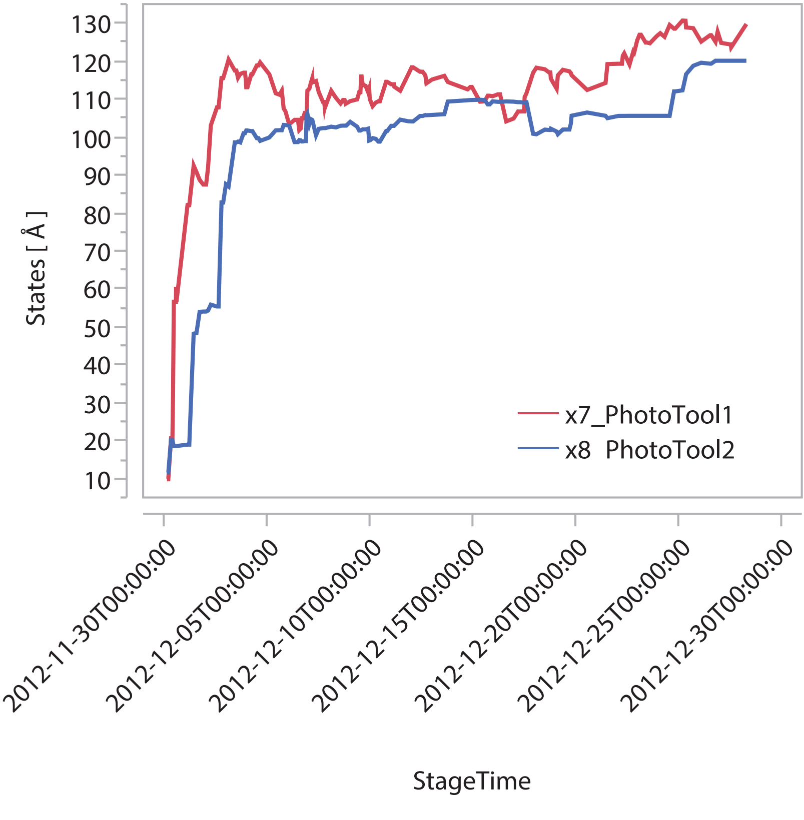


Figure 3.11. Photo tool to tool mismatch in the dry etch R2R control.

the non-threaded R2R control. Instead of separating each photo tool as an independent control thread, a non-threaded R2R controller was designed to solve the metrology dilution problem. Two different non-threaded R2R algorithms, EWMA and model-based non-threaded (3.37) (3.47) and (3.48), were compared against the threaded R2R controllers performance. The threaded R2R controller was in active mode while the two non-threaded R2R controllers were in passive mode (the data collection mode). In the EWMA non-threaded R2R controller [83], the R2R controller computes the intercept related to each context for every measured lot.

$$y_k = mu_k + b_k^{Device} + b_k^{PhotoTool} + b_k^{EtchChamber} \tag{3.54}$$

$$\begin{aligned}\hat{b}_k^{Device} =& \hat{b}_{k-1}^{Device} + \lambda_{Device}(y_{k-1} - mu_{k-1} - \\ & b_{k-1}^{Device} - b_{k-1}^{PhotoTool} - b_{k-1}^{EtchChamber})\end{aligned} \tag{3.55}$$

$$\begin{aligned}\hat{b}_k^{PhotoTool} =& \hat{b}_{k-1}^{PhotoTool} + \lambda_{PhotoTool}(y_{k-1} - mu_{k-1} - \\ & b_{k-1}^{Device} - b_{k-1}^{PhotoTool} - b_{k-1}^{EtchChamber})\end{aligned} \tag{3.56}$$

$$\begin{aligned}\hat{b}_k^{EtchChamber} =& \hat{b}_{k-1}^{EtchChamber} + \lambda_{EtchChamber}(y_{k-1} - \\ & mu_{k-1} - b_{k-1}^{Device} - b_{k-1}^{PhotoTool} - b_{k-1}^{EtchChamber})\end{aligned} \tag{3.57}$$

Equation (3.54) showed how the intercept term is separated, and equations (3.55) to (3.57) described how each intercept state is updated, where λ is the EWMA damping factor. Two non-threaded R2R controllers were in passive mode for data collection, while the threaded R2R controller was in active mode to control the process. The estimated output of a non-threaded controller is computed by following.

$$\hat{y}_k^{NT} = y_k + m(u_k^{NT} - u_k^{TH}) \tag{3.58}$$

where u_k^{TH} is recommended setting of the threaded control and u_k^{NT} is the non-threaded recommended setting in passive mode. $\hat{y}_k^{NT}$ is the estimated output of a non-threaded R2R controller if it is in active mode.

The head-to-head comparison results are listed in Table 3.2.

Table 3.2. Head to head non-threaded R2R control comparisons

Controller Type	Cpk	PPM
Threaded	1.20	180
EWMA Non-threaded	1.21	174
Model-based Non-threaded	1.24	123

The model-based non-threaded R2R performed the best in terms of Cpk and the threaded R2R controller performed the worst. PPM is parts per million, which is the probability of out of control events in the SPC chart.

3.7 Discussion

A few interesting topics will be discussed in this section; some of them might lead to the future research.

3.7.1 State Estimation of the Non-threaded R2R Controller

One of the advantages of the non-threaded R2R control is that the intercept states of the non-threaded controller is updated much more frequently than that of the threaded R2R controller. Data in Figure 3.12 are real production data that we obtained from a non-threaded controller, which showed that the state of a non-threaded controller is updated more frequently. A better process control can be realized in non-threaded R2R control because the process and equipment drift can be captured more quickly. On the other hand, the non-threaded R2R control module is tuned less aggressively, compared to the threaded R2R module, as discussed earlier. Referring to equation (3.47), in the wafer level non-threaded R2R controller we built, the R/Q ratio is set at 0.01, which is much smaller, compared to 0.1 typically used in threaded R2R controls. The non-threaded R2R controller tuning technique can be a future research topic. Figure 3.13 showed that it took about 10 valid metrology events for the controller to reach steady state from the zero initial state. During such transition period, non-threaded R2R control is automatically downgraded to threaded controller, due to business rules validation. Recently, Stuber [84] suggests that historical data can be used to estimate the initial states, which can help the controller reach optimal states quickly.

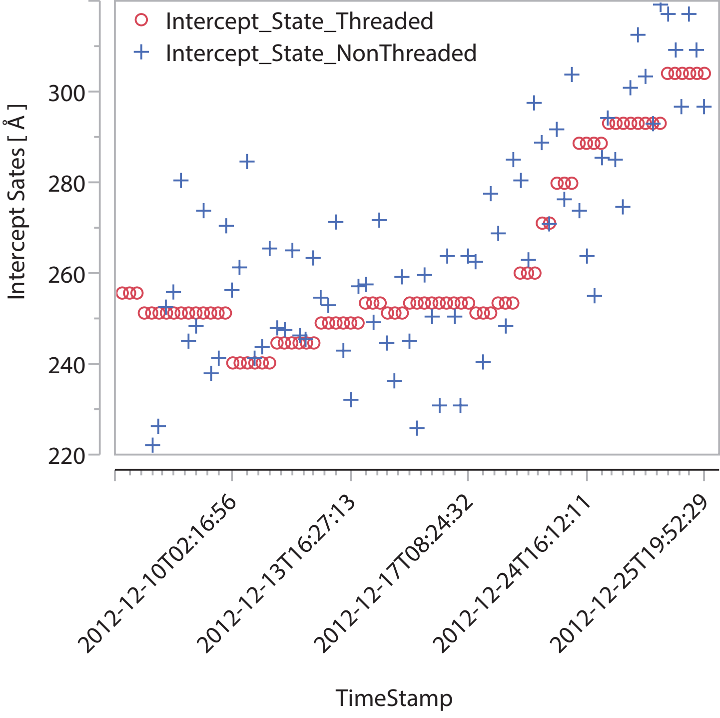


Figure 3.12. Comparison of threaded and non-threaded intercept states updating frequency.

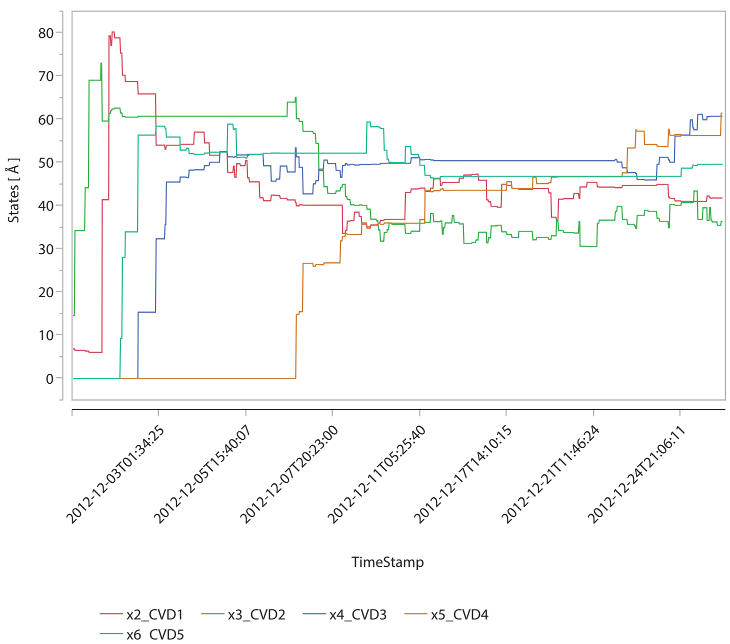


Figure 3.13. It took about 10 runs from zero initial state to steady state for CVD tool states.

3.7.2 Tuning Non-threaded R2R Control

The non-threaded R2R control tuning process is outlined as the flow chart of Figure 3.14. The non-threaded R2R performance can be estimated and compared with that of the threaded R2R controller, using equation (3.58). It might take several tuning iterations to get the controller tuned to the optimum state, and one should not turn on the non-threaded control module, unless significant Cpk gain is observed. If this tuning process can be automated, then a self-tuning of the non-threaded R2R control can possibly be achieved.

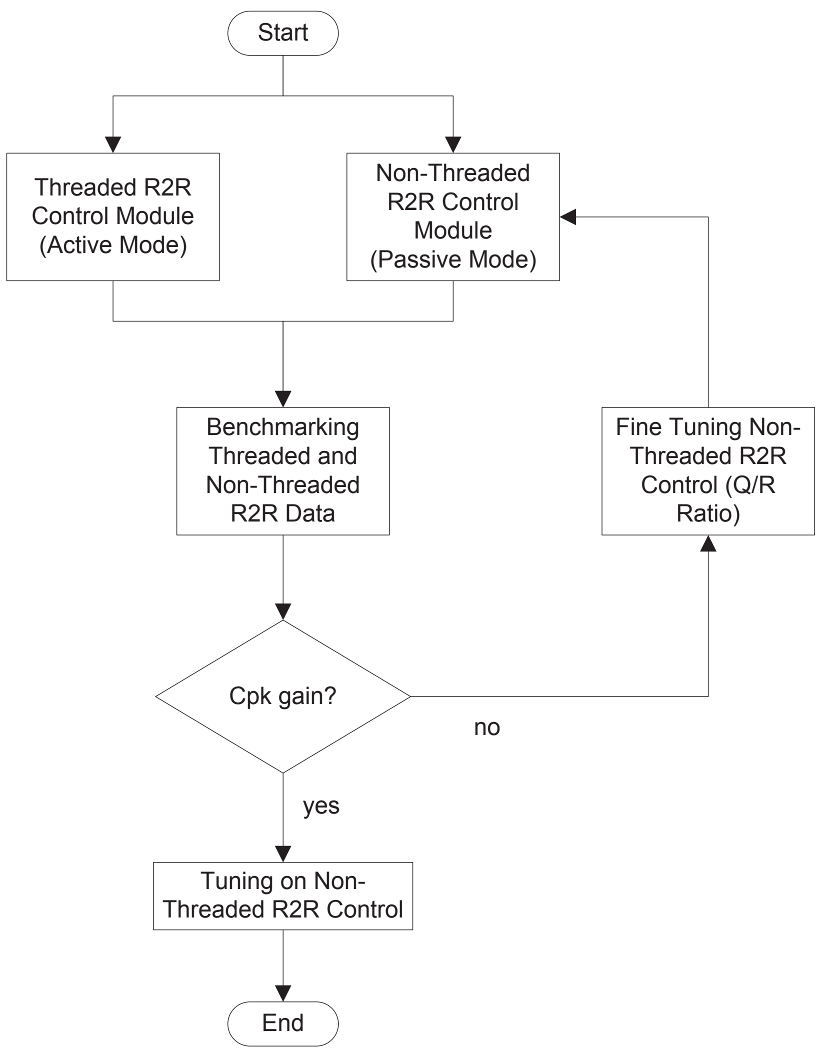


Figure 3.14. Benchmarking performance between threaded and non-threaded R2R control.

3.7.3 Reduction of Qualification Runs Using Non-threaded R2R

Other research [87] demonstrates that a non-threaded R2R controller can be possibly used to minimize qualification runs after tool maintenance. Historically, a lot of qualification runs (or target wafers) have to be conducted post maintenance events; sometimes it can be as many as one for every recipe. Such qualification runs will increase the cost of production and impact the cycle time. Production wafers used in qualification runs usually have to be downgraded and possibly scrapped. Increased cycle time is often seen, because the tool has to be offline for production until all qualifications are passed. However, if one can estimate tool state and recipe state in the following equation,

$$y_k = mu_k + b_k^{tool_i} + b_k^{recipe_j} \tag{3.59}$$

then we do not have to run as many qualification runs. Perhaps, we only need a single recipe qualification to release all recipes, so that a reduced number of qualification runs can be realized.

3.8 Summary

The unobservable states issue is a road block to build reliable non-threaded R2R controllers, and such problems have been addressed in a new hybrid non-threaded R2R controller design by automatically downgrading the controller from non-threaded mode to threaded control mode. This method proved to work with high reliability in a complex manufacturing environment. For a long time, people claimed that model-based non-threaded R2R controllers are not practical, because the number of states changes when new contexts are added, for example adding new tools or adding new reticles. We solved this issue by reserving dummy contexts in the non-threaded R2R controller without adding additional complexity. After the non-threaded R2R controller was deployed in a real production environment, we observed some negative impacts on servers. For example, long execution time of the state estimation. Limiting the number of dummy states and balancing the load among servers are proposed to overcome such problems.

We demonstrated the above methods in a DE R2R control, which has been deployed on one of the most critical processes in a production Fab. Threaded,

EWMA-based non-threaded and model-based non-threaded R2R controller performances were compared head to head on the same process and the same tool in production, and the data collection showed that the model-based non-threaded controller outperforms the other two control methods, EWMA non-threaded and threaded R2R controls.

Finally, we discussed how to benchmark the performance of non-threaded R2R control with that of threaded R2R control. A framework of an automatic tuning non-threaded controller is proposed, which leads to future research topics.

CHAPTER 4

ETCH RATE PREDICTION OF SILICON DIOXIDE FILM IN A DILUTED HF SOLUTION

4.1 Abstract

In order to reduce the cost introduced by metrology steps, while still allowing for advanced control with its high requirements on the availability of up-to-date measurements, there has been an increased interest in Virtual Metrology (VM) as an approach that can predict the metrology data without physically conducting the measurements. VM utilizes process trace data from the fault detection (FD) system of the current process step and selected data from previous steps, including pre-metrology data or other data from either product or process, to predict the post-metrology data. In this research, we propose to incorporate a multiphysics model into semiconductor virtual metrology to improve the prediction quality and accuracy, and VM prediction is integrated into the Run-to-Run (R2R) control system to improve the process capability. Furthermore, we demonstrate that the benefits of VM can be realized in high volume production, including variation reduction, excursion prevention, yield improvement, cost of ownership reduction and cycle time improvement.

4.2 Introduction

There can be hundreds of process steps in semiconductor wafer fabrications, depending on the complexity of the device, the circuit and its connections. To ensure good production yield, the semiconductor manufacturing requires frequent monitoring of both tools and processes at each step. Tool and process monitoring using statistical process control (SPC) often involves metrology operations, which

are deployed at almost every step in the advanced wafer fabrication facilities (Fab). In recent years, R2R controllers have been adopted by Fab to improve process capability, and R2R controllers optimize recipe parameters from lot-to-lot or wafer-to-wafer using feedback or feed-forward metrology data. Advanced process control and monitoring require a large amount of metrology data, but more metrology operations will increase manufacturing cost and increase cycle time. Therefore, there has been an increased interest in virtual metrology as an approach that can predict the metrology data without physically conducting the measurements.

The international technology roadmap for semiconductors (ITRS) identifies virtual metrology as an increasingly critical technology for improving productivity and reducing waste [88]. The movement from reactive to predictive in process control solutions has become a new standard in industry. Referring to Figure 4.1, VM is meant to predict postprocess physical and electrical quality parameters of wafers and/or devices from the information collected from the manufacturing tools and any other information available in the preprocessing steps and current steps (e.g., process trace data in FD) in Fab. The wafer level metrology predictions together with real metrology data can be fed forward to a wafer level R2R controller in the downstream process step. Such a control scheme would be beneficial for reducing wafer level variations in semiconductor manufacturing.

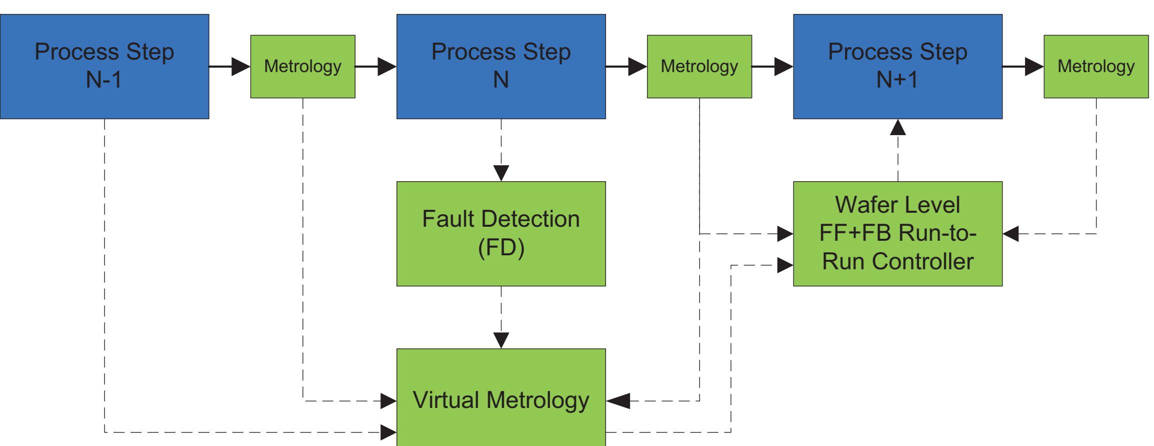


Figure 4.1. Place of execution of VM in a process flow

VM is commonly used in other fields, often known by different names, but this approach is relatively new in semiconductor manufacturing. In process industries, inferential or soft measurements are common. A classic example of inferential measurements is using the pressure and temperature measurements to predict composition of key components in distillation [89]. Chemometrics can be applied to solve the predictive problems in chemistry [90]. In predictive applications, properties of chemical systems are modeled with the intent of predicting new properties or behavior of interest. In both cases, the datasets can be very large and highly complex, involving hundreds to thousands of variables, and hundreds to thousands of cases or observations. Multivariable analysis, such as principal components analysis (PCA) and partial least squares (PLS), has shown their effectiveness at empirically modeling the low rank structure, exploiting the interrelationships or latent variables in the data. Partial least squares in particular was heavily used in chemometric applications such as pattern recognition and signal processing [91] for many years before it began to find regular use in other fields [49, 92], such as semiconductor manufacturing.

In semiconductor manufacturing, VM systems are built at different metrology steps [93], which can be classified by the following types:

1. Film thickness measurements [Easy]
2. Critical dimension (CD) measurements [Medium]
3. Electrical parameters measurements [Medium]
4. Defect inspection and scans measurements [Difficult]

Film thickness metrology is relatively easy to predict, because the film thickness is usually determined by a single processing step, such as chemical vapor deposition (CVD) and physical vapor deposition (PVD). CD and electrical parameters are harder to predict, because they are usually related to multiple process steps and multiple feed-forward components. The data on defect inspection and scans or real time defect analysis (RDA) is very difficult to predict, because it is related not only to the current process, but also to the whole process module or even the whole process integration [93].

4.3 Motivations

In this research, we use physical and chemical reaction models in building VM systems. Such an approach is substantially different from current approaches of purely empirical modeling, which identify correlations based exclusively on process trace data or FD data. Most of the published VM systems in literature were built on regression algorithms such as Neuron Networks, PLS and Kalman filter. In recent years, moving horizon PLS has gained popularity for constructing linear VM models, while Neural Networks seems to be the dominant approach for modeling non-linear systems. The challenge for these statistical methods in our research is that the predicted metrology is not consistently accurate enough to be used by a R2R controller at critical process steps. Therefore, we hope that the prediction quality can be improved further by incorporating multiphysics and process knowledge into the VM system.

The other motivation for this research is to integrate VM systems into R2R controllers. R2R controllers can be optimized in that R2R model, process gain, is updated by VM systems from time to time. Without VM, the process gain is usually set at constant, which is obtained in a DOE. Furthermore, we would like to compare multiphysics integrated models with pure statistical regression models and identify pros and cons of these two methods.

VM data has been proposed to be integrated into wafer-to-wafer (W2W) R2R controllers [2,94]. However, with the current metrology and manufacturing execution system (MES) system in a manufacturing Fab, it is still hard for W2W control to be realized, except for some tool types having onboard metrology. The predicted quality parameter in this research is the etch rate of silicon dioxide in a diluted HF solution, and the tool type running such a process is a batch tool instead of a single wafer processing tool. Therefore, our intention is not to build a W2W R2R control using the VM system, but to demonstrate additional VM benefits for this batch process, including R2R model optimization, excursion prevention by monitoring product wafers while processing and online or off-line metrology sample reductions.

4.4 Background

4.4.1 Virtual Metrology Using Statistical Models

Statistical models such as principle component analysis (PCA) and partial least squares (PLS) have been successfully used for process monitoring, fault detection and reconstruction [49]. Both PCA (4.1) and PLS (4.2)(4.3)(4.4) are linear regression methods [3]. The input matrix of X can be decomposed into a loading P, a score T and residual:

$$X = TP^T + \tilde{X} = TP^T + \tilde{T}\tilde{P}^T \tag{4.1}$$

where X is input data matrix, T is score matrix, P is loading matrix and $\tilde{X}$ or $\tilde{T}\tilde{P}^T$ is the residual. Such decomposition can be done through either singular value decomposition (SVD) [95] or nonlinear iterative partial least squares (NIPALS) [50] procedure.

The objective of PCA is to maximize the variance along the loadings and minimize the residual, while the objective of PLS is that PLS not only tries to minimize the residual, but also wants to maximize the correlations between the scores of input matrix X and output matrix Y. In other words, it also maximizes the correlations between T and U.

$$X = TP^T + \tilde{X} \tag{4.2}$$

$$Y = UQ^T + \tilde{Y} \tag{4.3}$$

where X is a matrix of features for the independent variables (or process data), Y is a matrix of features for the dependent variables (or metrology data), P is the matrix of X loadings, Q is the matrix of Y loadings, T is the matrix of X scores and U is the matrix of Y scores.

First, X and Y are scaled to be zero mean and unit variance. Then the matrix X is decomposed into a score matrix T and a loading matrix P, while at the same time, matrix Y is decomposed into a score matrix U and a loading matrix Q. If Y has only one variable, then the decomposition of Y can be omitted by simply setting $Q = 1$. The outer models perform the principal component regression of both X and Y by equations (4.2) and (4.3).

The inner model establishes a linear relationship between scores u_i and t_i using least square regression (4.4):

$$u_i = m_i t_i + e_i$$
$$m_i = \frac{u_i^T t_i}{t_i^T t_i} \tag{4.4}$$

where u_i and t_i are one of the scores of Y and X, respectively, m_i is one of the regression coefficients of the inner model and e_i is one of the intercepts of the inner model regression. The NIPALS procedure of PLS decomposition can be found in [3].

Let's denote u_1 as the first column of score vectors for the Y matrix, and t_1 as the first column of score vectors for the X matrix. The second projection scores pair (t_2 and u_2), or second columns of T and U, are correlated, but usually less than the first pair (t_1 and u_1). The third projection scores pair (t_3 and u_3) are less correlated than the second projection scores pair (t_2 and u_2)$\cdots$, and so on. By inserting a new observation in X space, one can obtain t_1, t_2, t_3 $\cdots$, which give predicted values of u_1, u_2, u_3 $\cdots$ via equation (4.4), which leads to the predicted value of Y via equation (4.3).

PLS regression methods are often chosen for VM models [2, 96, 97]. The advantages of PLS regression include:

- PLS is able to handle high dimensional and collinear data
- It is easy to interpret the results
- Online implementation is straight forward

Neural Networks is the other empirical model for VM [5, 98, 99]. Compared with linear regression PLS, it handles nonlinear systems. The Neural Networks model was used to predict CVD thin film thickness, as shown in Figure 4.2. The VM model consists of one input layer, one hidden layer and one output layer. There are m neurons in the input layer ($x_1, x_2, \cdots, x_m$), and corresponding m selected variables after data preprocessing. z_1, z_2, $\cdots$, z_n are neurons in the hidden layer, which corresponds n training samples. Three neurons (y_1, y_2 and y_3) in the output

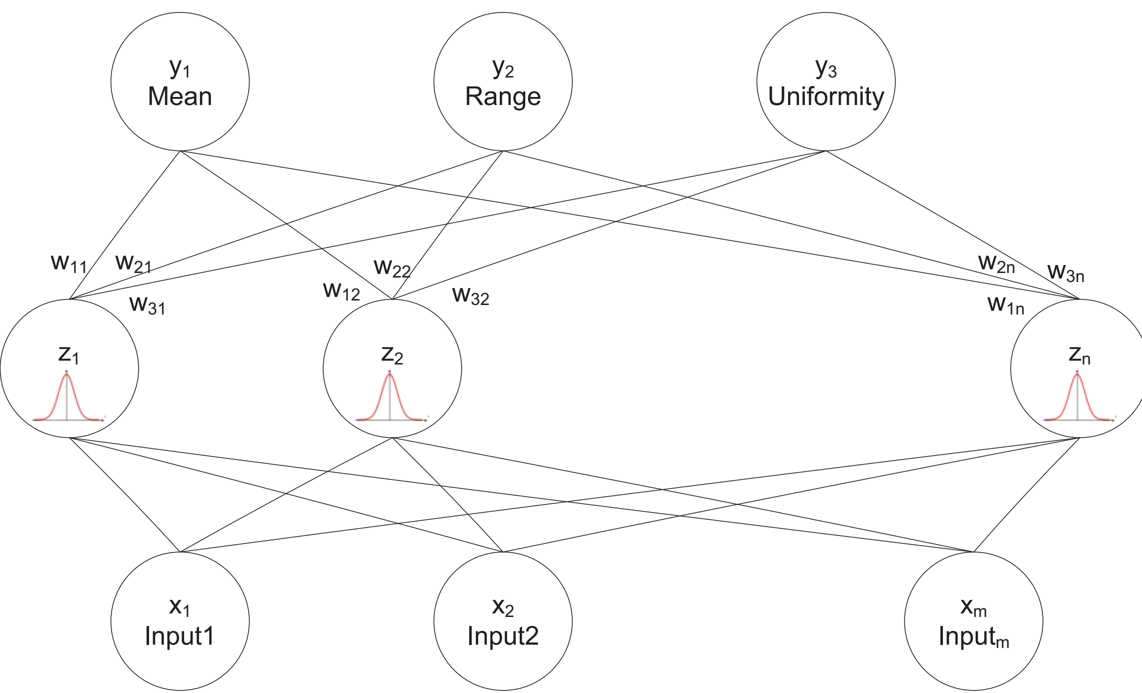


Figure 4.2. Neural network based VM model for CVD thickness [5]

layer are used to predict the Mean, Range and Uniformity, respectively. Gauss function is chosen for the transfer function in the hidden layer, while the output transfer function can be a simple linear function. Back-propagation neural networks (BPNN), piecewise linear neural networks (PLNN), fuzzy neural networks (FNN), simple recurrent neural networks (SRNN) and radial basis function neural networks (RBFN) have been tested for modeling VM [5, 57–61]. It is true that Neural Network is capable of representing complex nonlinear functions, while PLS and PCA have the advantages of easily handling variable collinearity and data dimension reductions.

In a typical data transformation and data flow for VM, the process data and metrology data are collected from production equipment and metrology equipment respectively. First, the raw data are checked to ensure its completeness to assure data quality, then it is normalized to be zero mean and unit variance. The data dimension reduction is performed next by eliminating insignificant parameters. VM prediction and reliance index (RI) are then calculated. RI indicates the confidence of VM predictions and a greater value of RI represents a better credibility of VM

prediction results. Such data flow structures have been proposed since 2007 [59].

The importance of data quality in virtual metrology was emphasized to build accurate VM models [98,100]. There are three pieces related to data integrity. The first piece is the quality of the data. The second piece is the quality of the data processing system. A good data processing system should be able to perform data "cleaning" and transformation. Data "cleaning" fills in the missing values and eliminates outliers. Data transformation usually means normalization of collected process data. The last, but not least, is the data dimension reduction, which extracts the important parameters from many process data items.

Besides the data quality research, more VM algorithms were explored. The recursive moving window PLS [2] gained its popularity. A typical PLS model is based on historical data and such a model is then updated, when enough data points have been accumulated. Using all the data points from the beginning would lead to computational problems. In a recursive PLS method that uses a moving window, which can be used to calculate a new PLS model every time, a new data point is obtained, and old data points that are out of that window are discarded. The recursive moving window PLS was designed to be used by W2W R2R controllers. W2W R2R control can be only realized on limited tools with onboard metrology (or integrated metrology) installed; however, VM data can enable wafer level R2R controls on process tools without onboard metrology [2]. A plasma etch process is relatively difficult to be modeled, because it is often associated with incoming variations. Etch bias, the difference between two CD patterns, was modeled in [67], and the Neural Networks algorithm was chosen, due to its fast computations. Etch rate modeling of plasma etch was also explored in 2009 [52]. In this work, several variables selection and modeling techniques were examined, and the best performing model was Neural Network. Linear and nonlinear algorithms were compared to predict CVD film thickness [101,102]. The best VM model was neural network (nonlinear modeling). In 2010, Kalman filter-based VM [103] was introduced, and support vector regression (SVR) [104] was evaluated in 2011. A promising VM model, canonical variate analysis (CVA), was benchmarked with PLS in 2011 [105]. It showed that CVA outperforms PLS,

because CVA captures the directions of maximum correlation between process sensor variable input x and quality variable y, while a PLS model may include the directions representing variations in the process sensor variables that are irrelevant to predicting quality variables. CVA aims to find a set of canonical vectors W_x and W_y (4) that maximize the correlation between the output quality variables and input process variables that are orthogonal to each other. After finding the canonical vectors, a model prediction of the quality variables at the time point $k_t h$ can be written as:

$$\begin{aligned} \tilde{\mathbf{y}}(k) &= (\mathbf{W}_y)^{-1}(\mathbf{B} \cdot \mathbf{x}(k)\mathbf{W}_x) \\ \mathbf{B} &= diag(b^1, b^2, \cdots, b^N) \end{aligned} \tag{4.5}$$

where $\mathbf{W}_x$ and $\mathbf{W}_y$ are canonical vectors. The coefficients of this model $b^1, b^2, \cdots, b^N$ can be obtained by standard least square estimation:

$$\begin{aligned} b^i &= (\mathbf{X}^T \mathbf{w}_x^i)^T \mathbf{Y}^T \mathbf{w}_y^i pinv[(\mathbf{X}^T \mathbf{w}_x^i)^T (\mathbf{X}^T \mathbf{w}_x^i)] \\ &= \frac{\rho^i \sqrt{(\mathbf{w}_y^i)^T \Sigma_{yy} (\mathbf{w}_y^i)}}{\sqrt{(\mathbf{w}_x^i)^T \Sigma_{xx} (\mathbf{w}_x^i)}} \end{aligned} \tag{4.6}$$

where Σ_{xx} and Σ_{yy} are the covariance matrices of input and output, and *pinv* is Moore Penrose pseudo inverse.

Recently, PLS-based VM was used to achieve real-time release of medical device components [96]. Another effort was made to model plasma etch using PLS in 2013 [97]. The main contribution of this work was to use T-PLS [106] to monitor data quality and to filter outliers. T-PLS has the advantage, based on total projection to latent structures, because the standard PLS structure has limitations in which the scores can contain variations not related to the output y. All above statistical methods, either linear or nonlinear, are all empirical data driven models.

Evaluating confidence levels for virtual metrology is critical, because the VM system is incomplete without a reliance index (RI). The RI and the global similarity index (GSI) [107] are proposed to gauge the degree of reliability. The RI and its threshold value are obtained by analyzing the process trace of the production tool to thereby determine whether the virtual metrology result is obtained with high

confidence. In order to help in gauging the degree of reliability further, the GSI and the individual similarity index (ISI) are proposed to define the degree of similarity between the input set of process data and all of the historical sets of process data.

4.4.2 Etch Rate Prediction Using PLS

The one bath (ONB) system, shown in Figure 4.3, is a 20 liter volume tank which is used to mix the HF 49% chemical with deionized (DI) water. To obtain precise concentration, more than 20 liters of DI Water are introduced into the tank, resulting in an overflow in the ONB tank. The reason for such overflow is to ensure that the mixing process is more uniform. As shown in Figure 4.4, the chemical charging time of this process is about four minutes, with the HF flow rate close to 100 ml/min and the DI wafer flow 50 $liter/min$, for the recipe of 500:1, in that 500 units of DI water are mixed with 1 unit of HF 49% by volume. The HF flow is turned on about one minute after the recipe starts (or trace data collection starts), and the window start is defined as a few seconds after HF flow is turned on to filter out flow meter noise, caused by the turbulence and the window ends at the end of chemical charging. The means and standard deviations within the defined window are calculated for each process trace data as the process indicators, which can be used in process monitoring in the FD system.

Besides bath the HF flow rate and the bath DI water flow rate, other process trace data are also collected, including bath temperature, N_2 pressure and so on. The prediction of the oxide etch rate in the diluted HF solution may be related to all of the available process data or only a subset of them. After several iterations of PLS evaluations, five process indicators are selected to the PLS model to predict bath etch rate with best prediction result:

1. Bath temperature mean
2. DI water flow rate mean
3. DI water flow rate standard deviation
4. HF flow rate mean
5. HF flow rate standard deviation

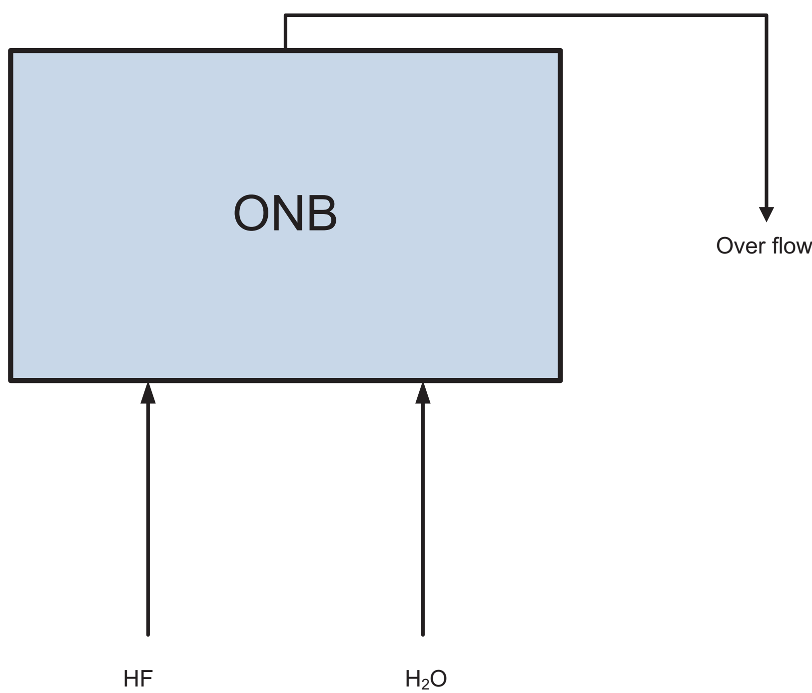


Figure 4.3. 20-liter ONB tank schematics

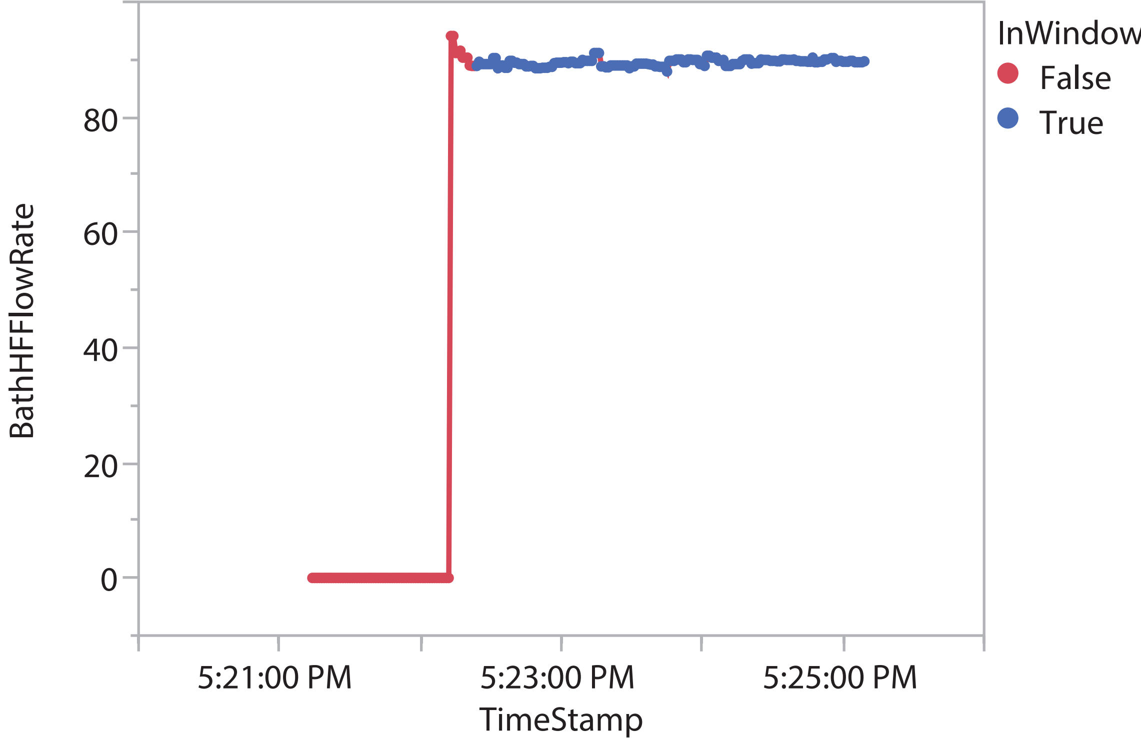


Figure 4.4. Chemical charge window of HF flow rate

PLS evaluation results are plotted in Figure 4.5. To obtain a reliable model, a minimum number of principal components should be selected in the model. The optimum number of principal components can be determined by both predicted residual error sum of squares (PRESS) and percent variation explained [92].

$$PRESS = \sum (y_{cal} - y_{act})^2 \tag{4.7}$$

where y_{cal} is the calculated value through the PLS model and y_{act} is the actual value. In our evaluation, four principal components are chosen, as shown in the top graph in Figure 4.3, which results that 54% variation can be explained for cumulative Y as PRESS approaches to a minimum value 0.782. The percent variations that can be explained by cumulative X are about 78.4%. The validation method we choose is "leave one out cross validation", computed for the validation sets based on the models constructed by leaving out one observation at a time, resulting in an R^2 of 0.54.

In the next chapter, we will show that an improved result can be obtained through a multiphysics model.

4.4.3 Limitations of Statistical Models

The traditional statistical VM models have the following limitations, which lead us to consider multiphysics-based models:

- **Poor input data quality issues.** Process trace data have a lot of noise. Without the physics knowledge, traditional statistical model-based prediction is downgraded by averaging out the noise.

- **Lack of physical correlation.** The process data items used by traditional statistical models are often the mean or standard deviation of process trace data, which does not carry any direct physical correlation with the output variable.

- **Large training data requirements.** A large training data set is often required for traditional statistical models.

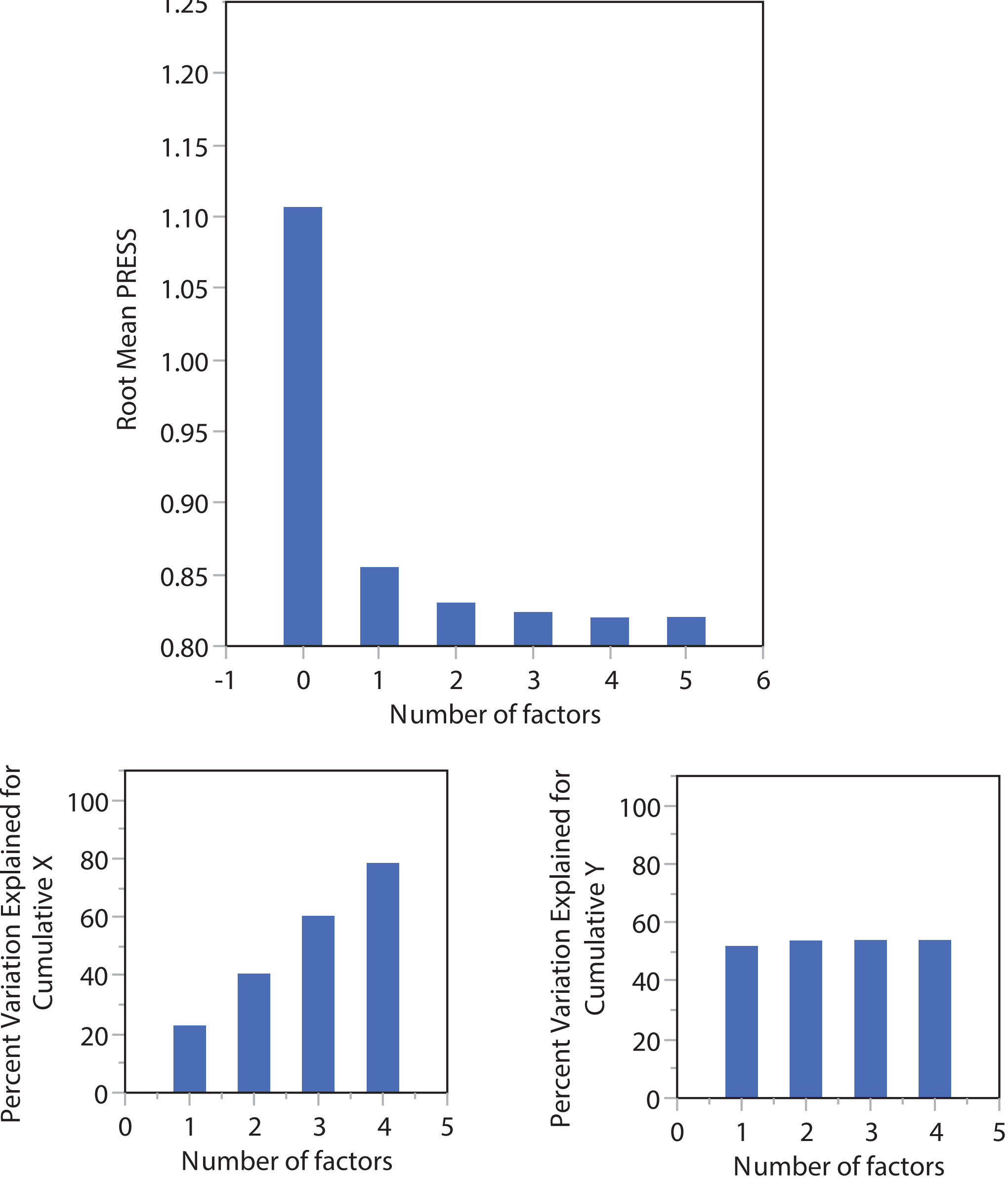


Figure 4.5. PLS evaluation results

- **Incoming material variations.** Traditional statistical models often cannot handle incoming material variations, including incoming chemical batch variations.

- **Statistical biases on selecting key process parameters.** Statistical methods can be biased on process data items reduction or selecting key process parameters, due to tool-to-tool or instrument-to-instrument mismatches.

4.4.4 Scope of This Work

The remainder of this paper is organized as follows. Section 4.5 explains the etch rate mechanism of silicon dioxide in a diluted HF solution, reliance index and the multiphysics-based model on etch rate predictions. Section 4.6 elaborates a new approach to integration of virtual metrology into R2R control. Next, Section 4.7 presents the benefit analysis of this virtual metrology project and finally, and Section 4.8 concludes the work with a summary.

4.5 Virtual Metrology Using a Multiphysics Model

4.5.1 Etch Rate Prediction Background

High volume semiconductor manufacturing facilities face challenges related to HF etch processing of wafers (or wet etch). For example, the high cost of test wafers, the large metrology sampling rate requirement and the considerable engineering time required make it difficult to achieve accurate control of etch rate. Taking chemical samples, measuring sample composition at a chemical lab and troubleshooting etch rate shift are time consuming. Depending on the process, the tools have to be taken off-line to sample etch rate measurement at least once a week; and the same thing occurs after each tool maintenance event. Such etch rate metrology has impacted the Fab cycle time performance. Using VM would improve cycle time. Other benefits of etch rate VM are listed as following:

- Much better process control and monitoring with virtual etch rate on every process run
- Less taxing on metrology capacity, because 100% sampling is not needed for process control
- Ability to set the framework to model all HF cleans (wet etch) with VM
- Lower use of less etch rate wafers, or non-process wafer (NPW) reductions
- Incorporation of VM into R2R control
- Decreased downtime postmaintenance events

In the traditional approach, the first task is to reduce data items in the prediction model. The principle component analysis can be used for this purpose; however, it will only provide the results, assuming linear dependence of the Etch Rate on the available process data. In our multiphysics approach, we first use the fundamental insight into the mechanisms influencing etch rate change, and then select the model structure to reflect such knowledge. This may lead to a nonlinear dependence of the etch rate on the available process data.

4.5.2 HF Etch Rate and Its Mechanism

We reviewed some previous work related to etch rate modeling of silicon dioxide in a diluted HF solution and its mechanism. Starting with basic reaction kinetics (4.8) of dilute HF solutions [108]:

$$
\begin{aligned}
[\mathrm{HF}] &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} [\mathrm{H^+}] + [\mathrm{F^-}] \\
[\mathrm{HF_2}^-] &\underset{k_4}{\overset{k_3}{\rightleftharpoons}} [\mathrm{HF}] + [\mathrm{F^-}]
\end{aligned}
\tag{4.8}
$$

where the equilibrium constants at 25 ^oC are listed in (4.9),

$$
\begin{aligned}
k_e1 &= \frac{k_1}{k_2} = 0.0013 \\
k_e2 &= \frac{k_3}{k_4} = 0.104
\end{aligned}
\tag{4.9}
$$

If we define, $A = [HF]$, $B = [HF_2^-]$, $C = [F^-]$ and $D = [H^+]$, then changes in the composition can be described by the following ordinary differential equations (ODE's):

$$
\begin{aligned}
\frac{dA}{dt} &= -k_1A + k_2CD + k_3B - k_4AC \\
\frac{dB}{dt} &= -k_3B + k_4AC \\
\frac{dC}{dt} &= k_1A + k_3B - k_2CD - k_4AC \\
\frac{dD}{dt} &= k_1A - k_2CD
\end{aligned}
\tag{4.10}
$$

Solving all these equations (4.10) together, the evaluation results of the initial concentration $[HF] = 0.05$ mol/liter is shown in Figure 4.6. Equilibrium constants at 25 ^oC can be looked up in the handbook [109].

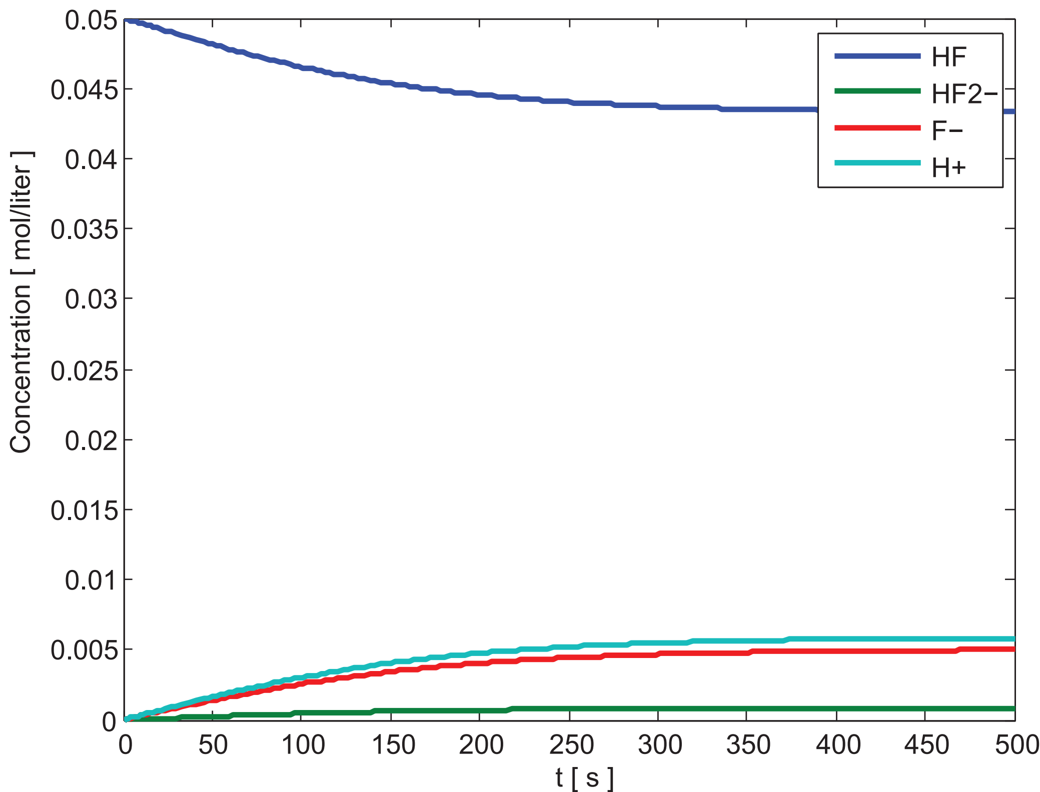


Figure 4.6. Composition changes of each species for initial concentration [HF] equal to 0.05 mol/liter

Given the composition in the etch tank, Judge [108] provided the following correlation for the etch rates,

$$\begin{aligned} R(Ang./sec) =& 5.0\times 10^{7}[HF_2^-]e^{\frac{-\triangle E_1}{RT}} \\ &+2.2\times 10^{6}[HF]e^{\frac{-\triangle E_2}{RT}} + C(T) \end{aligned} \tag{4.11}$$

where $\triangle E_1$ and $\triangle E_2$ are activation energy and $C(T)$ is a constant term depending on temperature.

An alternative etch rate model was proposed by introducing dimer term $(HF)_2$ into the HF etch rate mechanism [6]:

$$\begin{aligned} R =& a[(HF)_2] + b[(HF)_2]^2 + c[HF_2^-] \\ &+ d[HF_2^-]\times \log(\frac{H^+}{HF_2^-}) \end{aligned} \tag{4.12}$$

$$2\,HF \underset{}{\overset{k_e3}{\rightleftharpoons}} (HF)_2 \tag{4.13}$$

The equilibrium constant K_e3 of (4.13) is 2.7 liter/mol.

The equilibrium in the dilute HF solution considering dimerization (4.12) is plotted in Figure 4.7 by solving equations (4.10) and (4.13) together.

The conclusion from these chemical reaction models is that the SiO_2 film etch rate is dependent on total fluorine F concentration, activation energy and the temperature. The above etch rate models using chemical reaction are better than the empirical data-driven approaches such as PLS because they not only help us understand the fundamentals of chemical reaction and build better process indicators, but they also help us to correctly select key parameters (process variables) which should be part of the model. Next we will use this insight to develop and

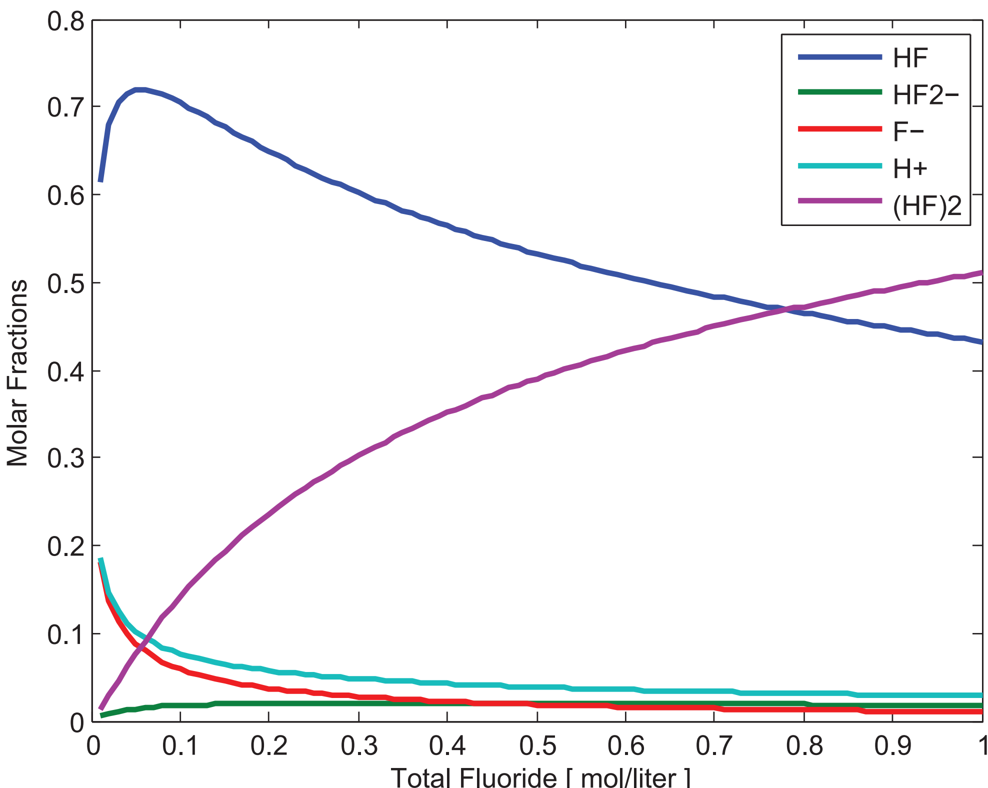


Figure 4.7. The calculated fraction of each component in an HF solution as a function of total fluoride concentration [6]

validate etch rate model using the correlation between predicted data and real HF weight percent (*wt*%) measurement from the chemical lab.

4.5.3 Etch Rate Prediction Using the Multiphysics-Based Model

According to equations (4.11) and (4.12), the silicon dioxide etch rate in a diluted HF solution depends on species concentration, activation energy and temperature. The temperature factor can be safely eliminated from the model because the temperature is very stable according to the collected process trace data, and the temperature is controlled precisely at $23 \pm 0.1^{o}C$. On the other hand, the activation energy of the chemical reaction remains the same for the same thin film. Therefore, our conclusion is that etch rate only depends on chemical species concentrations. Although it might be a nonlinear function, one can assume the linearity by eliminating higher order terms when HF solution is very diluted and the etch rate varies in a small range (e.g., from 22.5 Angstroms per minute to 26.5 Angstroms per minute):

$$R = b[HF]_{wt\%} \tag{4.14}$$

where $[HF]_{wt\%}$ is the weight percent of a dilute HF solution and b is a constant.

The prediction of etch rate actually becomes predicting the weight percent of a dilute HF solution. Experiments were designed for the prediction model for the weight percent of a dilute HF solution, and data collection was done successfully. The weight percent of the dilute HF solution can be measured accurately in the chemical analysis lab through a titration procedure [110], while the difficulties we encountered include the production constraints, tool safety interlock and chemical lab sampling logistics and so on. With the help from engineers from both the wet process and the process control system, the data collection and analysis were completed with the following conclusions. The weight percent of a dilute HF solution can be predicted by equation (4.15):

$$[HF]_{wt\%} = k\frac{Q_{HF}}{Q_{DIW}} \tag{4.15}$$

where Q_{HF} and Q_{DIW} volumetric flow rate means of HF and DI Water, respectively, and k is a constant, which can be different for every chamber or tank.

The measured and predicted weight percent of dilute HF solution is listed in Table 4.1. The predicted concentration using equation (4.15) is strongly correlated to the real measurement in the chemical analysis lab.

The governing equation of etch rate prediction is equation (4.16) after combining equations (4.14) and (4.15):

$$\hat{R} = K\frac{Q_{HF}}{Q_{DIW}} \tag{4.16}$$

where K is a coefficient, which depends on the chamber context and $K = bk$.

The coefficient K can be estimated for each context using an exponential weighted moving average (EWMA) filter (4.17) (4.18):

$$\hat{K}_N = \lambda K_{Cal,N} + (1-\lambda)\hat{K}_{N-1} \tag{4.17}$$

$$K_{Cal,N} = \frac{R_{meas,N} Q_{DIW,N}}{Q_{HF,N}} \tag{4.18}$$

where $K_{Cal,N}$ is the estimated coefficient using real etch rate metrology and the flow rates data from FD system at current run. Q is the mean value of FD traces within the window (the portion of blue color curve in Figure 4.4). λ is a weighting factor whose value is between 0 and 1.

4.5.4 Etch Rate Prediction Reliance Index

A reliance index (RI) is designed to gauge the reliability of the predictions and RI is also used to filter out the false alarms when VM is used for the propose of

Table 4.1. The measured and the predicted weight percent of HF solution

Tool	Lot ID	HF Wt% Measured	HF Wt% Predicted
P15T2	W960442.002	0.052	0.0539
P15T2	X122062.002	0.053	0.0533
P15T2	X155312.002	0.054	0.0541
P15T2	W289142.002	0.055	0.0540
P15T2	W307782.002	0.053	0.0537
P15T2	W326502.002	0.053	0.0534
P15T2	W334102.002	0.054	0.0540

processes monitoring. RI and global similarity index (GSI) are evaluated as the indicators of prediction quality [107, 111, 112].

Let $Z_{\hat{y}_i}$ denote the normalized prediction of VM and $Z_{\hat{y}_{ri}}$ denote the normalized reference prediction. The RI [111] is computed as,

$$RI = 2\int_{\frac{Z_{\hat{y}_i}+Z_{\hat{y}_{ri}}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \tag{4.19}$$

with $\mu = Z_{\hat{y}_i}$ if $Z_{\hat{y}_i} < Z_{\hat{y}_{ri}}$, $\mu = Z_{\hat{y}_{ri}}$ if $Z_{\hat{y}_i} > Z_{\hat{y}_{ri}}$ and $\sigma = 1$. In fact, the RI of (4.19) is the overlap area between two normally distributed bell curves, $Z_{\hat{y}_i}$ and $Z_{\hat{y}_{ri}}$, and $Z_{\hat{y}_i} \sim N(0,1)$, $Z_{\hat{y}_{ri}} \sim N(0,1)$. The reference prediction can be calculated using a least square multiregression method.

On the other hand, the GSI assesses the similarity of the input data sets between the model samples and recent data, and Mahalanobis Distance is used to quantify such similarity. Let r_{ij} denote the correlation coefficient between the parameters i and j in the input data, and there are k runs of input data and r_{ij} is computed as,

$$r_{i,j} = \frac{1}{k-1}\sum_{l=1}^{k} z_{il} z_{jl} \tag{4.20}$$

where z_{il} and z_{jl} are the normalized input data items and the correlation coefficients matrix for n parameters is defined as,

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{4.21}$$

The GSI can be computed by (4.22):

$$GSI_k = \frac{D_k^2}{n} = \frac{Z_k^T R^{-1} Z_k}{n} \tag{4.22}$$

where D_k^2 is the Mahalanobis Distance, and Z_k is the normalized input data of kth run, which includes n parameters.

Alternatively, we evaluate a new method to compute a RI through the probability of failure in a multicomponent system (4.23). Series systems function properly only when all their components function properly [17].

$$P(A^\wedge B^\wedge C) = P(A) \cdot P(B) \cdot P(C) \tag{4.23}$$

with the assumption that A, B and C are independent components. Therefore, proposed RI can be evaluated by (4.24):

$$RI = P_{DQ} \cdot P_{HF} \cdot P_{DIW} \cdot P_{HORIZON} \cdot P_{TIME} \tag{4.24}$$

and the explanation of each term in equation (4.24) is listed as following:

- P_{DQ} is the data quality value of FD data. This value is received from FD system directly, and it is related to missing data and other data quality factors during data collection in the FD system.

- P_{HF} is the two-tailed probability of the z-score of HF flow z_{HF} occurring by assuming that the HF flow rate is normally distributed. The population mean μ and the standard deviation σ can be computed by historical data:

$$z_{HF} = \frac{Q_{HF} - \mu}{\sigma} \tag{4.25}$$

- P_{DIW} is the two-tailed probability of the z-score of DI water flow occurring, which is similar to P_{HF}.

- $P_{HORIZON}$ is evaluated by the invalid records in the moving window horizon. The horizon length is a user defined value. The record is validated by a set of limits testing, for example the goodness of fit (GOF) of metrology is greater than a specified threshold (e.g., 0.9).

$$P_{HORIZON} = 1 - \alpha \frac{n_{inv}}{n_h} \tag{4.26}$$

  where α is the user-defined weight for the horizon factor, n_{inv} is the number of invalid records in the moving horizon and n_h is the user-defined horizon length.

- P_{TIME} is related to time elapsed between now to the time stamp of last model update. The longer the time since last model update, the lower the P_{TIME} value. P_{TIME} can be simply evaluated by a linear equation (4.27):

$$P_{TIME} = 1 - \beta \frac{t}{t_{min}} \tag{4.27}$$

  where β is the user defined weight for time factor, t is the time elapsed between now to last model update and t_{min} is minimum requirement of frequency for VM model update which depends on the speed of etch rate drift.

In the actual implementation, RI is not only used to disable the out of control action plan (OCAP) of predicted etch rate, but it is also used to monitor the health of the VM system. For example, if the RI is consistently lower than its threshold, then it would indicate that VM model update stops working for some reason. Furthermore, RI can be part of the dynamic tuning of a R2R controller [112].

4.5.5 Etch Rate Virtual Metrology Design

The main architecture is outlined in Figure 4.8, and the design of the VM involves two systems, FD and R2R. The connection between the two systems is through web services. The Run-to-Run system collects the flow rates data from the FD system to predict etch rate via equation (4.16). VM model parameters need to be updated from time to time to capture the chemical batch and chemical flow variations. Depending on the processes, some etch rate qualifications run once a week and some critical processes run every day. Whenever new etch rate actual metrology data (or etch rate qualification data) is available, the model parameters K will be updated by Run-to-Run, in other words, K will be re-estimated using equations (4.17) (4.18). K carries both the flow meter calibration and the variation of chemical batch HF 49% information, while the flow rate ratio carries the physics, or the multiphysics-based information. Both the etch rate predictions and the prediction reliance index are saved in FD system for OCAP.

4.5.6 Virtual Metrology Results Analysis and Discussions

The initial data collection results were very encouraging. As shown in Figure 4.9, the correlation (r^2) between actual measured etch rate and the predicted etch

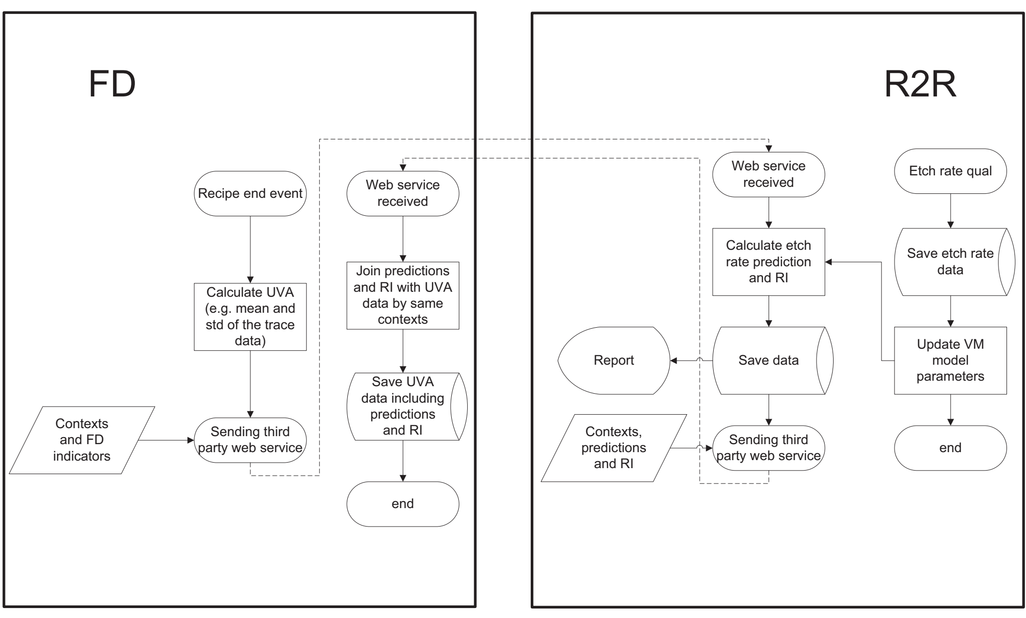


Figure 4.8. Virtual metrology design architecture

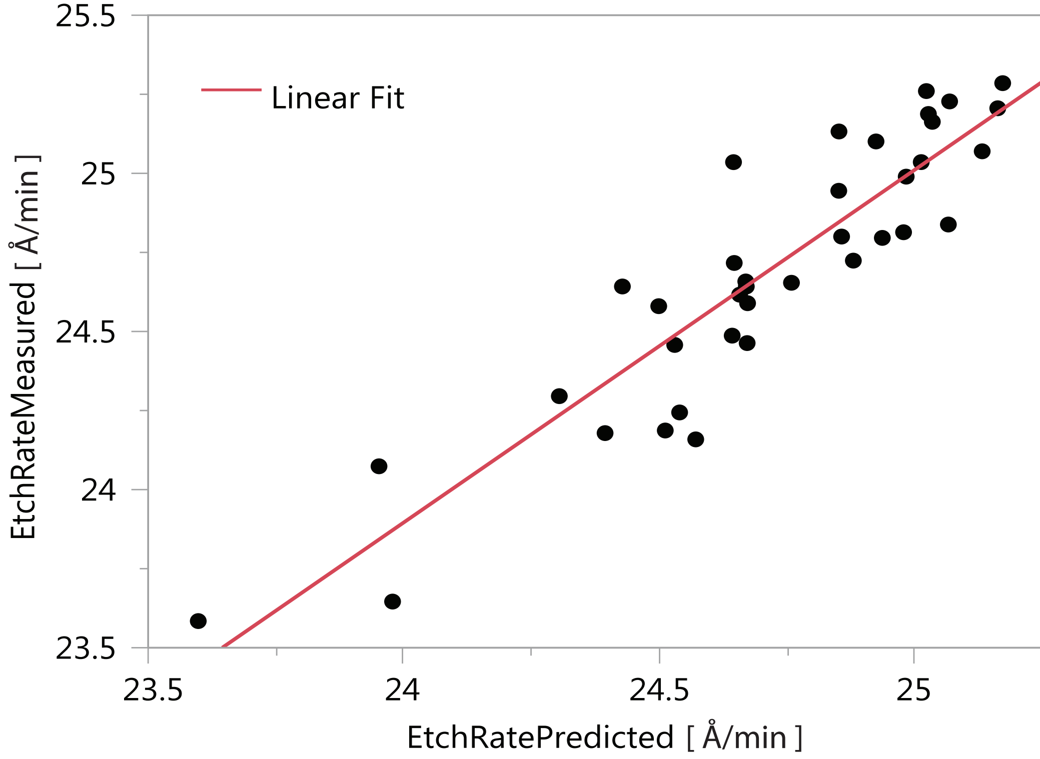


Figure 4.9. Initial results $r^2 = 0.83$: correlation between actual measured etch rate and predicted etch rate

rate is 0.83 for the first two weeks' data collected after this VM system was built, and the recipe that was used was 100 : 1, 100 units of DI wafer mixing with 1 unit of HF49% by volume. After collecting one year's data, the correlation r^2 value downgrades a bit to 0.64, as shown in Figure 4.10.

In an ideal situation, the slope and r^2 are equal to 1 and the intercept is equal to 0 for both Figure 4.9 and Figure 4.10. The results shown in Figure 4.10 are still compelling for the long-term VM data collection, and our results collected in the high-volume production are at least comparable to or better than the results published. After some troubleshooting of the outliers in Figure 4.10, we found that some of them are caused by test wafer preparation problems and the others are unknown events. Recently a logic was implemented as a solution for fast tracking the sudden changes: if the absolute prediction error, $|ER_{predicted} - ER_{measured}|/ER_{target}$,

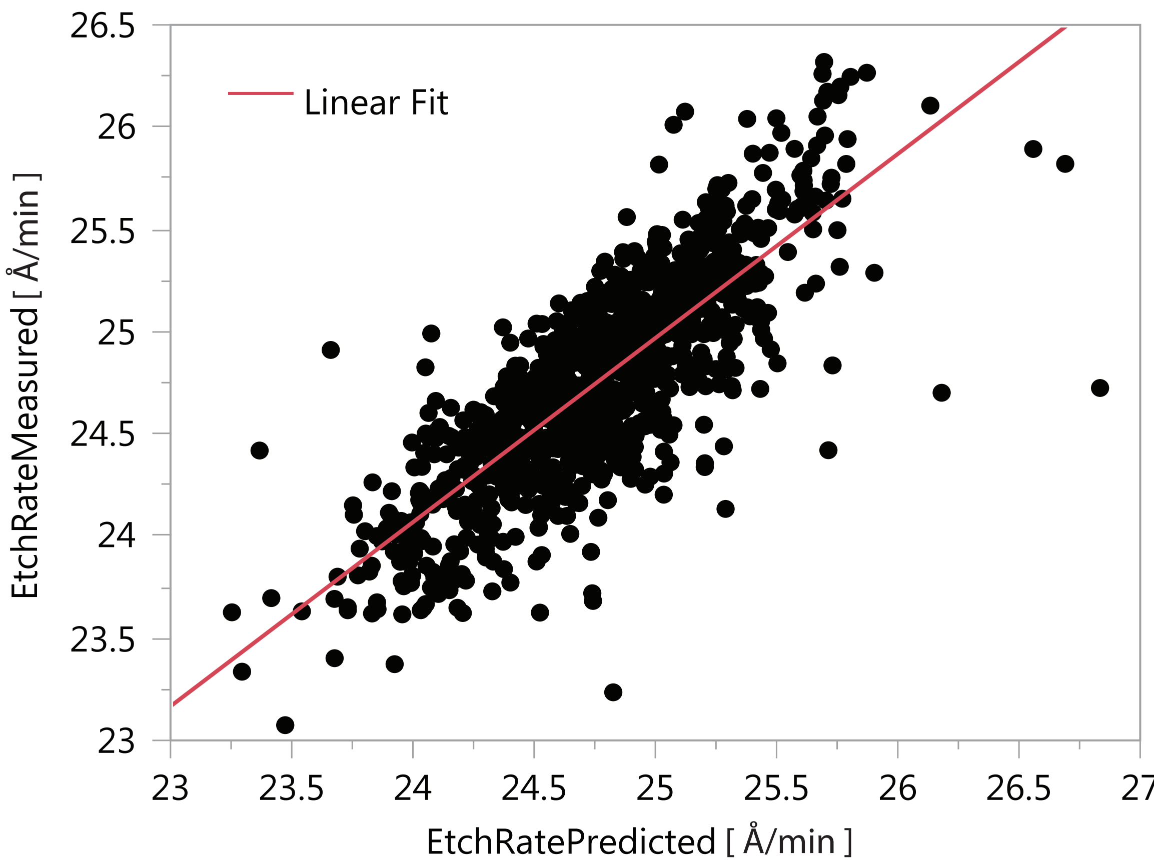


Figure 4.10. Long-term results $r^2 = 0.64$: correlation between actual measured etch rate and predicted etch rate

is beyond 10%, then the moving horizon of the VM will be reset, which means that VM needs to collect fresh data to make any new predictions, and the RI drops to zero immediately after reset. One of the benefits of implementing this is to better track dramatic shifts by unknown events, for example, some types of maintenance events. It would be even better if the maintenance event were integrated into the VM system for automatic reset.

4.6 Integration of Virtual Metrology into Run-to-Run Control

Typically an R2R controller adjusts recipe parameters based on metrology data to improve process capability, while metrology operations have constraints due to high cost and long cycle time. In a high-volume semiconductor manufacturing environment, wafer level metrology sampling is basically avoided due to these constraints. In many cases, only 10% to 20% of lots are sampled for the metrology and furthermore, there is a delay between process step and metrology steps, and as a result, R2R performance is downgraded. VM becomes a promising system to improve R2R control performance in two main aspects, real time metrology and 100% availability. Feeding VM data into R2R controllers is one of the usages to realize VM benefits, which includes feed-forward and feedback at following applications:

- **Wafer level feed-forward application.** By far, this is one of the most promising benefit models because wafer level R2R control is demanded for many critical steps. The wafer level predictions (VM data) can be fed forward to a wafer level R2R controller at the downstream process steps, refer to Figure 4.1, in such a way that all wafers in a lot are compensated to improve the process control performance. Such R2R systems requires high demand on prediction quality because typically, there is no dampening for the feed-forward components in the R2R control.

- **Lot level feedback application.** In most lot level feedback R2R control systems, only a small portion of the lots are sampled due to the cost and cycle time constraints. With the help of VM, 100% of the lots' metrology become available, either actual metrology or predicted metrology. Furthermore, the

R2R control performance is often downgraded due to metrology delays [79], while VM data become real time as it takes only seconds to compute metrology predictions. As a result, the lot level R2R control performance can be improved when 100% lot level data are available with VM data in real time.

- **Wafer to wafer (W2W) R2R application.** There are many benefits for W2W R2R control through VM, including minimizing metrology delay and improving wafer level process capability [2, 51, 113]. However, most of MES systems do not allow wafer level recipe adjustment after the lot is committed on the tool, so the wafer to wafer (W2W) control is very challenging to be implemented except some process tools having onboard metrology. Therefore, the W2W R2R implementations in the high-volume production are limited to the tools with onboard metrology systems today, while wafer level VM predictions will enable more W2W control applications in future.

- **R2R control model update application.** It's well known that R2R control models drift over time but it is assumed to be constant. For example, process gain (or slope) is assumed to stay constant, so the plant and model mismatch downgrades R2R control performance. In this research, R2R model (or slope term "etch rate") is updated through the VM, so that the R2R control can remain optimal as another benefit model of VM. We will discuss this more next.

In this VM research project, the outputs of the VM system are etch rate prediction and its reliance index. On the process control side, the process module controlled by R2R using VM data is called "local oxidation of silicon"(LOCOS) [114, 115], which is a typical process in the isolation structure of devices in semiconductor fabrication, and a conventional LOCOS isolation structure is illustrated in Figure 4.11, which includes the topographies of a semirecessed and fully-recessed LOCOS structure. An accurate process control of locally oxidized silicon process is essential for electrical performance of the isolation structure. The LOCOS process flow and steps are listed in Figure 4.12.

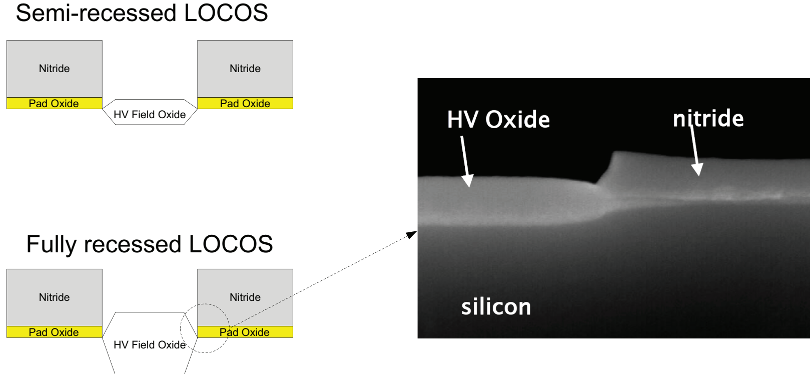


Figure 4.11. Shape and topography in a semirecessed and fully-recessed LOCOS structure

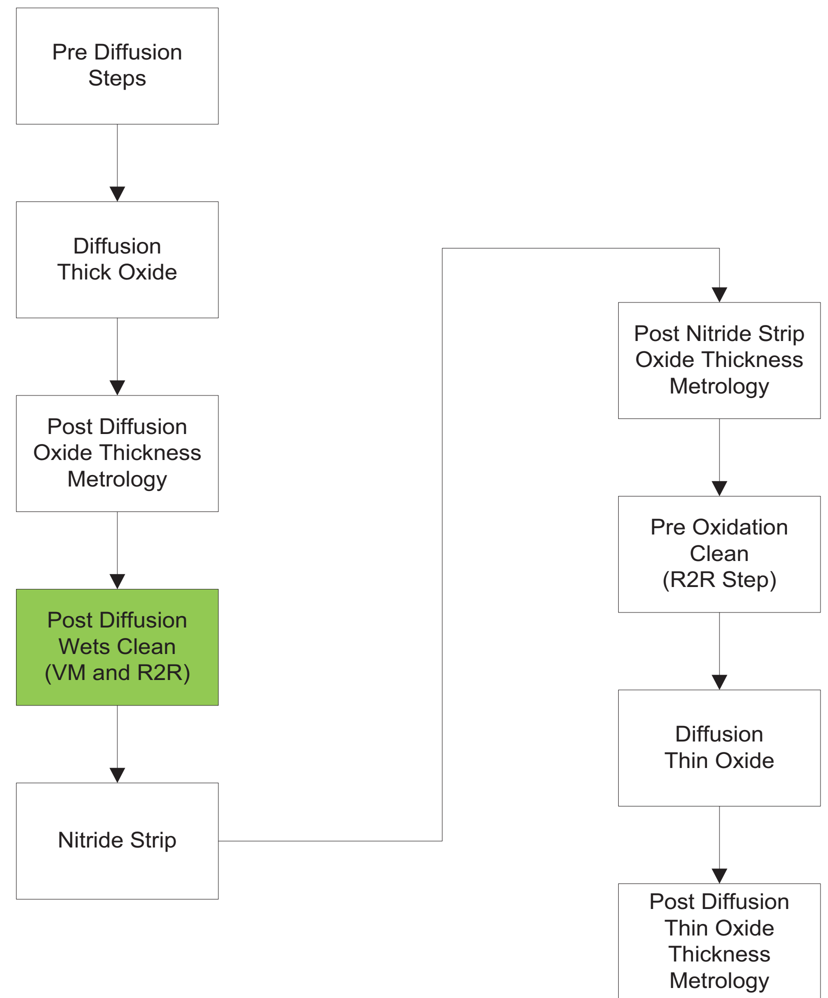


Figure 4.12. LOCOS process steps

Before the implementation of VM, the only R2R control step is "Pre Oxidation Clean," and problems of such a control scheme are listed in the following:

- Poor Cpk ($Cpk < 1$) was obtained at the final step "Post Diffusion Thin Oxide Thickness Metrology" even when 100% of the lots were sampled (so-called "over" sampling) at step "Post Nitride Strip Oxide Thickness".

- Long metrology tool time is spent at step "Post Nitride Strip Oxide Thickness" because 100% of lots were sampled, and this increased the cycle time of production line.

- Furthermore, the metrology tool was operated at its constrained capacity, and the cost of an extra metrology tool is close to two million dollars.

A better control scheme was developed with the help of VM at step "Post Diffusion Wets Clean" and the block diagram of R2R control using VM is described in Figure 4.13. The R2R control type is a feed-forward and feedback batch control system: the intercept state, b_k, is continuously updated by post metrology step "Post Nitride Strip Oxide Thickness" via the EWMA filter for feedback control:

$$b_{k+1} = \lambda(y_m - \hat{R}u_k - f_k) + (1 - \lambda)b_k \tag{4.28}$$

where λ is the damping factor, which is a value between 0 and 1, y_m is the metrology at k^{th} run, $\hat{R}$ is the process gain, u_k is the manipulated variable at k^{th} run and f_k is the feed-forward disturbance.

On the feed-forward side, "Post Diffusion Oxide Thickness Measurement",f_k, is a feed-forward component to R2R control. The R2R model (or process gain), $\hat{R}$, is continuously updated by VM system using equation (4.16). Such model update is critical for incoming feed-forward disturbance compensation as mentioned earlier, because there is no dampening factor for the feed-forward component f_k in (4.29). The etch rate $\hat{R}$ needs to be accurate, otherwise the controller recommended setting in following would be biased:

$$u_{k+1} = \frac{y_t - b_{k+1} - f_k}{\hat{R}} \tag{4.29}$$

where y_t is the control target of step "Post Nitride Strip Oxide Thickness" and u_{k+1} is the R2R recommended settings.

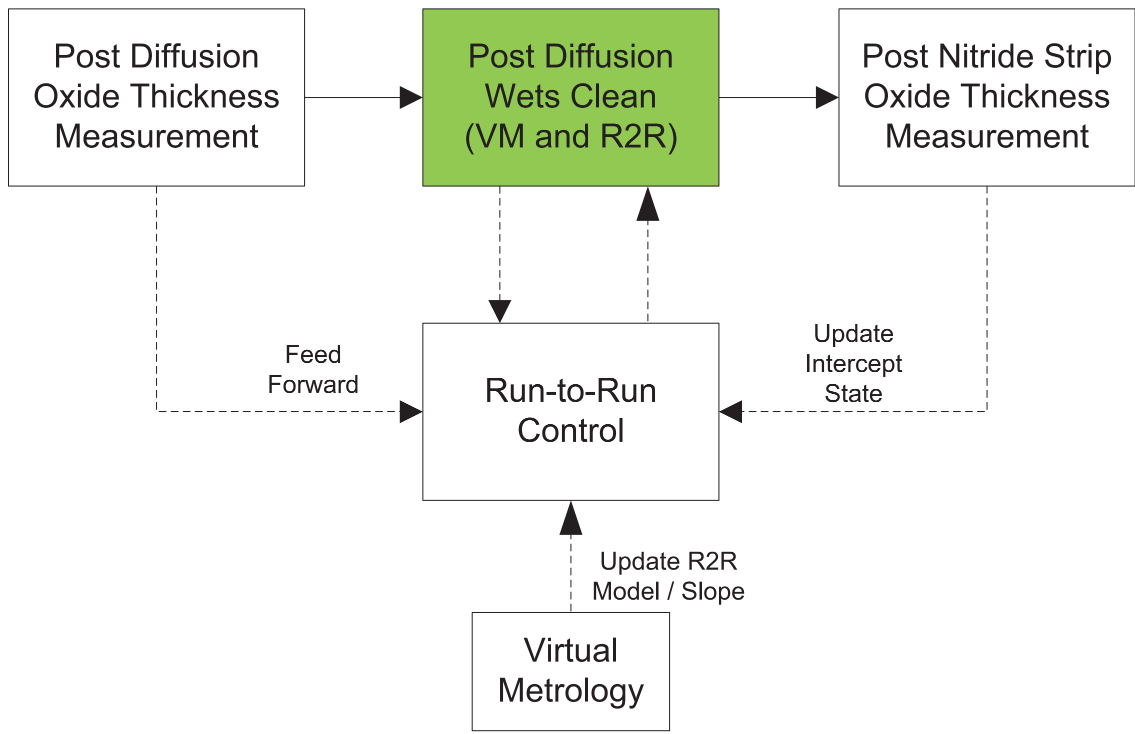


Figure 4.13. Run-to-Run controller model update through virtual metrology

For dampening noise impact from VM system, R2R process model $\hat{R}$ is updated only when the mismatch between plant and model is greater than certain threshold and also RI is high (for example, the mismatch is greater than 2% and RI is greater than 70%). We have demonstrated that such R2R and VM control scheme has improved the process capability greatly in high-volume semiconductor manufacturing production. Furthermore, it also improves other metrics, including excursion prevention, yield improvement, cycle time reduction and cost reduction. We will discuss them in detail in Section 4.7.

4.7 Benefit Analysis

Cheng et al. have proposed the VM benefit model [93], which includes the following aspects and cost reduction measures, thus providing a competitive edge for semiconductor manufacturing companies:

- Cycle time reduction by skipping on-line or off-line metrology
- Yield enhancement through process capability improvement

- Cost saving of nonproduct wafers
- Capital expenditure (or capex) reduction

In fact, the VM system in this research project has realized the benefit of every single aspect in above benefit list, which has reduced the manufacturing cost significantly.

4.7.1 Excursion Prevention

The predicted etch rate and its reliance index provide additional process indicators to prevent process excursions. Those new process indicators are saved in the FD system with limits and OCAP, which can be used for process monitoring in real time and total inspection of all batches. For example, it was the newly predicted etch rate and its reliance index that highlighted differences between the actual etch rate and the predicted etch rate. After investigating further, a leak was found in the tank. This is one of the examples that VM can be used for excursion prevention.

4.7.2 Process Capability Improvement

After VM and R2R control strategies were deployed in real semiconductor manufacturing, a significant improvement of process capability index, in terms of Cpk, has been observed. As shown in Figure 4.14, only one tool was released first as a pilot for this new control scheme starting from June 2nd, and all tools were released on August 5th based on the excellent results of this pilot line. The Cpk was improved from 0.8 to 1.48 for the SPC chart of “Post Nitride Strip Oxide Thickness” step, an 85% Cpk improvement. Moreover, the performance of the final LOCOS metrology step “Post Diffusion Thin Oxide Thickness Metrology” was also improved, as shown in Figure 4.15. The Cpk was improved from 0.57 to 1.21, a 110% Cpk improvement. Figure 4.15 collected much less data compared with Figure 4.14 due to a low sampling rate. Since “Post Nitride Strip Oxide Thickness” is a pre-metrology step for the R2R controller at “Pre Oxidation Clean,” it needs more lots to be measured for feed-forward. We confirm that the Cpk improvement in Figure 4.15 is the result of incoming variation reduction by R2R and VM at step “Post Nitride Strip Oxide Thickness” in Figure 4.14.

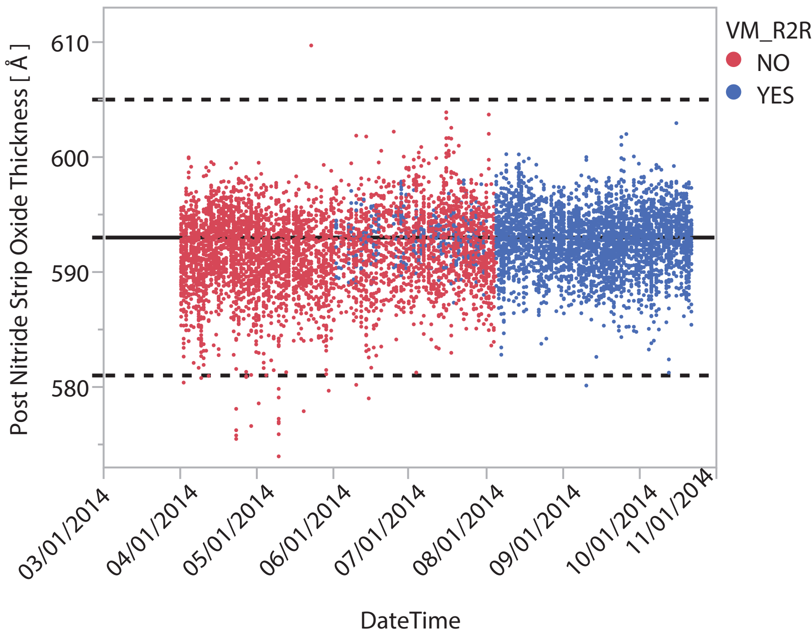


Figure 4.14. SPC performance for "Post Nitride Strip Oxide Thickness": VM R2R vs. no VM R2R. The sampling rate is 100% of the lots.

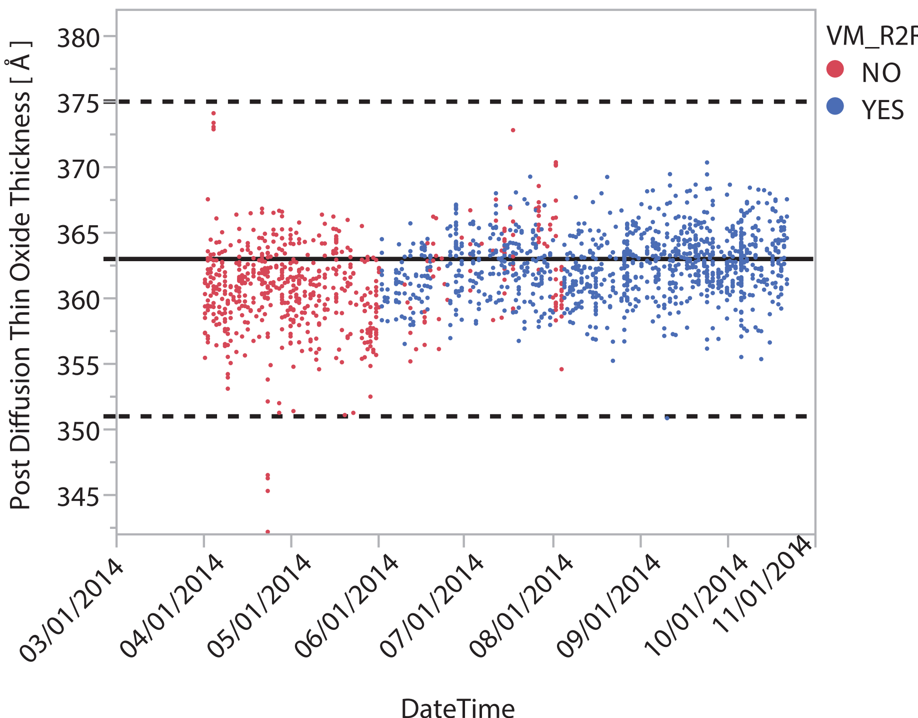


Figure 4.15. SPC performance for "Post Diffusion Thin Oxide Thickness": VM R2R vs. no VM R2R. The sampling rate is 10% of the lots.

4.7.3 Yield Improvement

Besides the Cpk improvement of in-line SPC charts of metrology data, data analysis showed that the variation of final electrical parametric data corresponding to oxide thickness is improved by 20%. As a result, such electrical parametric data improvement is translated into a 0.15 die per wafer yield gain.

4.7.4 Cycle Time Reduction

The cycle time reduction is achieved mainly from the sampling rate reduction at metrology step "Post Nitride Strip Oxide Thickness". The Cpk improvement at this oxide thickness metrology step has enabled us to reduce sampling rate from 100% to 50% at metrology step "Post Nitride Strip Oxide Thickness". As shown in Figure 4.12, the feed-forward and feedback R2R controller at the "Pre Oxidation Clean" step required a 100% lot level sampling rate to support process control and process capability in the past. After the implementation of VM and the R2R controller at step "Post Diffusion Wets Clean", the process capability is improved significantly, which is shown in Figure 4.14. Therefore, an over sampling requirement is no longer needed. As a result, a 50% sampling reduction can be translated into 8.6 minutes of cycle time reduction due to the escape rate improvement.

4.7.5 Cost Reduction

Cost saving is the driving force of semiconductor manufacturing improvement, and yearly cost saving can be estimated from the below equation [93]:

$$
\begin{aligned}
Saving = W_o * [\frac{1}{1-(\triangle CT_P + \triangle CT_M)} - 1] * (1+\triangle Y) \\
(P-C) + \triangle Cost_M + \triangle Cost_T - Cost_V - Cost_Q
\end{aligned}
\tag{4.30}
$$

where W_o is number of wafer output per year, $\triangle CT_P$ is % cycle time reduction due to VM allowing production wafers to skip metrology sampling, $\triangle CT_M$ is % cycle time reduction due to VM allowing less test wafers used in the off-line tool monitoring process and more intelligent dynamic metrology schemes, $\triangle Y$ is % enhancement on process capability, reduction in scrap and so forth, P is the average selling price, C is the average production cost, $\triangle Cost_M$ is the cost saving of test wafers per year when applying VM, $\triangle Cost_T$ is the capex reduction per year when applying VM,

$Cost_V$ is the maintenance cost of VM and $Cost_Q$ is the additional cost per year due to false alarms or missed detections by VM.

In this single virtual metrology project, our estimated cost saving is US$346,564 per year, which includes the cycle time reduction by skipping the metrology step, yield benefit, cost savings of the test wafers and the metrology tool capex reduction.

4.8 Summary

This work demonstrated that incorporating physics and a chemical reaction model into virtual metrology can improve virtual metrology prediction quality. It's a better method to select key process variables for building VM, and it's a better method to establish "meaningful" process indicators compared with traditional statistical regression models. We have also demonstrated that prediction results of multiphysics-based model are improved over those obtained in traditional statistical approaches. Furthermore, multiphysics models require less training data than other approaches due to fast convergence behavior. Finally, incoming variations and raw materials from chemicals and gases are to be accounted for in the VM model; without taking those into account, VM prediction accuracy can be biased or compromised sometimes.

Besides the previous proposed W2W R2R control enabled by VM in literature, the R2R control model parameter, which is the etch rate (or slope) in our case, can be updated through virtual metrology in "real time". In this way, R2R control is operated at its optimal state, especially for compensating feed-forward components. We also demonstrated in high volume semiconductor manufacturing that integrating VM into R2R controller can improve process capability and yield significantly.

Our contributions in this project include: 1) this work is the first one to incorporate a multiphysics model into semiconductor virtual metrology to improve prediction quality and accuracy; 2) this work is the first one to use a VM system to update R2R control model parameters; 3) this work is the first one to account for incoming chemical batch variations in virtual metrology models; and finally, 4) this work realized almost all of the VM benefits in high volume production including

excursion prevention, process capability and yield improvement, cycle time and cost reduction.

CHAPTER 5

A GENERIC DIFFUSION FURNACE VIRTUAL METROLOGY

5.1 Abstract

The recipe adjustments by a furnace R2R controller, in fact, only keep five monitoring wafers' thicknesses on target, so it improves the thickness uniformity along boat slots only to some extent. However, the thickness profile in each heater zone or across all five heater zones cannot be completely eliminated because of temperature uniformity problems and depletion effect. In this project, we use a design of experiment and a multiphysics model to predict the wafer thickness of each boat slot, which can be potentially used as feed-forward components by a wafer level controller at the downstream process step to improve wafer level variations. On the other hand, several challenges are encountered during this research project, for instance, the queue time effect and the compensation of R2R adjustments. New methods are proposed to solve these challenging problems, and the results obtained are discussed.

5.2 Introduction to the Diffusion Furnace

5.2.1 Introduction to Furnace Process

The name "diffusion" furnace was created long before ion implant was invented in the mid-80s [116, 117]. The main function of a diffusion furnace was a thermal diffusion process which introduces dopants into silicon. A typical diffusion process involved a mask of silicon dioxide, which defined the area to receive the dopant. The dopant was introduced to the wafer surface in either liquid or gas form, and then dopant was driven deep into the wafer through the high temperature process, for example, 900 ^{o}C. Over the last few decades, the original function of "diffusion"

furnace has been completely replaced by ion implant, and nowadays, diffusion furnaces function in three main applications [118]:

- Growth of a new layer of material, which consumes some of the substrate; for example, silicon dioxide grows on the substrate while consuming some of the silicon substrate.
- Deposit of new thin films without consuming silicon substrate; for instance, ploy-silicon deposition.
- Heat treatments including anneal and alloy.

The diffusion processes often involve chemical reactions, low pressure or "vacuum", and high temperature. The chemical reaction turns gas and substrate into another material, and the property of the new material is very different from the original substrate. For example, a "dry" oxidation [119,120] can be described as,

$$Si + O_2 \rightarrow SiO_2 \tag{5.1}$$

while a "wet" oxidation [119] can be described as,

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2O \tag{5.2}$$

"Dry" oxidation is used to obtain better film quality, while the deposition rate of "wet"is much faster than the "dry" oxidation. There are two reasons why "wet" oxidation is faster. Firstly, small molecules diffuse faster than big molecules: The H_2O molecule is smaller than the O_2 molecule, so the H_2O molecule diffuses faster to the reaction site, where the silicon substrate interface is. Second, the more reactive radicals provide higher oxidation rate: both O^{2-} and OH^- radicals are reactive radicals in "wet" oxidation, and these radicals are looking for combining and re-combining opportunities. Therefore, a better oxidation rate can be achieved in "wet" oxidation. Approximately 44% of the thickness of the grown oxide comes from substrate silicon, which is shown in Figure 5.1.

Almost all diffusion processes are processed in low pressure or vacuum for better film property, thickness controllability and safety reasons. In most cases, the

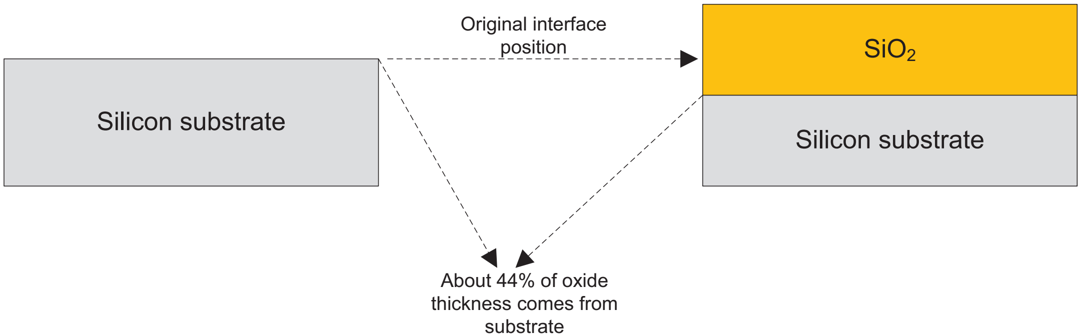


Figure 5.1. Growing oxide in diffusion furnace consumes some of the silicon substrate

diffusion processes involve high temperature (e.g., 800 to 1000 ^oC) to achieve the desired film property, but high temperature also causes dopant migration, as a side effect [118].

The deposition of new material using a diffusion furnace is often called low pressure chemical vapor deposition (LPCVD). The high temperature drives chemical reactions and low pressure prevents reactants from recombining or reacting with gasses in air and improves step coverage. For example, poly-silicon can be deposited via LPCVD in the diffusion furnace at a temperature of about 500 ^oC [121]:

$$SiH_4 \rightarrow Si + 2H_2 \quad (5.3)$$

and such deposition of new material without consuming any silicon substrate is shown in Figure 5.2.

Another application of a diffusion furnace is the heat treatment for either copper annealing or annealing after ion implants [118]. For the better quality of copper, a copper annealing process can transform the crystal structure into a more desirable grain structure for either resistance reduction or chemical and mechanical polishing (CMP) uniformity improvement. On the second application of heat treatment, the ion implant process damages the silicon surface, and devices do not function properly without repairing these damages. A rapid thermal process (RTP) has been developed to repair the implant damages as shown in Figure 5.3. At the same time, the implanted dopants can be also activated by moving them to the silicon lattice

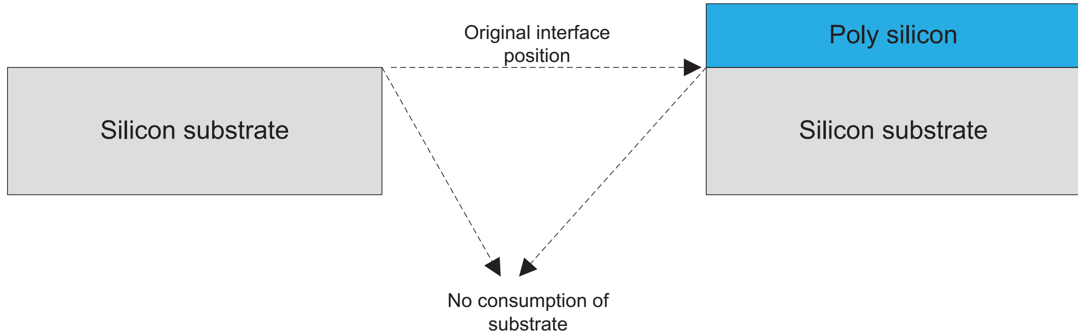


Figure 5.2. Depositing poly-silicon in diffusion furnace does not consume any silicon substrate

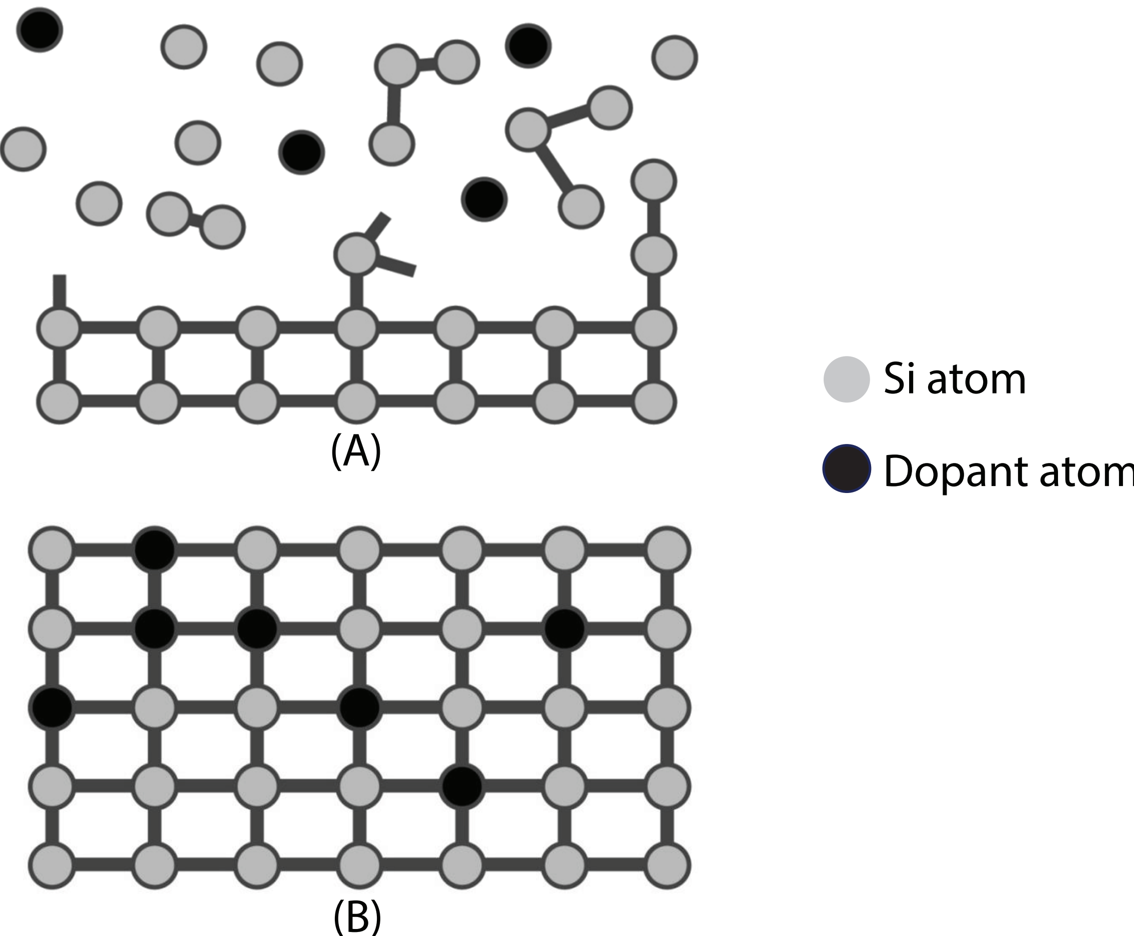


Figure 5.3. The damaged lattice and its repair: (A) The lattice is damaged by an ion implant process (B) The lattice is repaired by a rapid thermal process

site to improve the electrical property of devices being fabricated. Figure 5.3 shows differences before and after the "repair" by a RTP.

5.2.2 Introduction to Furnace Equipment

The diffusion furnace is a batch tool, which processes 4 to 6 lots together. Figure 5.4 describes a Kokusai vertical furnace which consists of a front opening unified pod (FOUP) storage rack, two FOUP load ports, one wafer transfer station, a boat where wafers are loaded and a tube where wafers are processed. As seen in Figure 5.5, the liner goes between the wafer boat and tube, which is also very important. Without a liner, gas would exit directly to exhaust.

Compared with many single wafer processing tools, there are many advantages of a diffusion batch tool in that the operating cost is cheaper and the process time (RPT) per wafer is shorter provided that it is a "full" batch. It does have some

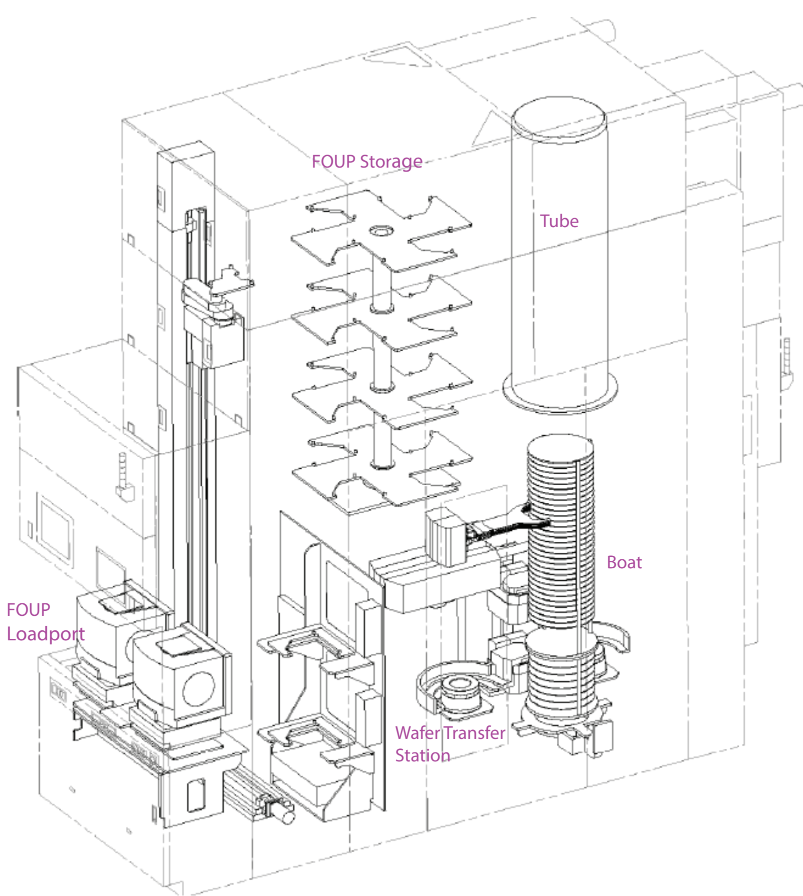


Figure 5.4. Kokusai vertical diffusion furnace

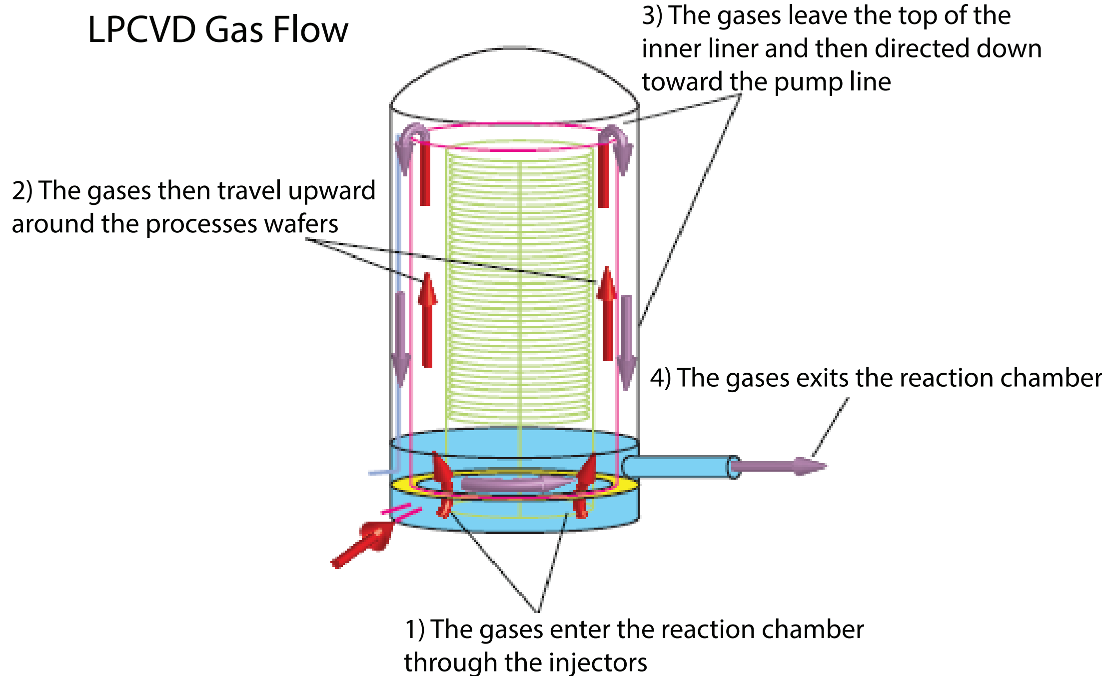


Figure 5.5. The configuration of boat, liner and tube

disadvantages. For example, the process is very slow, and sometimes it takes as long as 8 hours. Furthermore, it can impact multiple lots in a batch when equipment has problems like particle issues, pump failures or recipe aborts, etc. The other unique problem of a diffusion furnace is that the thickness profile in normal conditions exists due to the misalignments between the monitoring wafer locations and heater centers, refer to Figure 5.6, as well as the gas depletion effect of a vertical furnace where the gas is usually introduced from the bottom in Figure 5.5. R2R control of the diffusion furnace drives the thickness of all five monitoring wafers to the control target, while it does not address the thickness variation within a heater zone. In this paper, Section 5.3 illustrates R2R control of a diffusion furnace, Session 5.4 summarizes the motivations of developing diffusion furnace virtual metrology, Session 5.5 discusses the DOE data and evaluation results, Session 5.6 explains our multiphysics of furnace VM and some of its challenges and Session 5.7 gives our conclusions and the future work.

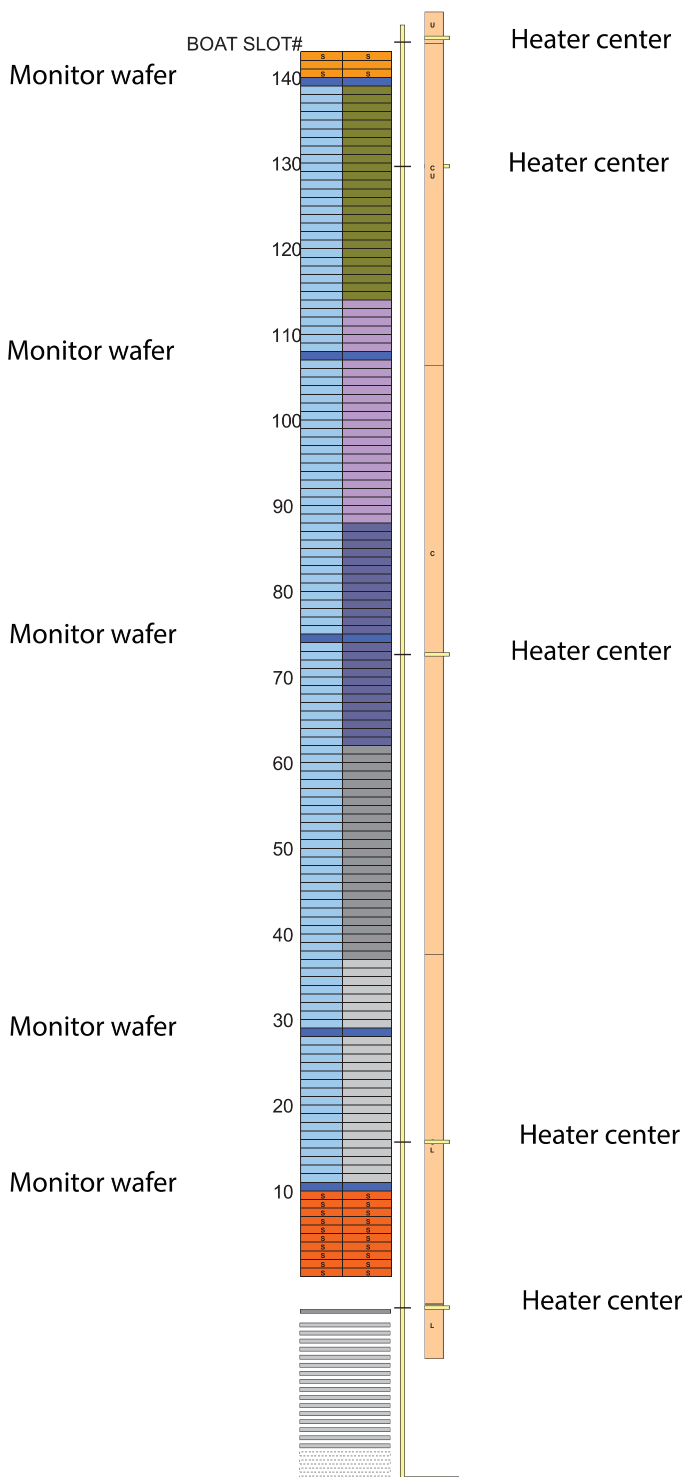


Figure 5.6. Heater center and monitoring wafer location in the boat drawing and their misalignment

5.3 Introduction to Diffusion Furnace R2R Control

5.3.1 Furnace R2R Control Model

Several diffusion R2R control systems have been proposed in the past [122–124]. In this section, we will propose a new furnace R2R controller which handles any number of the input and output combinations.

A typical input and output model of a diffusion furnace without the load size effect [123] is described below,

$$y_k = Mu_k + b_k \tag{5.4}$$

where y_k is the output, which is the thickness metrology of monitor wafers, M is the process gain matrix, which is obtained in the DOE, u_k is a vector of tuning knobs and b_k is a vector of intercept states, in matrix form,

$$y_k = \begin{bmatrix} TopThickness_k \\ TopCenterThickness_k \\ CenterThickness_k \\ BottomCenterThickness_k \\ BottomThickness_k \end{bmatrix} \tag{5.5}$$

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} \end{bmatrix} \tag{5.6}$$

$$u_k = \begin{bmatrix} TopTemperature_k \\ TopCenterTemperature_k \\ DepositionTime \\ BottomCenterTemperature_k \\ BottomTemperature_k \end{bmatrix} \tag{5.7}$$

and

$$b_k = \begin{bmatrix} b1_k \\ b2_k \\ b3_k \\ b4_k \\ b5_k \end{bmatrix} \tag{5.8}$$

The setpoint of "*CenterTemperature*" (from the center heater) of a furnace is usually fixed. In other words, it is not turned by the R2R controller. However the "*DepositionTime*" in (5.7) is added as another turning knob, which replaces the tuning knob of "*CenterTemperature*". Such a control scheme by fixing

"*CenterTemperature*" is relatively stable, because equation (5.4) becomes an "under-determined" case otherwise.

In general, there are three cases [23, 26] for equation (5.4) where $y_k \in \Re^l$ and $u_k \in \Re^n$:

- when $l = n$, a unique solution exists, also known as the exact case.
- when $l > n$, no exact solution exists, but there is still a least square solution, also known as the "over-determined" case.
- when $l < n$, there is not a unique solution, also known as the "under-determined" case.

The diffusion controller developed a few years ago is called a generic diffusion R2R controller, which solves all three cases above. This approach has been used to increase the R2R deployment pace and to minimize any negative impacts in the worldwide Micron Fab network [23, 24].

5.3.2 State Space Representation

A linear discrete state space model [30, 34, 85] used in the generic diffusion R2R controller is described below,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + F\omega_k \\ y_k &= Cx_k + \nu_k \end{aligned} \tag{5.9}$$

where x_k is the state vector, ω_k is the state noise and ν_k is the measurement noise. A is the state matrix, B is the input matrix, F is the state noise matrix and C is the output matrix.

For example, if we define state vector $x_k \in \Re^{10}$ as following [85],

$$x_k = \begin{bmatrix} Mu_k & b_k \end{bmatrix}^T \tag{5.10}$$

and assuming $u_{k+1} = u_k$, then the below matrices are obtained:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.11}$$

$$B = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.12}$$

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.13}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.14}$$

To build a more generic control system, we define a control system with a larger dimension in the actual implementation, where $y_k \in \Re^{10}$, $u_k \in \Re^{10}$ and

$x_k \in \Re^{20}$, which can handle up to 10 inputs and 10 outputs (10×10) in control system dimension. In the next section, we will discuss how to solve a "smaller" dimensional system, for example 3×3, in a generic diffusion controller.

5.3.3 State Estimation of Furnace R2R Control

For the state vector defined in (5.10), in fact, only intercept states b_k needs to be estimated. There is no state noise of the "Mu_k" term defined by the state noise matrix F. On the other hand, we do not recommend estimating parameters in the process gain matrix M, which should be obtained through the DOE, due to the lack of excitement of data during normal operation.

The state estimation is done by quadratic programming [34, 37] with the below objective function:

$$\min_{\omega_k, \nu_k} \left(J = \sum_{k=-1}^{h-1} \omega_k' Q \omega_k + \sum_{k=0}^{h} \nu_k' R \nu_k \right) \tag{5.15}$$

subject to the following constraints:

$$\begin{aligned} x_0 &= \bar{x}_0 + w_{-1} \\ x_{k+1} &= Ax_k + Bu_k + F\omega_k \\ y_k &= Cx_k + \nu_k \end{aligned} \tag{5.16}$$

Q and R are configurable weighting matrices used as part of the optimization cost function. The optimization function allows states to be calculated such that state noise ω_k and measurement noise ν_k are minimized while remaining within established model constraints. The states can be estimated based on historical run data using the control model and the state estimation optimization equations (5.15) (5.16), where h is horizon length and $\bar{x}_0$ is the initial state.

The objective function J can be transformed into the following form through algebraic manipulation [34]:

$$\begin{aligned} \min_{W_k} J &= \frac{1}{2} W_k^T H W_k + f^T W_k \\ W_{k,min} &\leq W_k \leq W_{k,max} \end{aligned} \tag{5.17}$$

where W_k is the state error vector, $H = \tilde{Q} + M_A^T \tilde{R} M_A$ and $f = (-2Y^T \tilde{R} M_A)^T$. For an example, when horizon length $h = 3$, the below matrices are obtained,

$$W = \begin{bmatrix} \omega_{-1} & \omega_0 & \omega_1 & \omega_2 \end{bmatrix}^T \tag{5.18}$$

$$\tilde{Q} = \begin{bmatrix} Q & 0 & 0 & 0 \\ 0 & Q & 0 & 0 \\ 0 & 0 & Q & 0 \\ 0 & 0 & 0 & Q \end{bmatrix} \tag{5.19}$$

$$\tilde{R} = \begin{bmatrix} R & 0 & 0 & 0 \\ 0 & R & 0 & 0 \\ 0 & 0 & R & 0 \\ 0 & 0 & 0 & R \end{bmatrix} \tag{5.20}$$

$$M_A = \begin{bmatrix} C & 0 & 0 & 0 \\ CA & C & 0 & 0 \\ CA^2 & CA & C & 0 \\ CA^3 & CA^2 & CA & C \end{bmatrix} \tag{5.21}$$

$$Y = y - M_B u - M_C \bar{x}_0 \tag{5.22}$$

$$y = \begin{bmatrix} y_0 & y_1 & y_2 & y_3 \end{bmatrix}^T \tag{5.23}$$

$$u = \begin{bmatrix} 0 & u_0 & u_1 & u_2 \end{bmatrix}^T \tag{5.24}$$

$$M_B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & CB & 0 & 0 \\ 0 & CAB & CB & 0 \\ 0 & CA^2B & CAB & CB \end{bmatrix} \tag{5.25}$$

$$M_C = \begin{bmatrix} C \\ CA \\ CA^2 \\ CA^3 \end{bmatrix} \tag{5.26}$$

A state noise vector W_k is obtained by solving this objective function (5.15) though quadratic programming (5.17) and W_k is defined as,

$$W_k = \begin{bmatrix} \omega_{k-h} & \omega_{k-h+1} & \cdots & \omega_k \end{bmatrix}^T \tag{5.27}$$

where h is the horizon length.

After the state noise vector W_k is obtained, all states in the moving horizon can be updated accordingly. It is also observed that the dimension of the above matrices is proportional to the horizon length: the longer the horizon length, the bigger the dimension of matrices, (5.18) to (5.26). Therefore, an increased computational cost has been observed by extending the horizon length h.

Typically, a diffusion R2R control has five inputs and five outputs, which is used in example (5.11) to (5.14). Since the diffusion controller is a generic control system, sometimes the number of inputs and outputs can be less than five. In this case, one can simply reduce the rank of Q and R to be equal to the number of actual outputs. In illustration, for a three inputs and three outputs control problem (3×3), Q and R are set as following for the state estimation:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.28}$$

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.29}$$

In summary, the state estimation of any combination of multiple inputs and multiple outputs (MIMO) problems up to a maximum dimension specified in the control system, (10×10) in our case, is solvable in the generic diffusion R2R controller.

5.3.4 Furnace R2R Control and Its Performance

The generic diffusion R2R controller was coded on the E3 platform from Applied Material, and the controller type is model predictive control (MPC) [125] outlined in

Figure 5.7: batches "$k-4$" through "k" have been processed in the furnace reactor and have post-metrology measurements. The filter horizon consists of a fixed number historical runs which is used to estimate the states. Batch "$k+1$" has not yet been processed, which constitutes the prediction horizon. The recommended setting u_{k+1} for the next batch without constraints is computed as following,

$$u_{k+1} = \frac{y_t - \hat{b}_{k+1}}{M} \tag{5.30}$$

where y_t is the control target and $\hat{b}_{k+1}$ is the estimation of intercept states. On the other hand, u_{k+1} can be also computed with respect to tuning knob's constraint through equation (1.13).

Excellent process capability improvements in terms of C_{pk} have been obtained after deploying such generic diffusion R2R controller. For instance, over 90% C_{pk} improvement is obtained for a poly-silicon deposition process in Figure 5.8, and the model dimension is a six inputs and nine outputs system, where the inputs include all five-zone heater temperatures and a deposition time. On the other hand, the outputs are the thickness measurements of monitor wafers from nine different furnace positions, or boat slots. The increased number of monitor wafers compared with five monitor wafers is part of the effort in flattening the thickness profile of a diffusion furnace.

Overall, we have seen an average of 40% C_{pk} improvement for all diffusion processes, besides other benefits including increased R2R deployment pace and reduced R2R controller excursions related to the controller development.

5.4 Motivations for VM of Diffusion Furnace

The thickness profile of a vertical diffusion furnace often introduces variations into process steps downstream. In the last section, we discussed the diffusion R2R

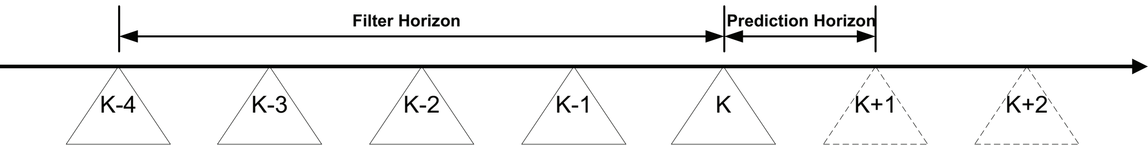


Figure 5.7. The filter horizon and predictive horizon for model predictive control.

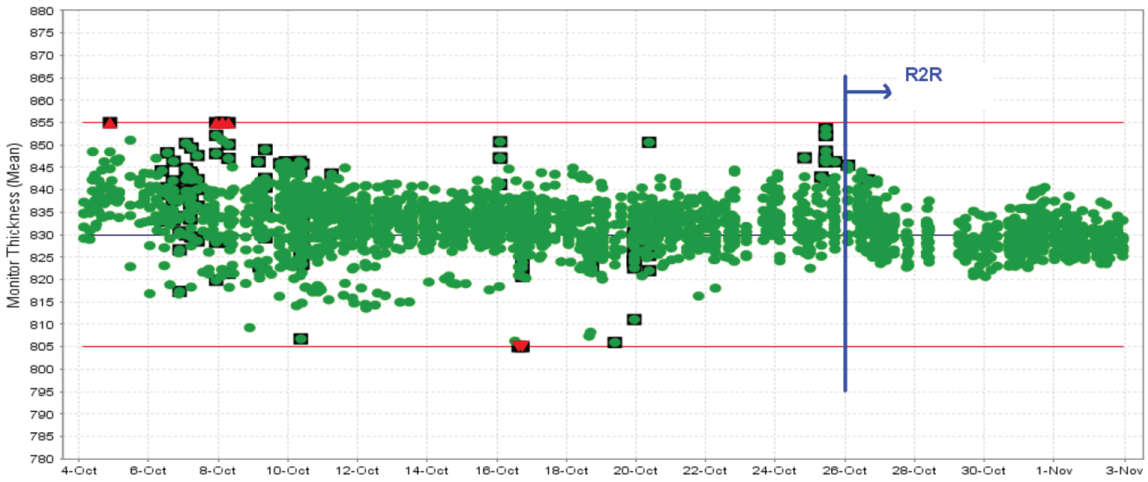


Figure 5.8. A furnace R2R control (6×9) improved process capability in terms of C_{pk} by more than 90%.

control which is used to improve the thickness profile of a vertical diffusion furnace. In the example of Figure 5.8, more monitor wafers (nine vs. five) are used to flatten the ploy-silicon deposition thickness profile, while using more monitor wafers means more manufacturing cost and long cycle time (almost doubled metrology time). On the other hand, the thickness profile in each heater zone or across all five heater zones cannot be completely eliminated by an R2R controller because of the temperature uniformity problems [126] and a depletion effect [127].

Our motivations for doing this research project include the following:

- Improve the thickness profile prediction through multiphysics model, so that it can enable wafer level compensation at the downstream process steps.
- Provide wafer level thickness data of all wafers in a diffusion batch for quality control and analysis.
- Understand the challenge and roadblock of this diffusion VM project and look for future research topics.

5.5 Incorporating Multiphysics into Furnace VM

5.5.1 Background

A method of predicting a diffusion furnace profile through neural networks was proposed by Bode and Toprac [127] and the thickness profile can be compensated

by an etch R2R controller in the downstream. They pointed out that deposition rate by a diffusion furnace is affected by normal variations in temperature, reactant flow rate and gases depletion, as shown in Figure 5.5. The gases enter the reactor from the bottom, then they travel upward around the process wafers. Therefore, it is possible that a decreased gas concentration occurs at an increased distance from the gases inlet. Such a phenomenon is called the depletion effect, and to overcome the gas depletion problem, the furnace is divided into four or five independently controlled heater zones, as seen in Figure 5.6. Most of the diffusion furnaces have five heater zones including top, top center, center, bottom center and bottom. For example, if the gas concentration is higher in the bottom of the reactor, then the temperature setpoint of the bottom zone can be lowered to slow down the deposition rate in the bottom zone.

Since the reaction rate is related to both temperature and gas concentrations, it is important to understand the temperature distributions of a furnace. Hirasawa et al. [128] analyzed the heat transfer mechanisms and the steady state temperature distribution of a vertical furnace. The gas velocity in the reactor is very slow, in the order of 0.01 m/s, so convective heat transfer can be negligible. The tube in Figure 5.4 is made from silicon carbide or quartz, and at steady state, temperature across the tube is constant, in that it is very thin, and the gap between the heating coil and the tube is very small, so the effect of the tube can be negated too. Therefore, the radiation dominates the heat transfer in steady state of a diffusion furnace.

In the steady state condition of a furnace without any wafers loaded, the radiative heat transfer is given by (5.31) (5.32),

$$q_i = -G_i + \sum_{k=1}^{m} G_k F_{ik} \tag{5.31}$$

$$G_i = \epsilon_i \sigma T_i^4 + (1-\epsilon_i) \sum_{k=1}^{m} G_k F_{ik} \tag{5.32}$$

where q_i is the net radiative heat flux absorbed in the i element, G_i is the radiosity of the i element, ϵ_i is emissivity, σ the Stefan-Boltzmann constant and F_{ik} is the configuration factor, which is described in [129].

The steady state temperature distribution [129] for a furnace without any wafer is,

$$T_i = (\sum_{k=1}^{m} G_k F_{ik}/\sigma)^{0.25} \tag{5.33}$$

where G_k is the radiosity of the k element and T_i is the steady state temperature distribution of a small sphere.

In the unsteady condition, the conduction also contributes to the heat transfer. The unsteady thermal conduction is described as [128]:

$$\rho C_p \frac{\partial T}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}(\lambda_1 r \frac{\partial T}{\partial r}) + \frac{\partial}{\partial z}(\lambda_2 \frac{\partial T}{\partial z}) + Q \tag{5.34}$$

where T is temperature, r and z are radial and axial coordinates of the furnace, ρ is density, C_p is specific heat, λ_1 and λ_2 are the radial and the axial thermal conductivity, and Q is the sum of the radiative heat absorbed and the heat generated.

Although heat transfer mechanism is important to understand the temperature distributions, the fundamental models (5.33) (5.34) were not used for the VM model in this research, because we are more interested in the thickness profile along the z direction. It is also difficult to obtain accurate model parameters, such as λ_1 and λ_2.

5.5.2 Design of Experiment

A DOE is created to understand the contributions of the deposition rate of each tuning knob. The DOE data were collected with the help of the diffusion process engineers, and they are plotted in Figure 5.9.

Figure 5.9 consists of six subplots, each corresponding to one of the six tuning knobs:

- Deposition Time: the deposition rate is about 0.07 Å/S across all boat slots.
- Top Temperature: the peak of deposition rate change by temperature is located at the heater center, boat slot 145. The shape is not very clear because it is one of the end zones of the furnace. It may be a "bell" shape curve but with a very limited confidence.

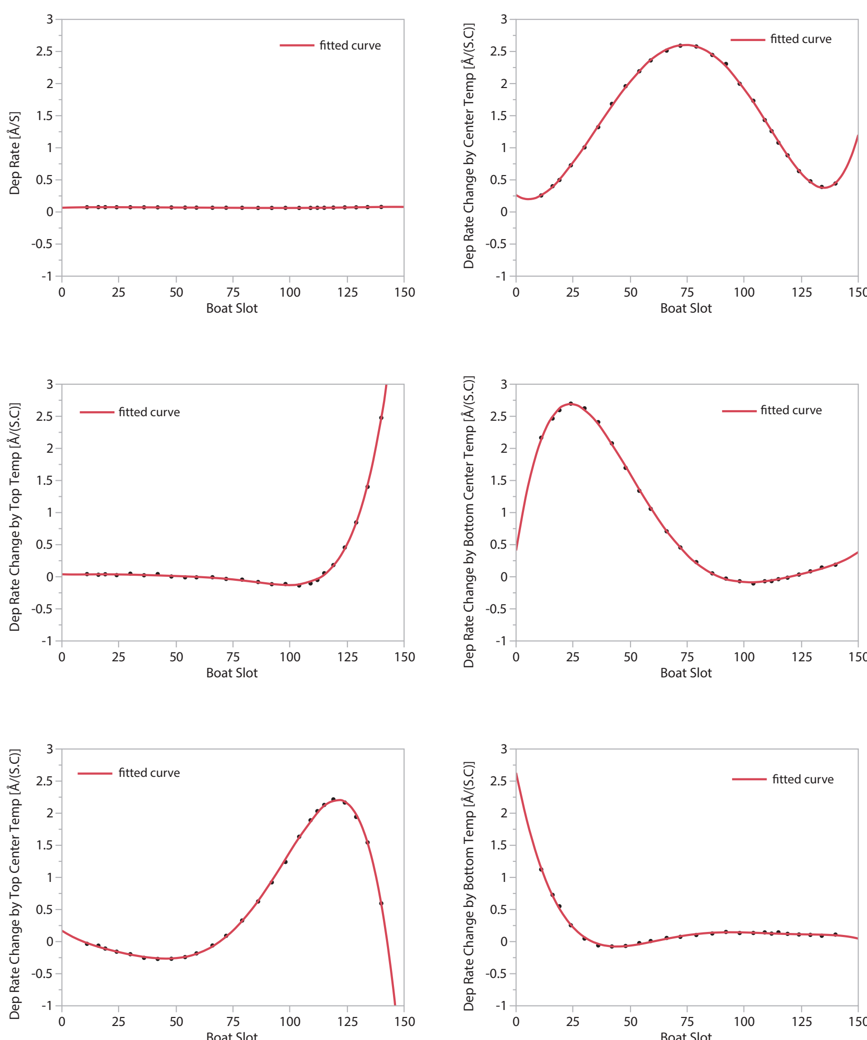


Figure 5.9. Design of experiment: deposition rate change by tuning knob at each boat slot (or furnace position)

- Top Center Temperature: the peak of deposition rate change by temperature is near the heater center, boat slot 130. The shape is close to "bell" shape for the left-hand side, but there is not enough data to justify the shape for the right-hand side of curve.
- Center Temperature: the peak of deposition rate change by temperature is located at the heater center, boat slot 73, and the shape of the curve is clearly a "bell" shape.
- Bottom Center Temperature: it is similar to "Top Center Temp", the peak of deposition rate change by temperature is near the heater center, boat slot 16.
- Bottom Temperature: it is similar to "Top temp", which is the other end of the boat.

After analyzing above DOE data, we give a hypothesis that the furnace temperature profile would be a combination effect of five Gaussian curves, one for each heater zone, and an intercept term contributed by the "deposition time" knob.

5.5.3 Curve Fitting Results

This hypothesis inferred by the DOE data is supported by the curve fitting results in Matlab. Figure 5.10 shows the thickness profile evaluations by adding on the heater zone contributions one by one. For illustration purposes, the upper graph of each subplot is the accumulated effect from all contributions, and the lower graph of each subplot are individual knob contributions.

Figure 5.11 is the overlay between fitting curve and the actual metrology, and the curve fitting is done by the curve fitting toolbox in Matlab. The general equation used in the curve fitting is,

$$y = x_t e^{\frac{-(p-p_t)^2}{2\sigma_t^2}} + x_{tc} e^{\frac{-(p-p_{tc})^2}{2\sigma_{tc}^2}} + x_c e^{\frac{-(p-p_c)^2}{2\sigma_c^2}} + x_{bc} e^{\frac{-(p-p_{bc})^2}{2\sigma_{bc}^2}} + x_b e^{\frac{-(p-p_b)^2}{2\sigma_b^2}} + x_{int} \tag{5.35}$$

where y is the thickness metrology; p is the boat slot of the measured wafer; x_t, x_{tc}, x_c, x_{bc} and x_b are the peak values of each Gaussian curve corresponding to the heaters from Top, Top Center, Center, Bottom Center and Bottom heater zone,

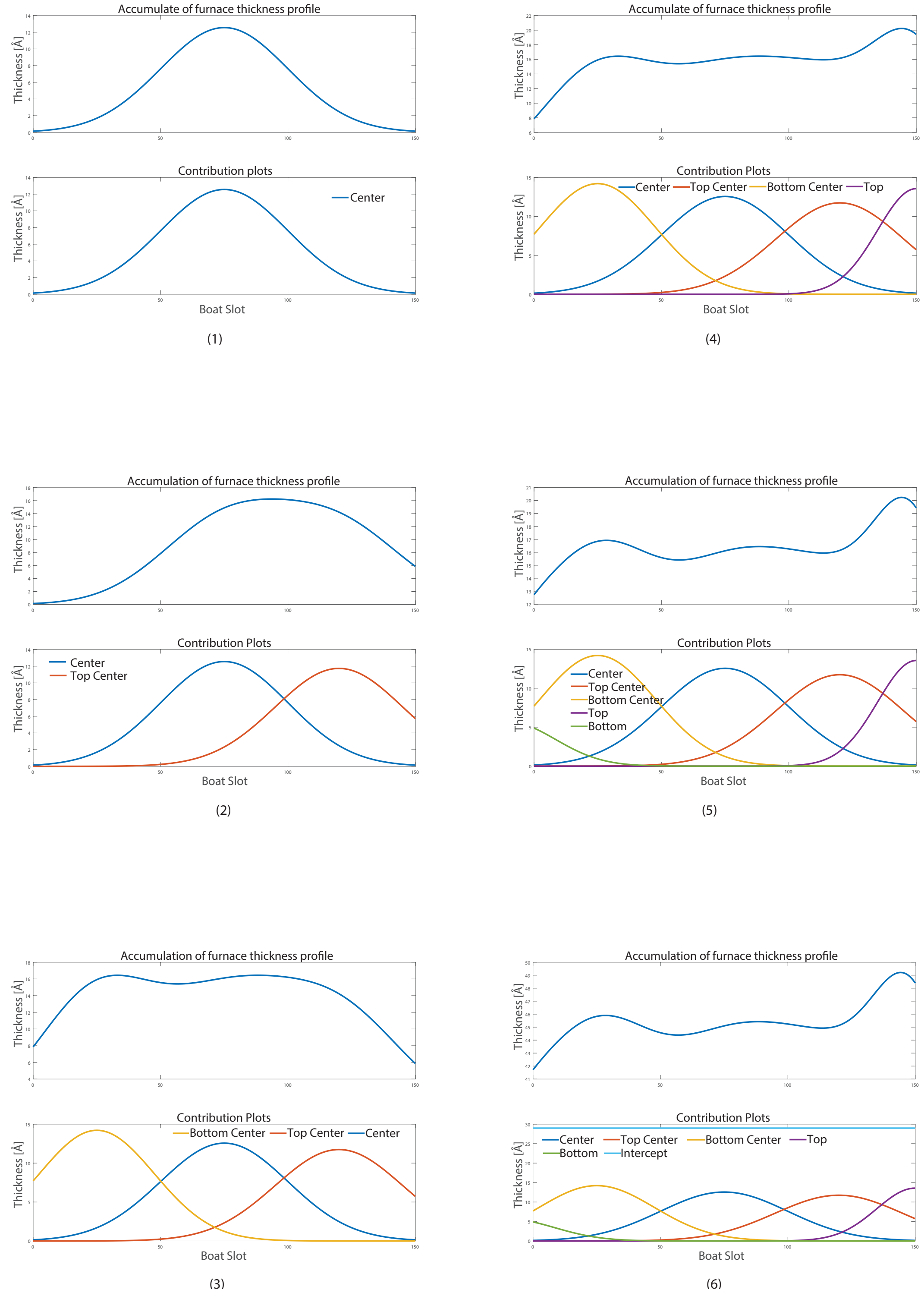


Figure 5.10. Accumulation effect and its contributions of a diffusion furnace thickness profile

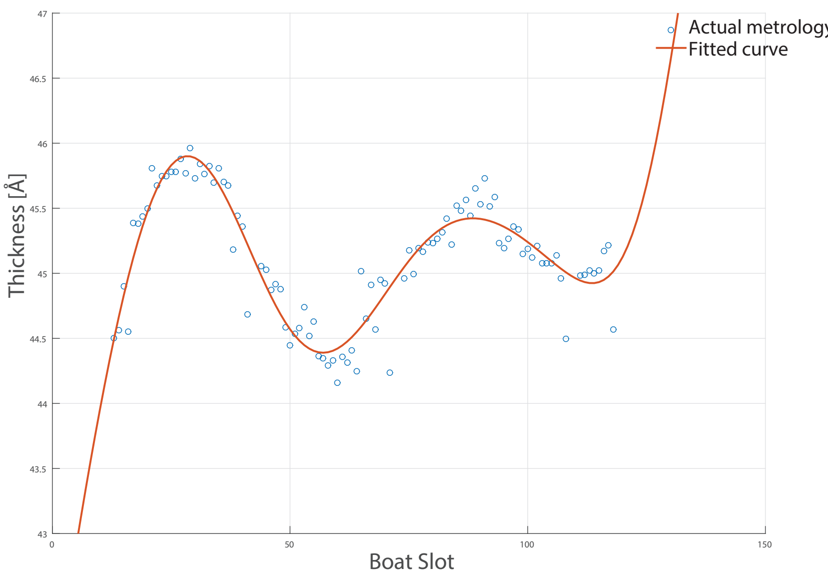


Figure 5.11. Overlay of actual metrology and fitted curve

respectively; σ_t, σ_{tc}, σ_c, σ_{bc} and σ_b are the standard deviations to define the "spread" of each Gaussian function; p_t, p_{tc}, p_c, p_{bc} and p_b are the boat slots of each heater center location; and finally x_{int} is the intercept term.

p_t, p_{tc}, p_c, p_{bc} and p_b are known values in the model (5.35). The inputs to the curve fitting toolbox in Matlab are p and y vectors, and the outputs of the curve fitting are peak values, x_t, x_{tc}, x_c, x_{bc} and x_b, as well as the standard deviations, σ_t, σ_{tc}, σ_c, σ_{bc} and σ_b.

5.5.4 VM Model Update Methods

Two model update methods are proposed for the diffusion furnace VM. If the thickness profile shape does not change over time, then only the intercept term is to be updated, as shown in Figure 5.12.

A new intercept state $x_{int,k+1}$ moves the whole thickness profile up and down without changing the shape of the thickness profile, which is a simpler case. The intercept state can be updated via an EWMA filter after obtaining new metrology data.

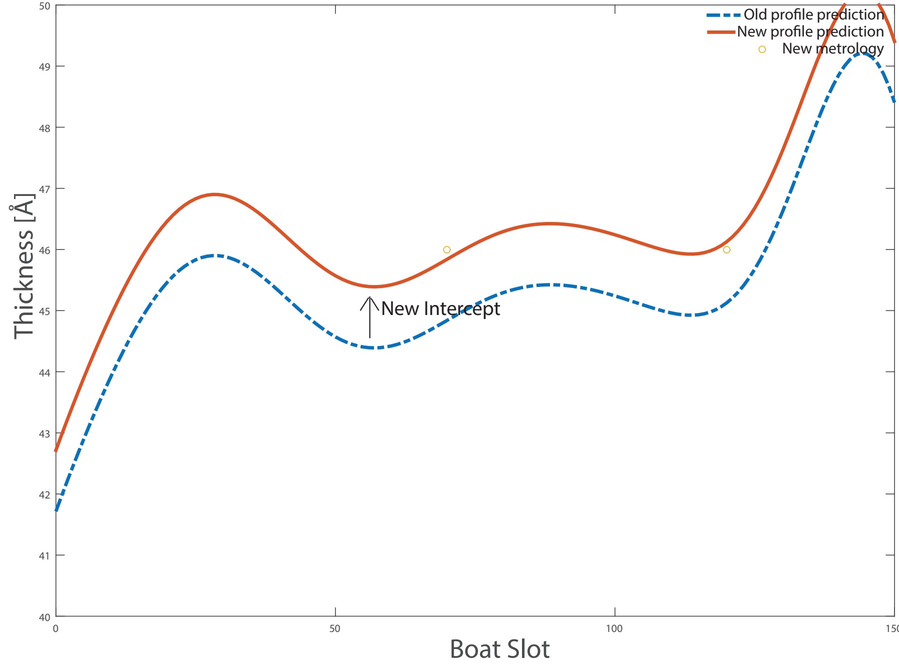


Figure 5.12. Model update by changing intercept state only

$$\begin{aligned}\hat{x}_{int,k+1} =& \lambda(y - x_t e^{\frac{-(p-p_t)^2}{2\sigma_t^2}} + x_{tc} e^{\frac{-(p-p_{tc})^2}{2\sigma_{tc}^2}} + x_c e^{\frac{-(p-p_c)^2}{2\sigma_c^2}} + x_{bc} e^{\frac{-(p-p_{bc})^2}{2\sigma_{bc}^2}} + x_b e^{\frac{-(p-p_b)^2}{2\sigma_b^2}}) \\ &+ (1-\lambda)\hat{x}_{int,k+1}\end{aligned} \tag{5.36}$$

where λ is the EWMA weight. Some assumptions have to be made for such a model update method that the peaks (x_t, x_{tc}, x_c, x_{bc} and x_b) and standard deviations (σ_t, σ_{tc}, σ_c, σ_{bc} and σ_b) are fixed values, and they stay the same as the values obtained at model fitting time. In other words, they do not drift over time. Since most of the diffusion processes are controlled by the R2R controllers, and the setpoint of each heater zone is adjusted by the R2R controller from time to time, this affects the peak values of the Gaussian curves. Therefore, such assumptions would be only correct for those processes without R2R controls.

The other model update method would be better, because we do not have to make such assumptions that the shape of a thickness profile stays the same. VM updates not only the intercept x_{int} but also the peak values (x_t, x_{tc}, x_c, x_{bc} and x_b). The standard deviations (σ_t, σ_{tc}, σ_c, σ_{bc} and σ_b) are assumed to be the fixed values

in order to simplify the system from a nonlinear system to a linear one (5.37):

$$y = \begin{bmatrix} x_t & x_{tc} & x_c & x_{bc} & x_b \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} + x_{int} \tag{5.37}$$

where $c_1 = e^{\frac{-(p-p_t)^2}{2\sigma_t^2}}$, $c_2 = e^{\frac{-(p-p_{tc})^2}{2\sigma_{tc}^2}}$, $c_3 = e^{\frac{-(p-p_c)^2}{2\sigma_c^2}}$, $c_4 = e^{\frac{-(p-p_{bc})^2}{2\sigma_{bc}^2}}$ and $c_5 = e^{\frac{-(p-p_b)^2}{2\sigma_b^2}}$, c_1 to c_5 can be computed as the boat slot of the measured wafer is known and the standard deviations (σ_t, σ_{tc}, σ_c, σ_{bc} and σ_b) are assumed to be constant. The assumption of a fixed standard deviation is reasonable because it is mainly determined by the heaters' design such as the length of the heater.

The model update becomes parameters estimation of x_t, x_{tc}, x_c, x_{bc}, x_b and x_{int}. Assuming that all parameters drift slowly over time in (5.38),

$$\begin{bmatrix} x_{t,k+1} \\ x_{tc,k+1} \\ x_{c,k+1} \\ x_{bc,k+1} \\ x_{bk+1} \\ x_{int,k+1} \end{bmatrix} = \begin{bmatrix} x_{t,k} \\ x_{tc,k} \\ x_{c,k} \\ x_{bc,k} \\ x_{b,k} \\ x_{int,k} \end{bmatrix} + \begin{bmatrix} \omega_{1,k} \\ \omega_{2,k} \\ \omega_{3,k} \\ \omega_{4,k} \\ \omega_{5,k} \\ \omega_{6,k} \end{bmatrix} \tag{5.38}$$

and also the linear system (5.37) can be rigorously transformed into (5.39).

$$y_k = \begin{bmatrix} c_{1,k} & c_{2,k} & c_{3,k} & c_{4,k} & c_{5,k} & 1 \end{bmatrix} \begin{bmatrix} x_{t,k} \\ x_{tc,k} \\ x_{c,k} \\ x_{bc,k} \\ x_{b,k} \\ x_{int,k} \end{bmatrix} \tag{5.39}$$

Therefore, this system can be described in a linear time varying (LTV) state space form [32,33],

$$\begin{aligned} x_{k+1} &= Ax_k + \omega_k \\ y_k &= C_k x_k + v_k \end{aligned} \tag{5.40}$$

and defining $x_k = \begin{bmatrix} x_{t,k} & x_{tc,k} & x_{c,k} & x_{bc,k} & x_{b,k} & x_{int,k} \end{bmatrix}^T$, below system matrices of state space model can be obtained,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.41}$$

$$C_k = \begin{bmatrix} c_{1,k} & c_{2,k} & c_{3,k} & c_{4,k} & c_{5,k} & 1 \end{bmatrix} \tag{5.42}$$

The states vector x_k, can be estimated by equations (5.15) (5.17), so this concludes the second method of VM model update.

5.5.5 Queue Time and Metrology

One of the challenges on this project is that ploy thickness growth begins when the FOUP is open. Figure 5.13 shows that native oxide grows much faster at the beginning, especially in the first 50 minutes, and then it settles down after 200 minutes. This problem is not observed at the diffusion furnace step, because the nitrogen purging is equipped at the load port of the furnace, while the metrology tool does not have a such nitrogen purge, where this queue time problem is discovered.

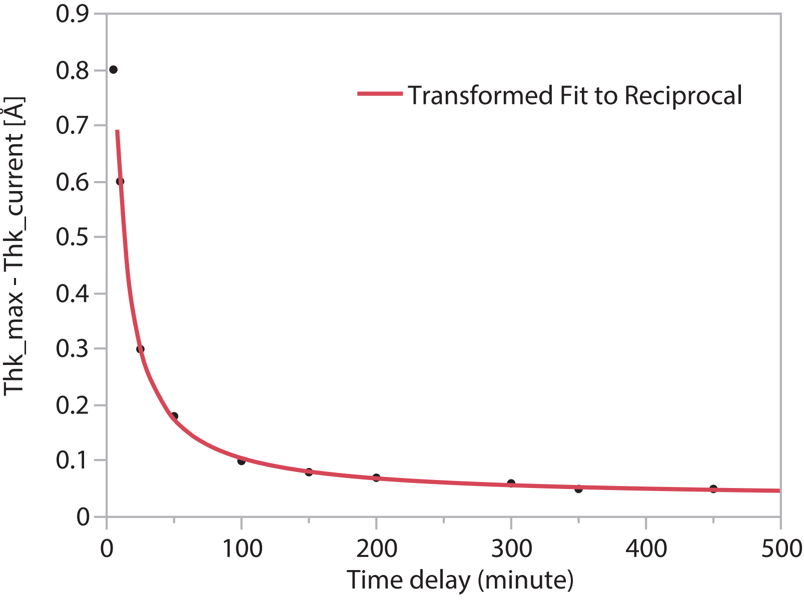


Figure 5.13. Metrology queue time

Exposing the wafers to air for two hours and then measuring seems to be a solution to this metrology queue time problem, but this solution increases the cycle time of the line and wastes metrology tool utilization. A better solution would "offset" the metrology data by accounting for how long it takes to measure the wafers by a metrology tool and the offset can be calculated by the model established in Figure 5.13.

5.5.6 Effects of R2R Control and Its Compensation

The other challenge in this project is how to account for R2R control adjustments. As shown in Figure 5.14, there is a gap between the two predictions at different times: one profile is predicted at the diffusion step and the other one is done at the post-metrology time. The prediction gap is from the R2R control adjustments, because the R2R controller makes the recipe adjustments as soon as it obtains the new metrology data. The VM strategy estimates new model parameters toward the thickness profile "inferred" by the metrology data before the R2R adjustment, while the thickness profile "deviates" from the metrology data after the R2R adjustments. Therefore, the R2R adjustments need to be accounted for when making predictions

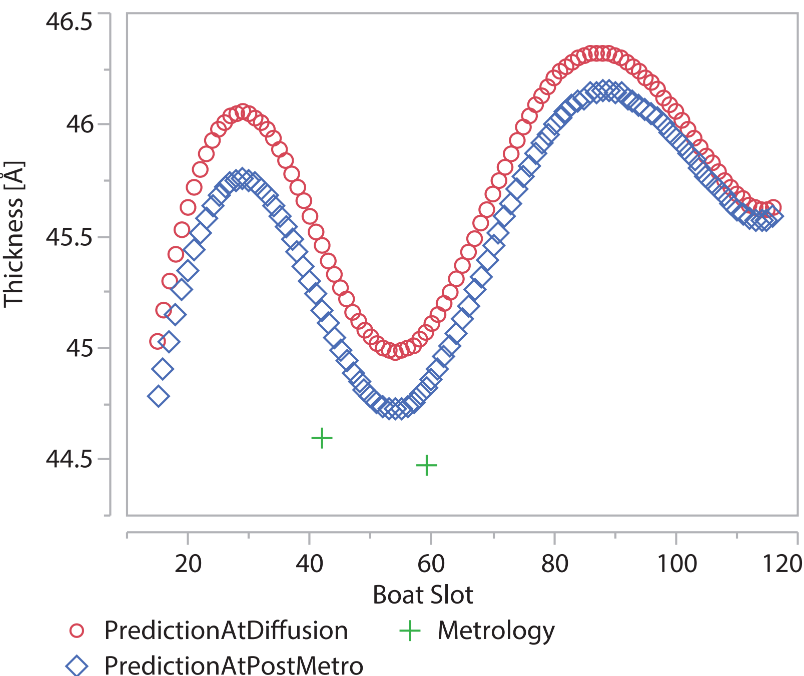


Figure 5.14. VM prediction gap at different times, furnace step vs. metrology step

for the next batch, when it's tracked out at the diffusion step. This would be especially useful when the post diffusion metrology sampling rate is not 100%.

The R2R effects can be accounted for by the peak values and the intercept adjustments. The peak values (of Gaussian curves) are mainly impacted by temperature knob adjustments, while the intercept is changed by the deposition time adjustment of the R2R control. The adjusted peak values and intercept can be estimated as the following,

$$x_{adjusted} = \Theta \triangle u + x_{current} \tag{5.43}$$

where $x_{adjusted}$ are the new peak values and new intercept after R2R adjustment, Θ is the gain matrix which can be obtained by a DOE, $\triangle u$ is the process tuning knob adjustment of R2R control and $x_{current}$ is the latest state estimation by the VM. In matrix form,

$$\begin{bmatrix} x_{t,adjusted} \\ x_{tc,adjusted} \\ x_{c,adjusted} \\ x_{bc,adjusted} \\ x_{b,adjusted} \\ x_{int,adjusted} \end{bmatrix} = \begin{bmatrix} \theta_t & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{tc} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_c & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{bc} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_b & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{int} \end{bmatrix} \begin{bmatrix} \triangle u_t \\ \triangle u_{tc} \\ \triangle u_c \\ \triangle u_{bc} \\ \triangle u_b \\ \triangle u_{time} \end{bmatrix} + \begin{bmatrix} x_{t,current} \\ x_{tc,current} \\ x_{c,current} \\ x_{bc,current} \\ x_{b,current} \\ x_{int,current} \end{bmatrix} \tag{5.44}$$

where θ_{int} is the process gain for the deposition time, which can be easily estimated from the R2R gain matrix in (5.6):

$$\theta_{int} = \frac{m_{13} + m_{23} + m_{23} + m_{43} + m_{53}}{5} \tag{5.45}$$

while θ_t, θ_{tc}, θ_c, θ_c, θ_{bc} and θ_b are the peak gains of temperature knobs, which may not be readily available in the R2R gain matrix depending on the heaters center locations and the monitor wafers' locations. However, a DOE in Figure 5.9 is very helpful to estimate them.

After compensating the R2R adjustments, an improvement on prediction gap between the furnace step and the metrology step has been observed as shown in Figure 5.15.

5.6 VM Results and Conclusions

Typically, only five wafers are measured at the metrology step post poly deposition of a diffusion step. However, more wafers can be measured to validate

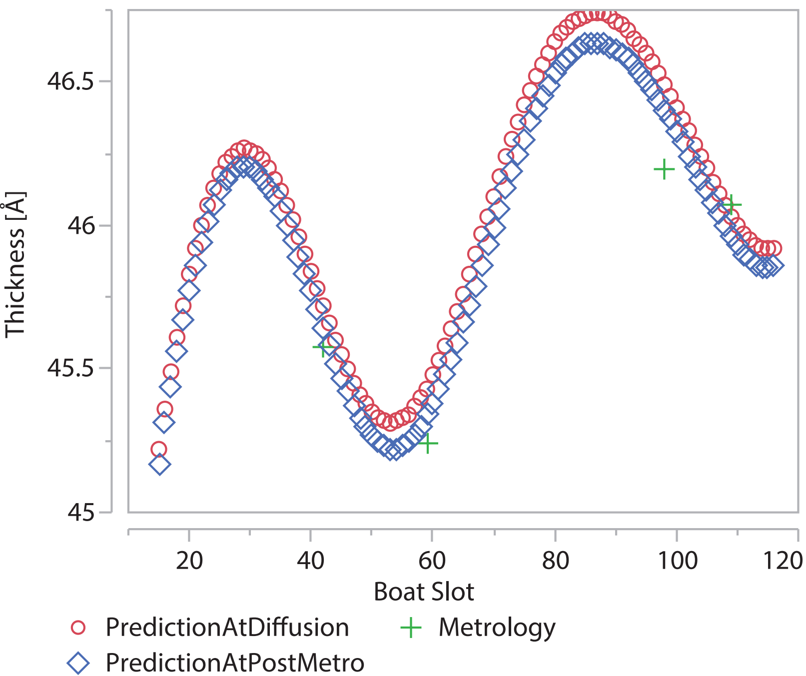


Figure 5.15. VM prediction gap at different times is improved between furnace step and metrology step

the predicted profile of a diffusion batch. Figure 5.16 is the overlay plot of actual metrology and VM prediction at the furnace step with queue time compensation, and in the case that the wafers of every other boat slot are measured. The correlation is very good with $r^2 = 0.72$, which is shown in Figure 5.17.

We demonstrated the diffusion VM capabilities in one of the critical diffusion processes, which has four Angstroms for the control window and only eight Angstroms for the specification window in the SPC control chart. Quite a few queue time related difficulties were discovered. One is related to the metrology tool without the nitrogen purge and the other one is related to the outgassing effect after a special dry etch step.

The other use case of this VM system is that the furnace prediction profile is used for the wafer level reliability analysis by quality engineers. For example, if a wafer in slot 16 is measured below the lower spec limit on the SPC chart, then the wafer in slot 15 would be also below the lower spec limit, because the wafer in boat slot 16 is thicker than that of boat slot 15, according to the thickness profile in Figure 5.15.

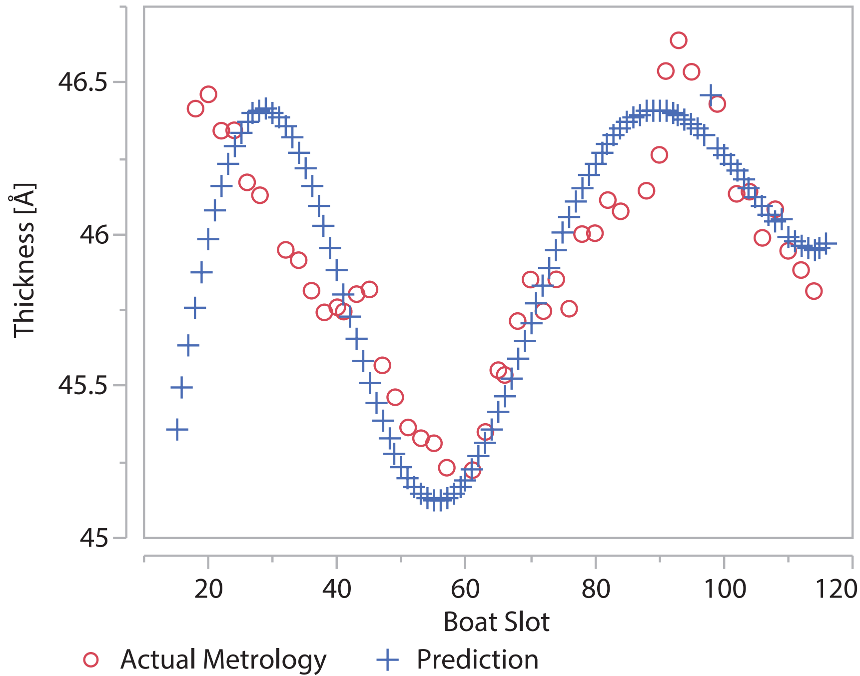


Figure 5.16. The overlay plot of actual metrology and VM prediction at furnace step

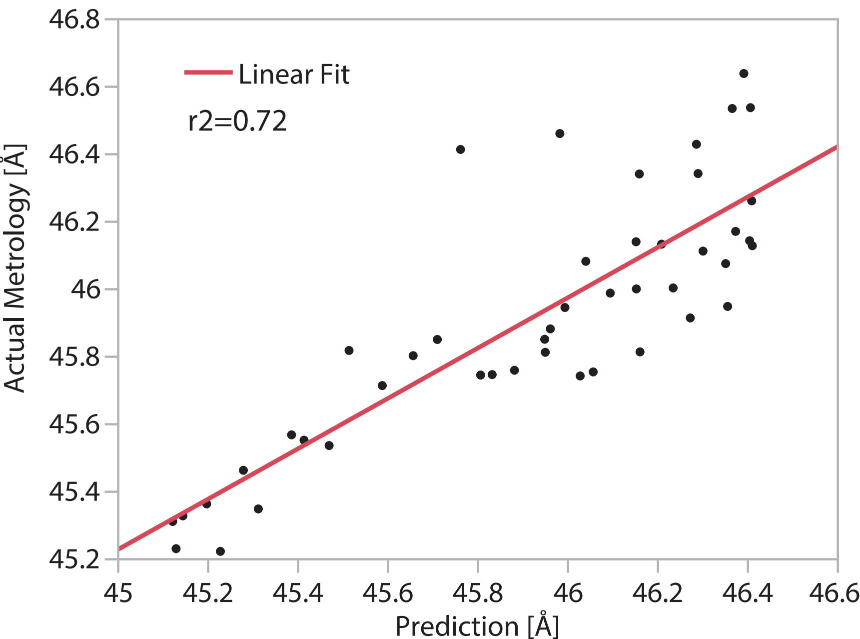


Figure 5.17. The correlation of actual metrology and VM prediction at furnace step

5.7 Summary and Future Work

The diffusion furnace processes wafers in a batch, typically five to six lots. The thickness profile along the boat slots exists because of the heaters' design and gas depletions. A neural network method was proposed to predict such thickness profile in the past, while we incorporate the physics insight, equipment knowledge and the experiments to develop a multiphysics model, which consists of five Gaussian curves and one intercept term, to predict the thickness profile. The model parameters are updated as soon as new metrology data become available. Furthermore, the R2R control adjustment can be accounted for to improve the prediction accuracy, and we propose a method to compensate the R2R adjustments. On the other hand, there are some challenges on building a reliable furnace VM model, such as the queue time effect of the metrology tool without nitrogen purge, which impacts the accuracy of the actual metrology data. Excellent prediction results were obtained in such multiphysical VM models after overcoming those challenges.

Since measuring all wafers in a batch is not feasible due to metrology constraints, future work should include feeding forward VM data to a wafer level R2R controller at downstream process steps. The benefit of doing this is that wafer level variations caused by the furnace position can be reduced or removed. For example, the operating recipe at the dry etch step can be modified based on the predicted thickness profile model to prevent over-etch or under-etch, which was proposed in the literature. We propose to extend this methodology to other process steps, such as the ion implantation step. In ion implant processes mask layer thickness can block ion penetration [118], so the film thickness profile of a diffusion process, oxide thickness or poly-silicon thickness can influence the poly-silicon resistance [14]. A tuning knob such as implant dose would help improve such wafer level variations.

Other future work would be to continuously improve the metrology accuracy, such as the queue time related problem and other incoming variations, so that the VM model can be "calibrated" better by a reliable metrology data source.

CHAPTER 6

THE OXYGEN PLASMA RESIST DESCUM VIRTUAL METROLOGY

6.1 Abstract

The resist descum step is a partial resist stripping step where residuals can be cleaned up. The photo resist residuals often cause yield loss because they block the following dry etch step. To prevent blocked etch or over etch of photo resist, the etch rate of resist descum is "monitored" by the etch rate qualifications, which could be performed a few days apart. In high-volume semiconductor manufacturing, thousands of wafers can be processed during the time window between the two etch rate qualification runs, and the risk of excursion due to poor etch rate is very high. Increasing the frequency of etch rate qualifications would impact the availability of production tools and increase the usage of non-product wafers. Therefore, it increases the manufacturing cost. A virtual metrology system is proposed to monitor the etch rate of all product wafers to mitigate the risk of blocked etch. In this research, we start with the background of the etch rate mechanism and use the insights about chemical reactions to select model parameters. The traditional PLS model is tested first, and then we propose a new method which conducts the "Zonal" data analysis to improve the prediction quality of the PLS. We demonstrate that the multiphysics-based model outperforms the PLS model through adding new process indicators which use the knowledge of chemical reactions. Finally, we conclude the discussions with the VM challenges, their potential solutions and the future work that might be done in this area.

6.2 Introduction

6.2.1 Introduction to Plasma Dry Etching

Wet chemical etch was used much earlier than a plasma etch in semiconductor manufacturing. However, it has now been largely replaced by the plasma etch, or dry etch, for two reasons [118]:

- The plasma process often generates very reactive species, and these species can etch more vigorously than those species in the nonplasma environment.
- The other important reason why plasma etch outperforms wet etch is that directional or anisotropic etching becomes possible with the help of plasma. Anisotropic etch is required to minimize underetching and etch bias which enables smaller and high density structures.

A simple plasma reactor is described in Figure 6.1: it consists of opposed parallel plate electrodes in a chamber that is pumped down to a low pressure, which is typically between 0.01 Torr to 1 Torr. After applying high frequency voltages between the two electrodes, usually 100 to 1000 volts for the driving voltage and 13.56 MHz for the driving frequency [130], a plasma is formed by current, which emits a characteristic glow. At the same time, reactive species are generated by this electrical discharge.

The plasma is an ionized gas with equal numbers of positive and negative charges, although the extent of ionization is very small, only one charged particle released per one million neutral molecules or atoms. The majority of negative charged particles are free electrons. Although the energy transfer from electrons to gas molecules is inefficient, electrons can obtain an elevated energy, typically many electron volts, in plasma. The increased temperature of electrons drives the collisions between electrons and molecules to form radicals in low temperature neutral gas. It will require very high temperatures, e.g., 1000^oK to 10000^oK, to create the same amount of reactive species without a plasma.

Electrons are light with very high mobility, so they diffuse the fastest and recombine with charged particles near the walls or boundary surface. A sheath (or a thin boundary layer) is then formed when such diffusions occur. The depletion

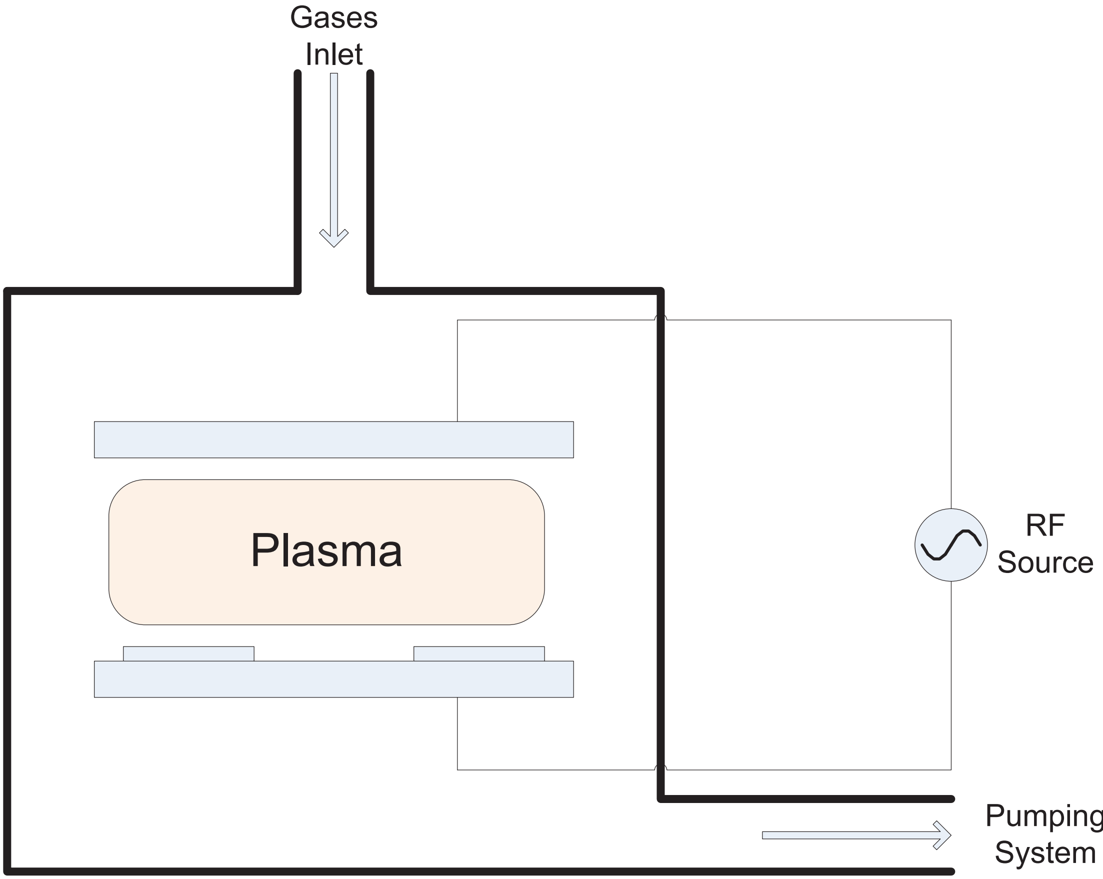


Figure 6.1. A simple plasma reactor

of electrons in the sheath leaves a positive potential to the walls. The positive ions are accelerated through the sheath potential and they strike the walls as the realization of ion bombardments. Three types of plasma etching include physical sputtering, chemical reaction and ion-assisted etching. In this research, O_2 plasma downstream etching is used, and sometimes it is also called chemical dry etching.

6.2.2 Introduction to O_2 Plasma Photo Resist Etching

Oxygen plasmas are used to etch photo resist isotropically, as shown in Figure 6.2, and such isotropically resist etch is extensively used in plasma etch applications of semiconductor manufacturing [8]. In the stripping process of photo resist, the plasma can cause some negative impacts. For example, impinging ions can cause electrical damage of devices, or the charging introduced by plasma induced potential can cause electrostatic punch through of the thin gate oxide. To solve the

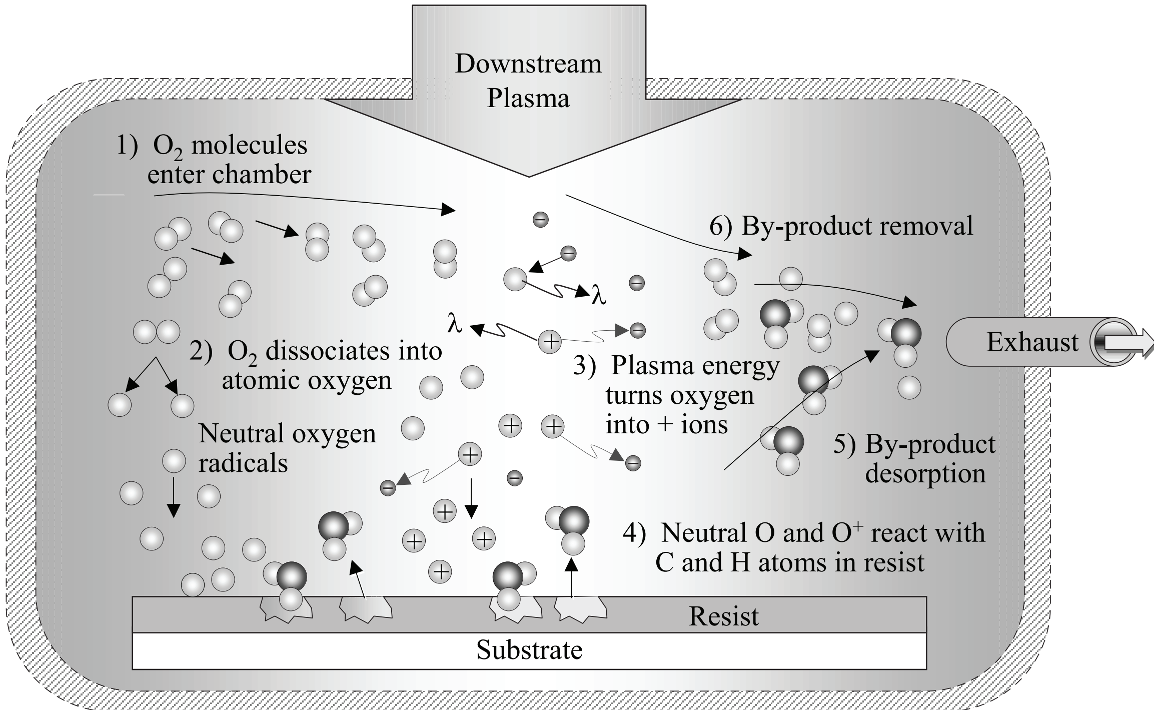


Figure 6.2. O2 plasma reactions with photo resist [7]

above-mentioned problems, downstream plasma is often used in the resist descum process. In the O_2 downstream plasma stripping, gas from the plasma source flows downstream into the chamber in a way that only neutral O and O_2 can get to the resist covered substrate. The reason for this phenomenon is that the charged species, electrons or positive O_2^+, have a much higher loss rate compared with the neutrals when plasma excitation is taken away [131]. The ideal downstream plasma etch is a spontaneous chemical reaction, which does not require ion, electron and photon bombardment.

Chamber pressure also plays an important role for the isotopical etch of the resist descum process [8]. Figure 6.3 illustrates that at pressure below 0.1 Torr, the characteristic potentials across the sheath and the voltage applied to the discharge increase significantly. Therefore, the physical sputter rate increases rapidly. When chamber pressure is increased to 1.0 Torr, the etching process becomes isotropical chemical reaction. However, most plasma etch is operated between these two extremes, which is ion-assisted reactive etch. The chamber pressure of a descum process is close to 1.0 Torr, and that is the reason why it is an isotropical etch.

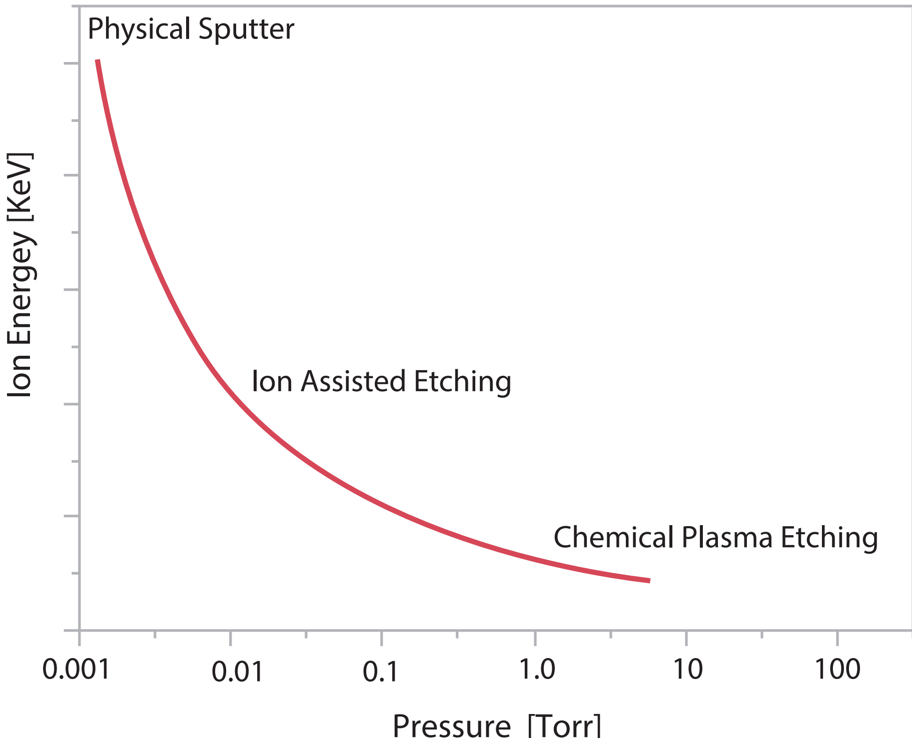


Figure 6.3. The chamber pressure influences the plasma etching types [8]

Compared with direct plasma etch, the downstream plasma etch is more difficult to control because the importance of wall collisions and recombinations in determining the composition of a species in the flux is vastly increased and the factors that control radical recombination at walls and the interaction between different species are not well understood [131].

At room temperature, the rate of reaction of atomic O with photo resist is very slow. An elevated temperature of 150 ^oC to 230 ^oC is required to obtain practical etching rates [131]. Similar to other chemical reactions, O_2 plasma resist strip conforms to the first order chemical reaction rate law with an activation energy $E_a = 11.8$ kcal/mole. The activation energy drops to $E_a = 9$ kcal/mole when hydrogen or water is added to the downstream plasma. In Arrhenius form,

$$R = A\exp(-E_a/RT) \tag{6.1}$$

where R is etch rate, A is the pre-exponential factor, E_a is the activation energy, R is universal gas constant and T is the absolute temperature. Figure 6.4 shows that temperature plays an important role in the etch rate, and at least 150 ^oC is

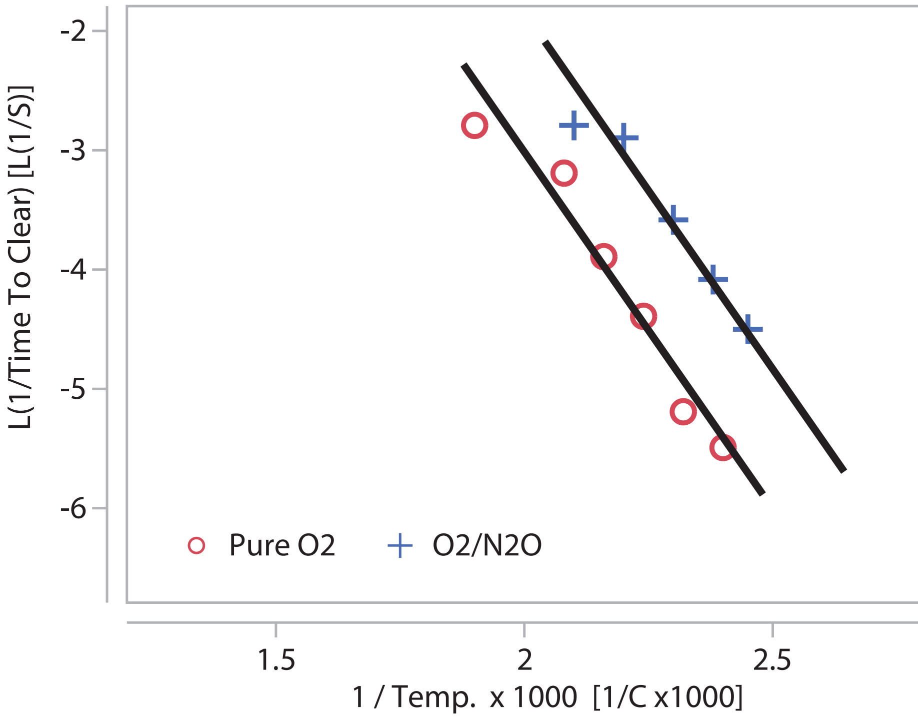


Figure 6.4. Resist etch rates in downstream plasma using O_2 or O_2/N_2O chemistries with different temperatures [9]

required to get a satisfactory etch rate. Mixing with other gases, like N_2O, would also increase the etch rate of photo resist and the activation energy for the reaction would drop to $E_a = 11.0$ kcal/mole compared with the pure O_2 plasma [9].

In summary, the two basic requirements of O_2 plasma photo resist etching are the damage free from radiations of plasma and a practical etch rate.

6.2.3 Chemical Reactions in Photo Resist Descum

Downstream O_2 plasma is used to isotropically strip photo resist from silicon wafers, and pure O atoms are highly selective to resist over silicon or oxide. Before we go into the chemical reactions in the plasma, let's have a look on the photo resist first. Photo resists are primary long-chain organic polymers consisting of mostly carbon and hydrogen [132], Figure 6.5 is a positive photo resist diazonaphthoquinone (DNQ), which is the photoactive compound (PAC), and novolac, a matrix material called *resin*.

In the discharge region, O atoms are produced by electron impact dissociation of the O_2, which is carried by flowing gas to the wafer. For a pure O_2 downstream

Figure 6.5. DNQ-Novolac photoresists

plasma discharge, there are inelastic electron-neutral collisions which supply the ions and radicals, which are continuously lost through chemical reaction and recombination. The formation of an ion, or ionization, can be described as the following collisions and reactions [8],

$$e^- + O_2 \rightarrow O_2^+ + 2e^- \tag{6.2}$$

the radicals can be formed by the below reaction,

$$e^- + O_2 \rightarrow O + O \tag{6.3}$$

and finally the heat and light are given by excitations,

$$e^- + O_2 \rightarrow O_2^* \tag{6.4}$$

$$O_2^* \rightarrow h\nu \tag{6.5}$$

$$e^- + O \rightarrow O^* \tag{6.6}$$

$$O^* \rightarrow h\nu \tag{6.7}$$

where e is electron, h is the Planck's constant, ν is the frequency, and O_2^* and O^* are the excited state of O_2 and O, respectively.

Chemical reactions of O_2 plasma resist strip can be described as the below [7],

$$C_aH_bO_c + O_2(plasma) \rightarrow CO_2 + CO + H_2O(gas) \tag{6.8}$$

$$C_aH_bO_c + O_2 + N_2 \rightarrow CO_2 + CO + H_2O + NOx(gas) \tag{6.9}$$

where $C_aH_bO_c$ is the photo resist.

While the by-products of such O_2 plasma resist descum can be much more complicated than those described in (6.9), because hydrocarbon "combustion" is a very complicated process, a spectroscopic study of O_2 plasma stripping of photo resisted is conducted [10] to analyze the behavior of byproducts. The stripping of photo resists from a silicon wafer using a downstream oxygen plasma has been monitored using the optical emission from electronically excited OH and CO species in the ultraviolet region of the spectrum. Besides the chemically stable species, such as CO_2 and H_2O, the CO and OH species are also produced in the excited electronic states with some radiative lifetimes in the order of 10^{-6} to 10^{-7} second. The spectrum band of (CO^*, OH^*), CO^* and OH^* are 283.0 nm , 297.7 nm , and 308.9 nm, respectively, and the intense is spectrally isolated from other systems arising from plasma-induced oxidation of photo resist. Figure 6.6 showed that the amount of resist etched is nonlinearly correlated with elapsed time by monitoring CO^* band, and similar behavior can be obtained for (CO^*, OH^*) and OH^*. Also CH^* is suggested as another byproduct in the etch rate and plasma power study [11].

The forming gas, H_2N_2, a mixture of hydrogen and nitrogen, is introduced to improve the etch rate of photo resist. Refer to equation (6.1), the activation energy drops from E_a = 11.8 kcal/mole to E_a = 9 kcal/mole when hydrogen or water is added to the downstream plasma. Similarly refer to Figure 6.4, which shows that adding N_2O also improves the etch rate of the photo resist, in other words, adding N_2O into O_2 plasma decreases the activation energy of the reaction. Some research showed that adding 1 % of nitrogen into the O_2 plasma can greatly improve the etch rate of photo resist [13]. Another benefit of the forming gas is that the wafer vapor is formed in the chemical reactions, which not only decreases the activation energy, but also protects semiconductor devices from sodium contamination by resists [12].

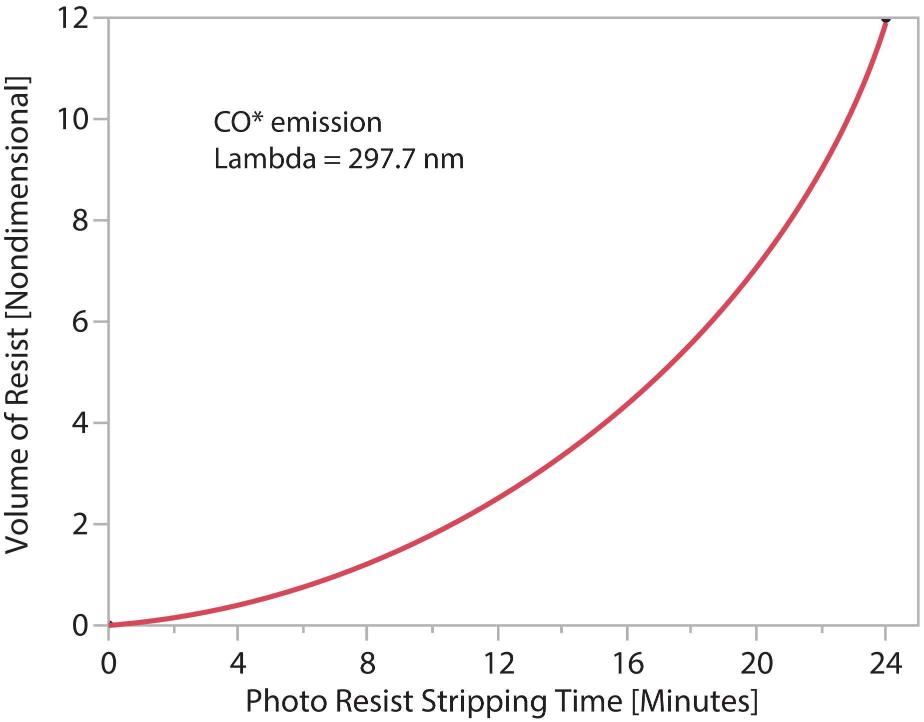


Figure 6.6. The volume of resist is stripped by time by monitoring CO^* band [10]

6.2.4 Physical Etching Models

Three different etching models are proposed for plasma dry etch [118], with different types of etching mechanisms including physical, chemical and ion-assisted etching.

The first model is a linear model, which is the simplest model by assuming that chemicals, radicals and ions act independently and they can be combined in a linear function. Some research shows that etch rate is a linear combination of each component in the flux [133–135].

$$R = \frac{S_c K_f F_c + K_i F_i}{N} \tag{6.10}$$

where R is etch rate, S_c is sticking coefficient which is between 0 and 1, F_c is the chemical flux including reactive neutral and free radicals and F_i is the ion flux at each point of the surface, while K_f and K_i are the relative rate constants for the two processes, which can include any stoichiometric factors.

The second model is the saturation or adsorption model, which is designed to account for the fact that chemical and physical components have interactions and

they do not act independently on etch rate R. There is a saturation effect which means that increasing one of the components in the flux will increase the etch rate to a maximum point. Then the etch rate is saturated due to lack of enough of the other components [136–138]. On the other hand, considering both reactive neutral species and ions that strike the substrate during etching process, the neutral species are adsorbed on the wafer surface and react with material on the surface, which forms the by-products, while the ions strike the wafer surface which can enhance the removal of by-products from the surface as well as improving the adsorption of neutral species. The mechanism is that sputtering the loosely held by-products can induce the reactions to convert the by-products into more volatile species [139,140]

$$R = \frac{1}{N} \frac{1}{\frac{1}{K_i F_i} + \frac{1}{S_c F_c}} \tag{6.11}$$

The equation (6.11) is analogous to one in which two "capacitance" are connected in series rather than in parallel as encountered in electrical engineering.

Finally, a more advanced model can address issues like sidewall inhibitor related ones in ion-assistant etching and physical sputter systems. While the below model tries to address redepositing sputtered material [141,142],

$$R = \frac{K_{sp} Y(\theta) F_i - K_{rd} S_{rd} F_{rd} - S_{cd} F_d}{N} \tag{6.12}$$

where K_{sp} is coefficients, $Y(\theta)$ is sputtering yield which is angle θ dependent, K_{rd} is direct redeposition coefficients, S_{rd} is sticking coefficients of direct redeposition, F_{rd} is chemical flux of direct redeposition, S_{cd} is the sticking coefficients of material which are sputtered off the surface but go into plasma and then drop back down to the wafer surface and F_d is the chemical flux of that.

6.2.5 Model Parameters Candidates

The chamber temperature is one of the good candidates for the VM model parameters, because temperature directly affects the etch rate, which is shown in Figure 6.4.

The chamber pressure needs to be considered as well since the atomic oxygen intensity is a function of chamber pressure, shown in Figure 6.7. The etch rate

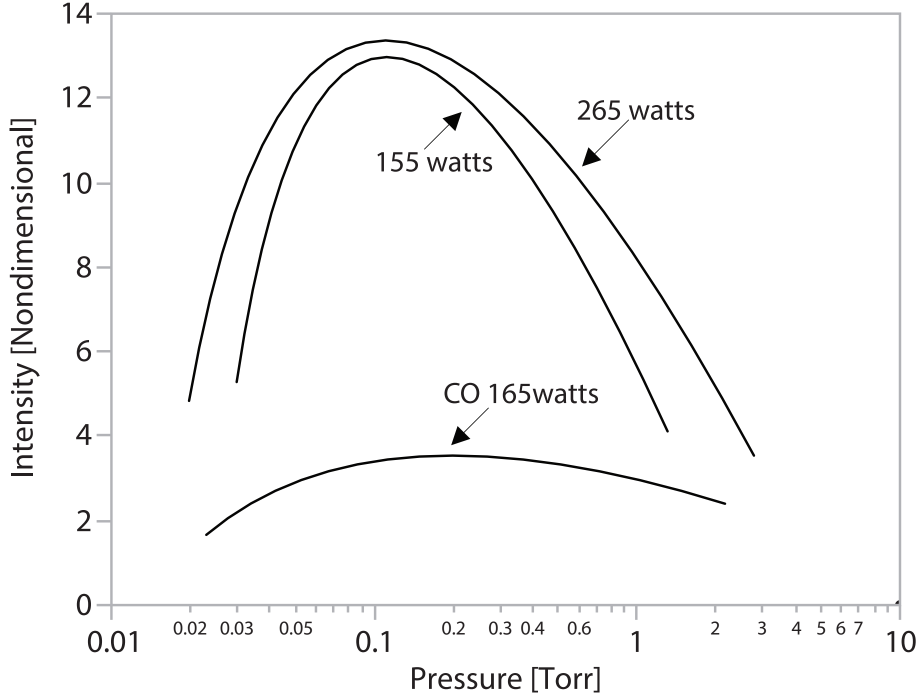


Figure 6.7. Spectral response of atomic oxygen as a function of chamber pressure [11]

increases with pressure to a maximum at approximately 0.1 Torr and then it starts to decrease [11].

Besides the temperature and chamber pressure, the plasma power is found to be strongly correlated to the photo resist etch rate in the downstream O_2 plasma. Figure 6.8 shows that the removal rate of resist is a linear function of applied RF forward power [11]. A higher removal rate occurs at a higher applied RF power.

Finally, O_2 gas flow and forming gas flow H_2N_2 are selected as the model parameters since O radicals are the main "etchant" of the photo resist. On the other hand, the forming gas H_2N_2 is used to improve the etch rate by decreasing the activation energy of the chemical reactions.

6.3 Partial Least Squares Model

6.3.1 Process Data and Model Parameters Selections

The process data items collected in FD are listed in Table 6.1.

All "SetPoint" items and "PinPosition" can be safely eliminated because they are straight lines, and "MicrowaveTime" and "WaferCountTotal" are timer and

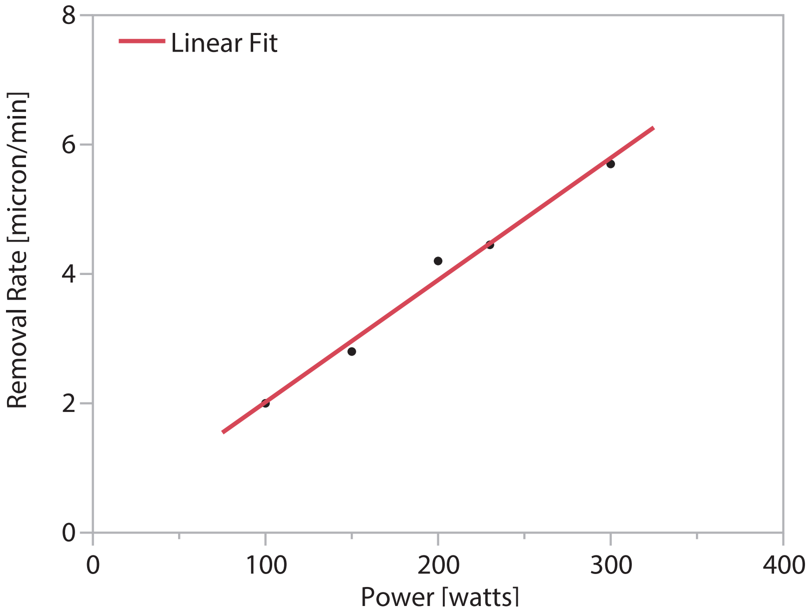


Figure 6.8. The plasma power effect on photo resist removal rate [11]

Table 6.1. Process data items of a descum process

Data Item Name	Units	Data Type
ChamberPressure	mtorr	double
ChamberPressureSetpoint	mtorr	double
ChamberTemp	celsius	double
ChuckTemp	celsius	double
ChuckTempSetpoint	celsius	double
ForwardPower	watts	doubles
ForwardPowerSetpoint	watts	doubles
H2N2GasFlowLow	sccm	doubles
H2N2GasFlowLowSetpoint	sccm	doubles
MicrowaveTime	hours	doubles
O2GasFlowHigh	sccm	doubles
O2GasFlowHighSetpoint	sccm	doubles
PinPosition	NA	double
ThrottleValveAngle	degrees	double
WaferCountTotal	NA	long

counter related to the tool maintenance cycle, which can be safely removed also. Although etch rate is strongly dependent on temperature, which is shown in Figure 6.4, "ChuckTemp" and "ChamberTemp" are not selected as the model parameters, because they are controlled perfectly, as shown in Figure 6.9. The reason for the accurately controlled temperature is that the wafer chuck is a "big" piece of aluminum which improves the temperature uniformity and stability of the system.

In summary, the parameters used in the PLS model evaluations in the next section include the following,

- ChamberPressure
- ForwardPower
- H_2N_2GasFlowLow
- O_2GasFlowHigh

6.3.2 PLS Model and Validation

The detailed PLS algorithm is introduced in Chapter 4. "Leave-One-Out" method is used for model validation, and some model validation methods will be introduced in next section. Figure 6.10 is the result of the validation.

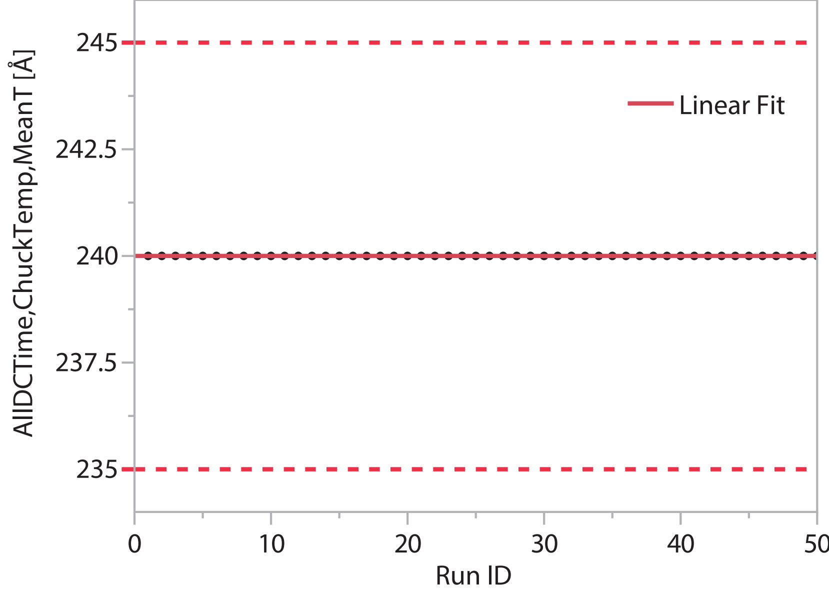


Figure 6.9. Chuck temperature of fifty consecutive runs

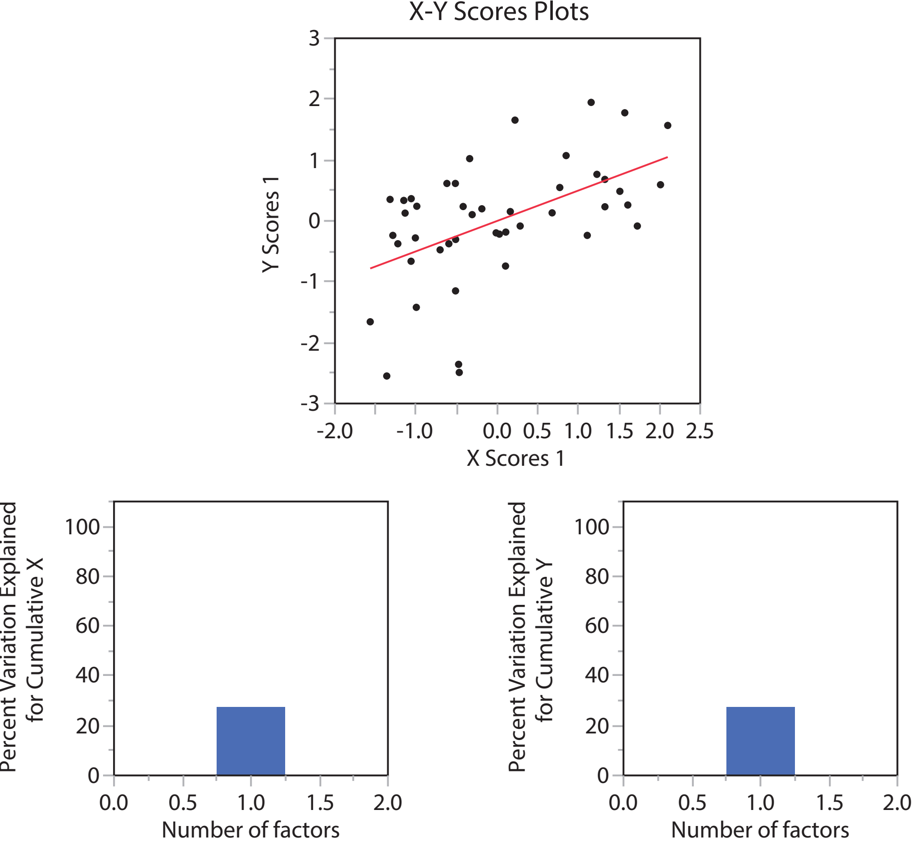


Figure 6.10. PLS model evaluation results

As a result, only 26% of the variation of input data X and 25% of the variation of output data Y can be explained by the PLS model. In the normal operation region, all input data items (or process indicators) are hovering around their setpoints, and there are not enough variations in the data for building an accurate model. That is why a DOE is required to force recipe setpoints to their extremes, so better metrology responses can be obtained. Therefore, the results of the PLS model through the inline FD data and metrology data without any classification can be downgraded. We will propose a "Zonal" data analysis in the next section, which would be able to overcome such problems.

6.3.3 "Zonal" Data Analysis

In the "Zonal" data analysis, we purposely classify the metrology data into different zones. Based on the metrology data's distribution, the data is divided into a "High Zone" and a "Low Zone", as shown in Figure 6.11: Q1 is called the

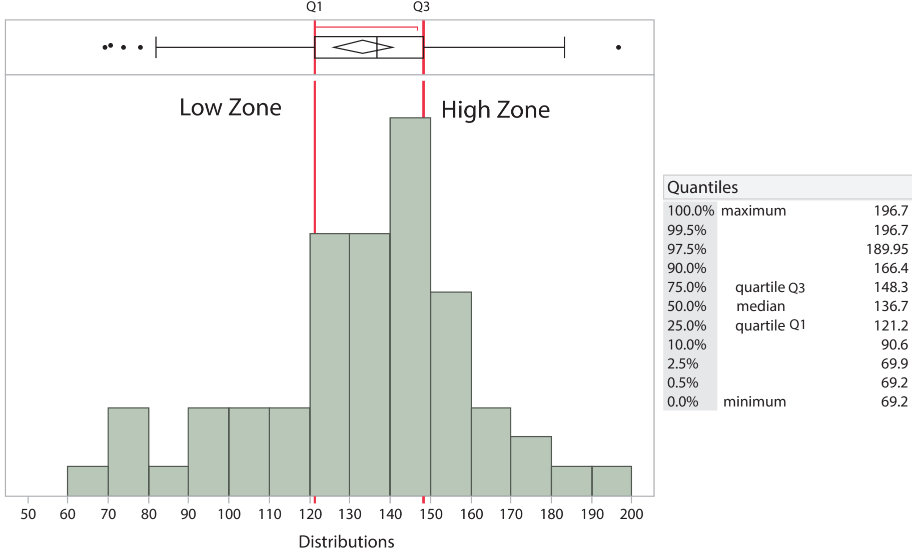


Figure 6.11. Metrology data are classified into high and low zones

first quartile, which is the 25th percentile of the data set; Q3 is the upper quartile, which is the 75th percentile of the data set. Data between Q1 and Q3 is called the interquartile range (IQR). In the example in Figure 6.11, the data in the "High Zone" is greater than Q3, while the data in the 'Low Zone" is less than Q1.

We discovered that the PLS regression model can be improved in some cases if we exclude the data points that fall into the interquartile range. Figure 6.12 is the new validation results after such "treatment" using the same data set of Figure 6.11, and the percent variation can be explained for cumulative Y is improved from 27% to 50%. Such improved result could be from the noise reduction by excluding data in the interquartile range by the assumption that metrology data carries more noise when it is close to the control target.

6.3.4 Model Validation Methods

Three cross validation methods [143, 144] can be used for the PLS evaluations:

- **Holdout**. It partitions the data into two mutually exclusive data sets: one is the training data set and the other one is the validation data set. It often uses two-thirds of data as training data, and the rest is used as validation data.

- **K-fold**. It is also known as rotation estimation, which randomly divides the

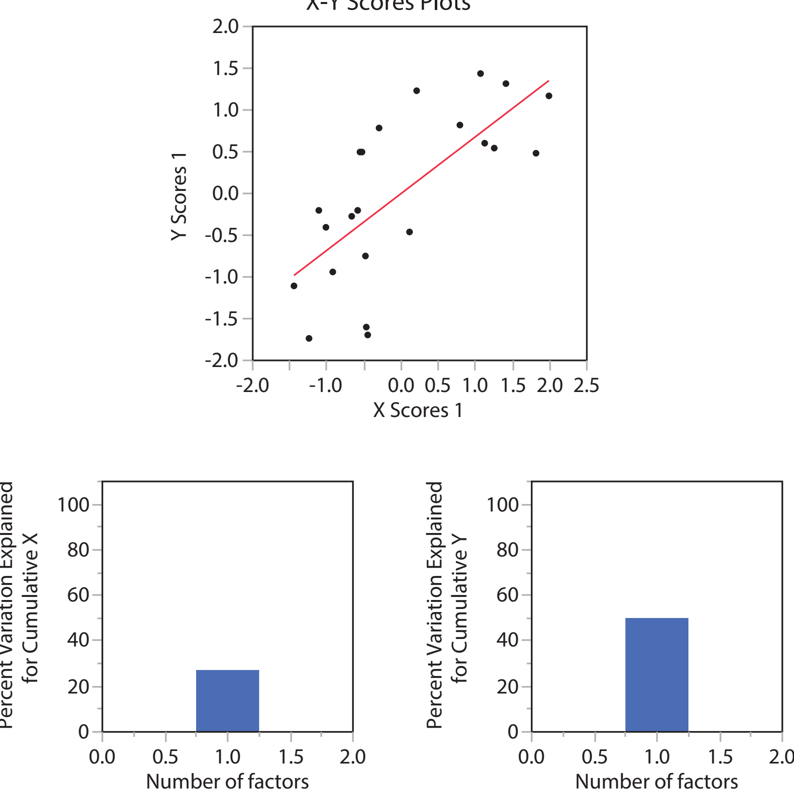


Figure 6.12. Improved PLS model evaluation result is obtained after excluding data in interquartile range

data into k subsets, each of them is used to validate the model which is fit by the rest of the data. So a total k models will be evaluated. The one with the best validation result will be the final model.

- **Leave-one-out**. Similar to k-fold, it estimates k-fold cross validation by using a single data sample into the fold.

6.3.5 Model Update Methods

A recursive PLS was proposed to adapt the process changes for on-line modeling [145]

For a PLS regression model in the original data scale,

$$y_k = mx_k + b \tag{6.13}$$

where y_k is the output or metrology, m is the model coefficients obtained in PLS regression, x_k is the input data and b is the intercept.

Let's denote y as the metrology vector in the moving horizon, the $\bar{y}$ as the mean of y vector and σ_y as the standard deviation of y vector. Similarly, we denote x as the input vector in the moving horizon, the $\bar{x}$ as the mean of the x vector and σ_x as the standard deviation of x.

y_k and x_k can be normalized as,

$$y_{s,k} = \frac{y_k - \bar{y}}{\sigma_y} \tag{6.14}$$

$$x_{s,k} = \frac{x_k - \bar{x}}{\sigma_x} \tag{6.15}$$

where $y_{s,k}$ is normalized of y_k with zero mean and unit variance and similarly $x_{s,k}$ is normalized of x_k, then one can obtain that,

$$y_k = y_{s,k}\sigma_y + \bar{y} \tag{6.16}$$

$$x_k = x_{s,k}\sigma_x + \bar{x} \tag{6.17}$$

Solving equations (6.13) (6.16) and (6.17), a VM model equation in scaled form can be obtained,

$$y_{s,k} = \frac{m\sigma_x}{\sigma_y}x_{s,k} + \frac{b_k - (\bar{y} - m\bar{x})}{\sigma_y} \tag{6.18}$$

The following three cases are proposed to be compared through Matlab evaluations,

- Case 1: a model whose inputs and outputs are in the original scale, $y_k = mx_k + b_k$, with b_k is updated via the EWMA filter,

  $$b_{k+1} = \lambda(y_k - mx_k) + (1-\lambda)b_k \tag{6.19}$$

  where λ is the EWMA weighting factor.

- Case 2: a model whose inputs and outputs are centered and scaled (refer to (6.20)) assumes that the intercept term is equal to zero while $\bar{y}$, $\bar{x}$, σ_y and σ_x are continuously updated by the data in the moving horizon.

  $$y_{s,k} = m_s u_{s,k} \tag{6.20}$$

- Case 3: a model whose inputs and outputs are in the original scale, $y_k = m_k x_k + b_k$, while m_k and b_k are updated through the following,

$$m_k = \frac{m_s \sigma_y}{\sigma_x} \tag{6.21}$$

$$b_k = \bar{y} - m_k \bar{x} \tag{6.22}$$

where m_s is the model coefficients obtained via the PLS regression in the form of centered and scaled data. Note that since there are multiple components in the input data, we need to estimate the process gain of each component using equation (6.21).

Evaluation results are plotted in Figure 6.13 and Figure 6.14. It turns out that Case 2 and Case 3 produce the same prediction results. The result of Case 1 is better than that of Case 2 and Case 3, because the root mean square error cross validation (RMSECV) [146] is smaller($Case1 = 18.2$ vs. $Case2 = 20.2$). The RMSECV is defined as following,

$$RMSECV = \sqrt{\frac{\sum (y_{pred} - y_{act})^2}{n}} \tag{6.23}$$

where y_{pred} is the VM prediction, y_{act} is the actual metrology and n is the validation sample size.

The reason updating intercept produces a better prediction result is related to its better tracking of the drift for incoming variations. We have implemented such intercept state estimation in the multiphysics-based model, which will be discussed in Section 6.4.

6.3.6 PLS Model Discussions

The challenges of the VM model using a moving horizon PLS include unknown incoming variations and the difficulty of obtaining the right signal or right model, because, as we have shown, all input data or process indicators are hovering around their setpoints and there is not enough variation in the data for building an accurate model. In addition, other difficulties include the process conversions, each chamber with its own VM model and the model likely changes over time.

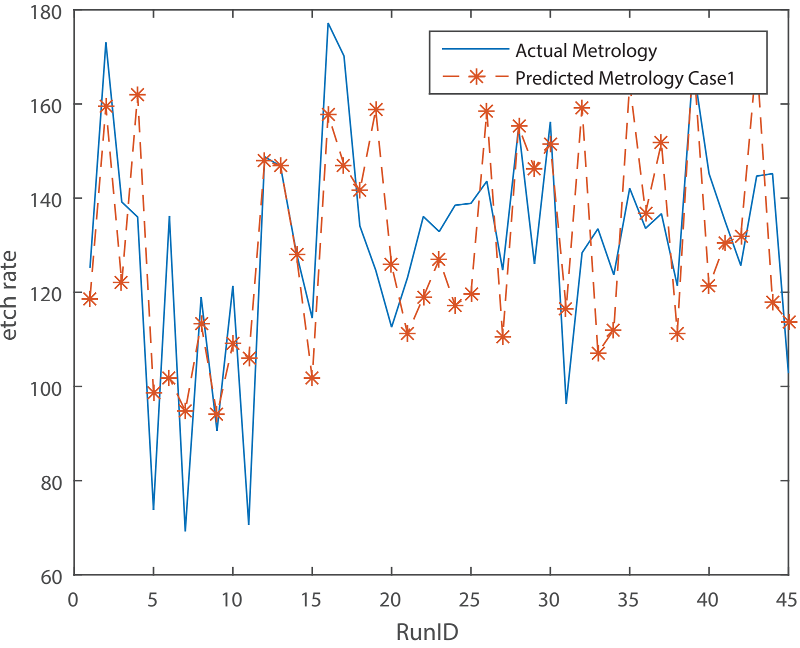


Figure 6.13. Actual metrology and predicted metrology overlay for Case 1.

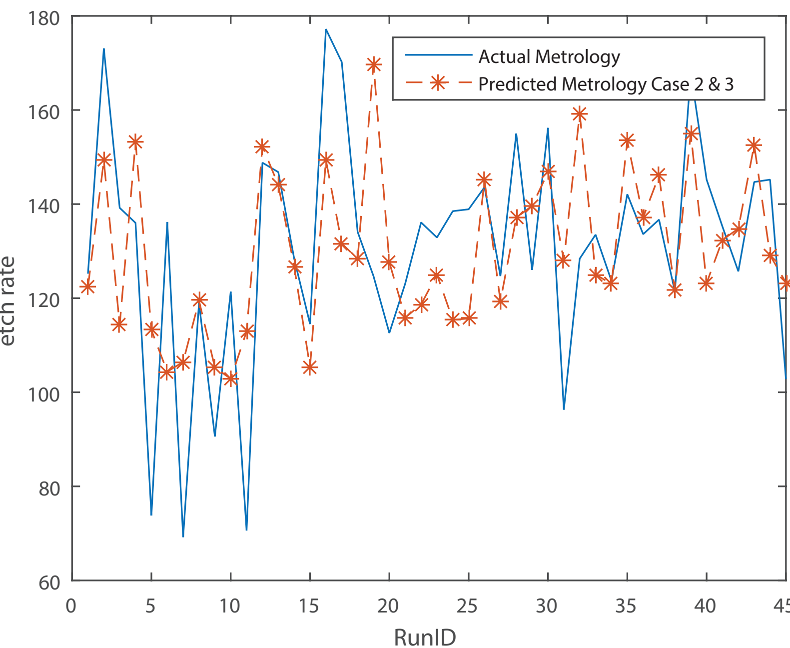


Figure 6.14. Actual metrology and predicted metrology overlay for Case 2 and Case 3.

In order to build a causal model, a DOE is often required because we can intentionally push the tuning knobs to their constraints and at the same time, the most important factors can be screened out.

6.4 New Methods

6.4.1 Introduction of New Model Parameters

In the earlier discussions, we explained why forming gas H_2N_2 is added to improve the stripping rate of photo resist. Fujimura et al. [12] analyzed the etch rate of resist when additive H_2O vapor or H_2 is mixed with the O_2 plasma, and they report that water vapor and hydrogen as the additive gas decrease the activation energy of resist stripping. Figure 6.15 shows that H_2 improved the etch rate rapidly at relatively low concentration and etch rate starts to decline after it reaches the peak, when the hydrogen mixing ratio is about 10%. Since 96.2% of the forming gas is nitrogen and the rest is hydrogen, the corresponding mixing ratio in our case is about 0.43%. The relative concentration of atomic oxygen was calculated from the emission intensity ratios OI(6158)/ArI(7067).

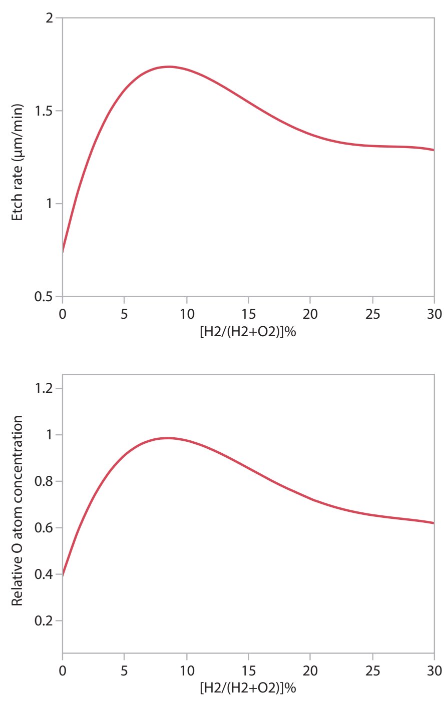


Figure 6.15. Etch rate and relative O atom concentration changes with hydrogen and oxygen mixing ratio [12].

On the other hand, Premachandran [13] reported that the etch rate of photoresist is greatly enhanced by adding 1% of nitrogen into the oxygen plasma. The improved etch rate is mainly from the increase of atomic oxygen concentration and certain impurity gases can improve the dissociation of oxygen molecule in the plasma [147, 148]. Figure 6.16 shows that the etch rate of photo resist is increased by a factor of 2 when 1% of nitrogen is mixed with oxygen [13]. Based on the above background, a new process indicator, which is the gases ratio H_2N_2/O_2, is proposed to be added to the VM model.

On the other hand, a DOE forces recipe process knobs to their extremes, and it often provides the variation needed for constructing an accurate model. Refer to Appendix C. A DOE was conducted to estimate the gains of the model parameters of the multiphysics-based model, while the PLS analysis was used to justify selection of the model parameters. The results of the PLS analysis are shown in Figure 6.17, where the gas ratio H_2N_2/O_2, forward power and chamber pressure are selected as the model parameters, which explains over 80% of the etch rate variations. This

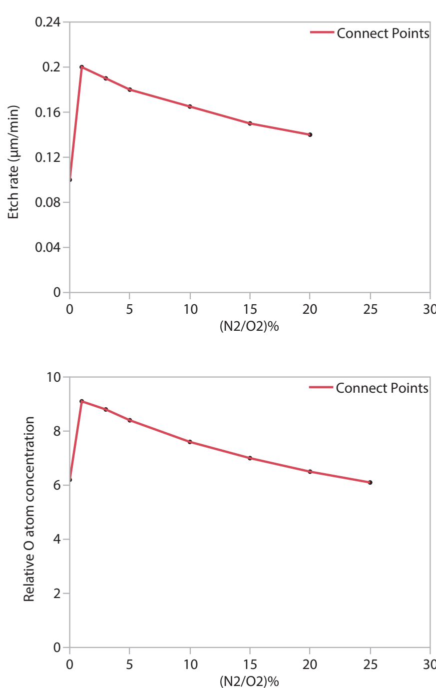
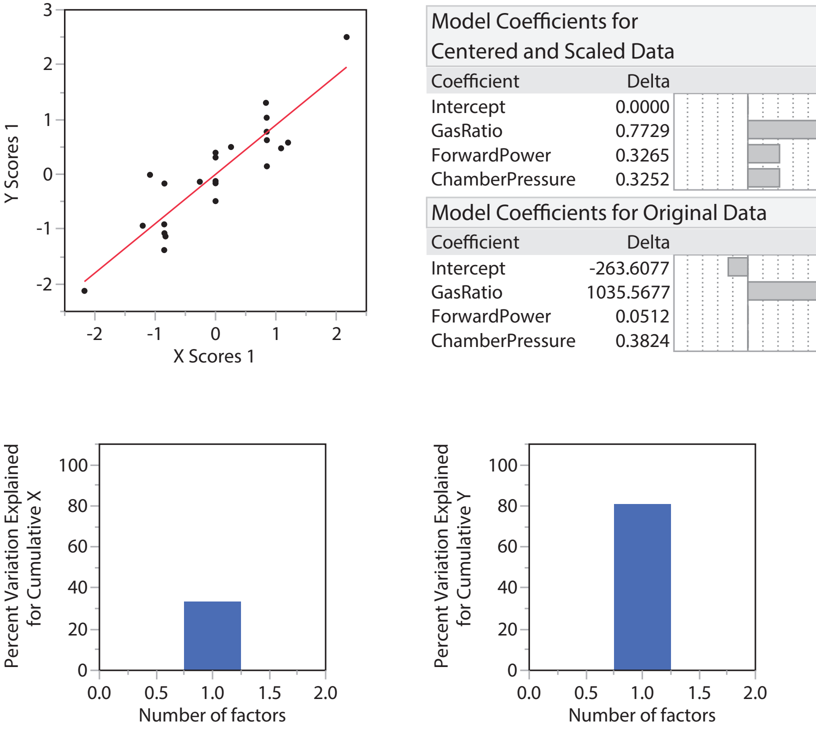


Figure 6.16. Etch rate and relative O atom concentration changes with various percentages of nitrogen in the oxygen plasma [13]



Model Coefficients for Centered and Scaled Data

Coefficient	Delta
Intercept	0.0000
GasRatio	0.7729
ForwardPower	0.3265
ChamberPressure	0.3252

Model Coefficients for Original Data

Coefficient	Delta
Intercept	-263.6077
GasRatio	1035.5677
ForwardPower	0.0512
ChamberPressure	0.3824

Figure 6.17. PLS analysis of descum DOE data.

result is improved over the previous results obtained through the data in normal production conditions. The etch rate response of gas ratio H_2N_2/O_2 is plotted in Figure 6.18 and the etch rate response of the other process knobs can be found in Appendix C.

6.4.2 The VM Model Based on DOE and Its Results

The virtual metrology model derived from the DOE can be described as following:

$$EtchRate = m_1(ForwardPower) + m_2(ChamberPressure) + m_3(\frac{H_2N_2}{O_2}) + Intercept \quad (6.24)$$

in a simpler form,

$$R_k = mx_k + b_k \quad (6.25)$$

where R_k is the etch rate metrology, m is a fixed gain (row vector), x_k is the input data (column vector) and b_k is the intercept, which can be estimated through the EWMA filter.

$$b_{k+1} = \lambda(R_k - mx_k) + (1-\lambda)b_k \quad (6.26)$$

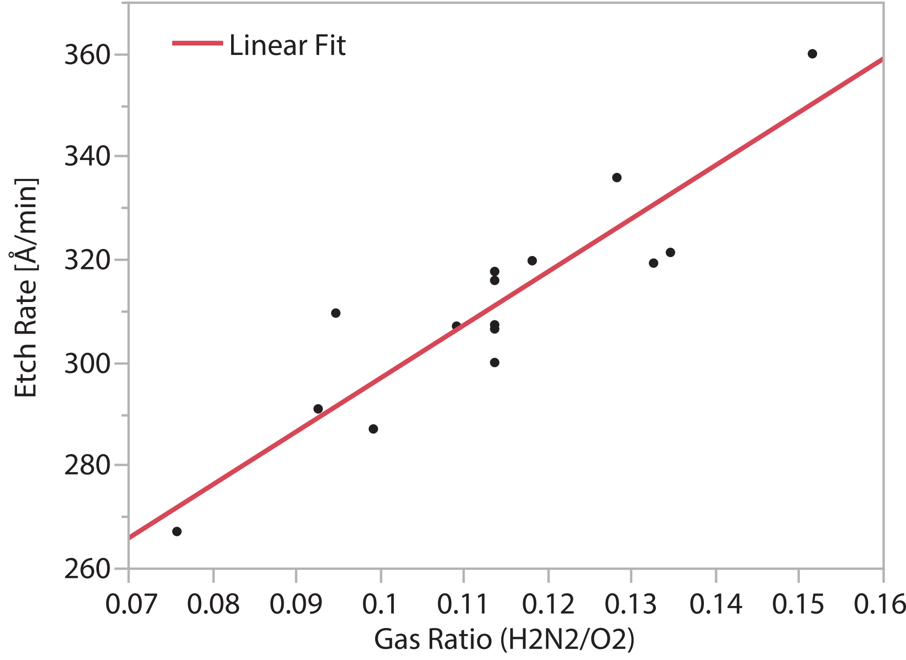


Figure 6.18. Etch rate response with the gas ratio factor

Compared with Figure 6.12, an improved correlation, $R^2 = 0.708$, is obtained on the VM model based on the DOE in Figure 6.19.

6.5 Challenges and Discussions

There are many factors which add variations to the etch rate in a dry etch process [148]. Besides the plasma power, chamber temperature, gas flows, chamber pressure and impurity presented in the plasma chamber, we believe that there are other incoming variations which are difficult to be accounted for. The photo resist variation might be one of the unknown incoming variations, for example, the resist batch. Different resist batches or resists made by different manufacturers may have different film properties, which result in different film stress. The other unknown variation is the elapsed time from the photo step to the etch step, and our theory is that the longer the elapsed time, the more solvent in the resist can be lost, although the data analysis shows no correlation. The VM model often shifts with a maintenance event. A good VM system would be able to account for all maintenance events, no matter at current step or at upstream steps.

A diffusion analysis on the O_2 plasma resist etch shows that the etch rate at the

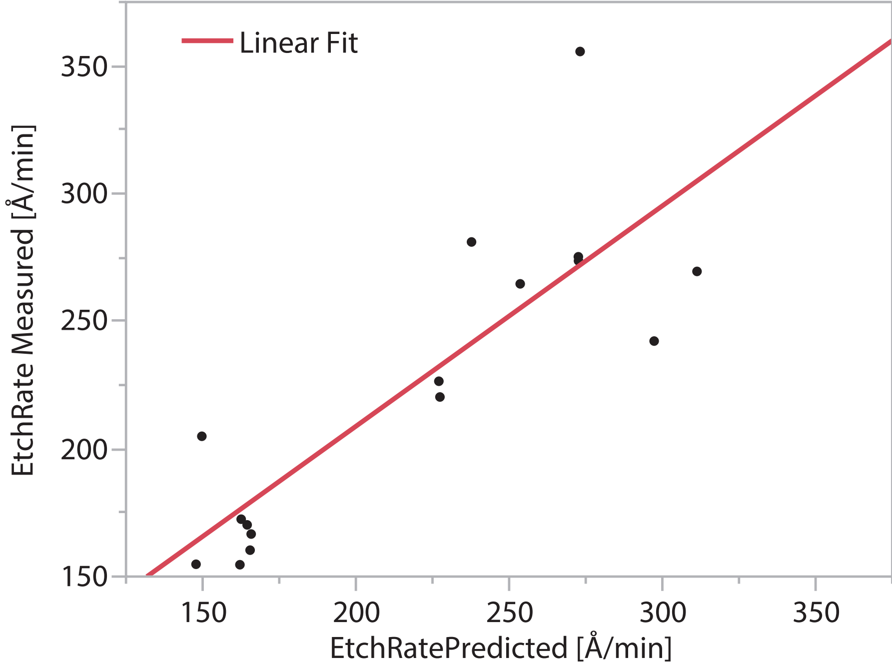


Figure 6.19. The prediction and measurement correlation of the VM model based on DOE.

edge is independent of wafer diameter, but the etch rate at the center is inversely proportional to the square of wafer diameter [149]. This explains the uniformity problem of etch rate within a wafer. The other source of variation within the wafer is called the "loading effect", which is related to distribution and the fraction of surface area of the film being exposed. We did not account for the within-wafer variation in this project.

6.6 Conclusions and Future Work

Although the "Zonal" data analysis improved the prediction quality of the traditional PLS model to some extent, a VM model using chemical reactions and equipment knowledge, which is the multiphysics-based model, outperforms the traditional statistical regression models. Not only are key process indictors selected or removed as model parameters using multiphysics knowledge, but also a new process indictor is created through extensive chemical reaction information.

The main contributions of this work include the following:

- Developed the "Zonal" data analysis, which is a promising method to obtain

accurate model parameters without a costly DOE.

- Created new process indicators using the chemical reactions background knowledge.
- Conducted evaluations and compared three different model update methods.
- Achieved the wafer level etch rate monitoring.

The linear etch rate model is used in this research with the assumptions that the gas and the plasma power act independently. In future work, I want to explore the nonlinear VM model, because the chemical reactions in the O_2 + H_2N_2 downstream plasma are very complicated and some nonlinear etch rate behaviors of plasma power or chamber pressure are expected. Other improvement opportunities include the handling of incoming material variations, e.g., resist batch variations.

CHAPTER 7

CONCLUSIONS AND PROPOSED FUTURE WORK

7.1 Conclusions

Non-threaded R2R control and virtual metrology are important components of process control systems, and both of them are designed to address the costly metrology operations. The threaded R2R controls do not share any information among different threads and the metrology data are diluted, so a larger metrology sampling rate is required to maintain the performance of the process control. However, more metrology operations increase the cost of semiconductor manufacturing due to cycle time and metrology tool cost. In contrast, a non-threaded R2R control does not require a high sampling rate, because the metrology information among different control threads can be shared.

On the other hand, virtual metrology predicts the metrology data without conducting the actual measurements. The predicted metrology data can be used for either process monitoring or process control. The metrology related cost can be reduced by skipping on-line or off-line metrology operations. If the prediction quality is high enough, then this predicted metrology data can be fed into the R2R controllers. Process control and perhaps yield can be improved through variation reductions by VM and R2R controllers.

7.1.1 Hybrid Non-threaded Run-to-Run Control

The major problems associated with a non-threaded R2R controller include unobservable control systems, the continuous change of model dimensions and the high computational cost of state estimation. We addressed the "unobservable control system" problem by a novel hybrid non-threaded R2R controller design, where the controller mode can be downgraded from the non-threaded mode to

the threaded control automatically, when the unobservable problems occur. The changing model dimension issue was solved by reserving dummy contexts in the non-threaded R2R controller without adding any other complexity. After the non-threaded R2R controller had been deployed in the real production environment, we encountered the same high computational costs of the state estimation, in terms of long execution time and the crash of the software execution engine (SEE). We proposed to balance the workload among servers through web services and we also limited the number of dummy states in order to address the high computational costs. Such hybrid non-threaded R2R has been successfully deployed in one of the most critical processes in the high volume production environment, and we have demonstrated its improved process control performance and as well as its robustness.

Threaded, EWMA based non-threaded and the model-based non-threaded R2R controller performances were compared head to head on the same process and the same tool in production, and our data collection showed that the model-based non-threaded controller outperforms the other two control methods, the EWMA non-threaded and the threaded R2R controls. The data collection was done with the threaded control in the "active" mode, which means it actively controlled the process, and the other two non-threaded R2R controllers were in the "passive" mode. Such a framework allows us to compare the performance of controllers in real time, so an automatic on-line tuning of the non-threaded R2R control can be realized.

7.1.2 Etch Rate Prediction of Silicon Dioxide Film in Diluted HF Solution

Virtual metrology is commonly built upon traditional statistical regression models, such as PLS and neural networks. Since the chemistry of etching silicon dioxide in the diluted HF solution is well described in the literature, in this research project, we incorporated physics and chemical reaction models into the virtual metrology. The comparison between traditional PLS regression model and the multiphysics-based model suggested that the multiphysics-based model is a better method to select key process variables for the VM model and it is a better method

to establish meaningful process indicators. We also discovered more advantages of a multiphysics-based model. For instance, it requires less training data because of its fast convergence and it can account for variations of chemical and gas batches and VM prediction accuracy can be biased or compromised without taking account of those.

The VM data can be fed into R2R controllers. For example, VM thickness predictions can be used as a feed-forward component of a wafer level dry etch R2R controller downstream to improve wafer level variations. In our research, we used the VM system in a different way where VM is used to update the R2R model. An R2R control model, typically the slope, is usually fixed, so the R2R control performance may be downgraded after process conversions at the current step or the upstream steps. A frequent model update through the VM system ensures the optimal state, especially when there is a feed-forward component.

We have demonstrated in high volume production that the multiphysics-based VM model produces better prediction quality than the traditional statistical models. The predicted etch rate is also used to update process gains of the R2R controller, so that the R2R controller compensates feed-forward disturbance better. The process control was still capable after a 50% sampling rate reduction. This project has realized almost all the VM benefits, which include excursion prevention, yield improvement, process capability gain, cycle time and the cost reduction.

7.1.3 A Generic Diffusion Furnace Virtual Metrology

No matter how well a furnace R2R controller performs, the thickness profile exists because of the design of the heaters and the gas depletion effect. Such thickness profile introduces variations into semiconductor manufacturing. We developed a multiphysics model through the equipment knowledge and the design of experiment. Five Gaussian curves and one intercept term are used to produce the final thickness profiles. If the shape of the thickness profile changes over time, then the peak magnitudes of the five Gaussian curves can be updated through metrology data. In the case that the shape of the thickness profile remains the same, only the intercept term is to be updated when new metrology becomes available. To get rid of nonlinear complexity, we propose to assume that the standard deviation terms

remain the same, and such an approximation is acceptable, because the standard deviations are "mainly" determined by design of heaters such as the length of the heater, which is fixed. Such assumptions have made the state estimation much simpler.

Two difficulties, queue time effects and the R2R control adjustment, were discovered when we deployed this strategy in the actual production environment. An offset curve or model was established to solve queue time problem. On the other hand, R2R adjustments can be modeled or offset by adjusting the peak value of the Gaussian curves or the intercept term. We obtained excellent prediction results after solving these problems.

One of the major benefits of this project is feeding forward diffusion thickness profile data to a wafer level R2R control at the downstream process steps, so wafer level variation caused by different furnace positions can be reduced or removed. The wafer level dry etch R2R controller application was proposed in literature, while we propose to extend this methodology to an ion implantation step, which will be discussed in detail in Section 7.2.

7.1.4 Oxygen Plasma Resist Descum Virtual Metrology

As of today, the plasma physics and phenomenon are not well understood and this makes virtual metrology of dry etch very challenging. We studied the major characteristics of O_2 plasma resist descum process including the sodium contamination free, electrical charge and radiation free through the downstream plasma and the etch rate improvement by introducing a forming gas H_2O_2. Model parameters were selected through the background of physics and chemical reactions. First we invented a new method called "Zonal" data analysis, which can improve the prediction quality of a PLS model by almost 50%. Three recursive PLS model update methods were simulated on the same data set, and we concluded that intercept update performs best, because it can adapt to the incoming variations.

We created a new process indictor through the extensive chemical reaction and process knowledge to improve the multiphysics-based etch rate model. The model parameters were obtained through a DOE. Based on the experience of the PLS evaluations, the model update is through the intercept state estimation,

which captures the drift of incoming variations. A $r^2 = 0.708$ was achieved in the production environment, which is sufficient for real time wafer level etch rate monitoring. The multiphysics-based we used is a simple linear model and we believe that the prediction can be improved further through exploring nonlinear models or handling incoming material variations like resist batch.

7.2 Future Work

7.2.1 Context Matching and Relaxation

Every wafer within a lot has a context history, and two wafers having the same context history tend to have the similar metrology data. Such assumptions in most situations are valid because two wafers in a lot were processed at a similar time frame and by the same gas or chemical batch. As long as the chamber does not drift too quickly, then we can safely assume that the two wafers going through the same contexts (or chambers) have the same metrology. Based on this assumption, we would be able to develop a new VM model called "context matching and relaxation."

Referring to Table 7.1, there are four steps affecting the post metrology data, where "step 1" has the most significant contribution, "step 2" has the second most significant contribution, and so forth. If two wafers in the same lot went through the same context (or process chamber) for all four steps, then the absolute prediction error is relatively small, while the prediction error bar grows when the context matching level is relaxed [150] as shown in Figure 7.1.

Let η denote the confidence of prediction which is associated with the context matching level analysis, and the prediction of unmeasured wafer can be described

Table 7.1. The context match and relaxation

Matching Level	Step 1 Match	Step 2 Match	Step 3 Match	Step 4 Match
Full contexts match	Yes	Yes	Yes	Yes
Relaxed match 1	Yes	Yes	Yes	No
Relaxed match 2	Yes	Yes	No	No
Relaxed match 3	Yes	No	No	No

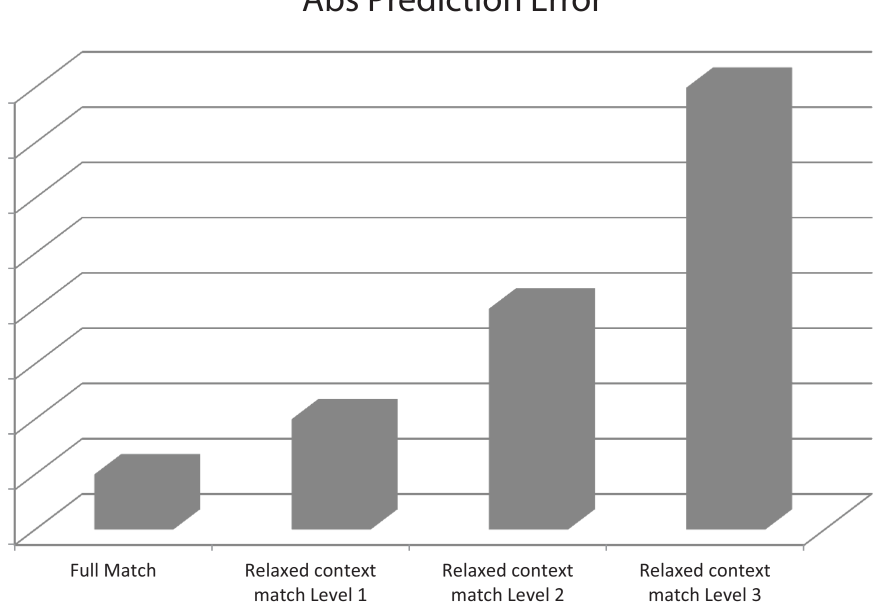


Figure 7.1. Prediction error bar grows when context matching level is relaxed.

as,

$$\hat{y} = \eta y_{wafer} + (1 - \eta) y_{lot} \tag{7.1}$$

where $\hat{y}$ is the prediction of an unmeasured wafer, η is the confidence factor related to the context relaxation level, y_{wafer} is the actual metrology of a wafer which has the best context match with the wafer being predicted and y_{lot} is the lot average of all measured wafers.

This VM model is able to handle multiple sources of variations from upstream steps. The future work includes the research of estimating η through the analysis of variance (ANOVA) or a bootstrap, and testing of this algorithm in the real production environment.

7.2.2 Auto Tuning of Non-threaded R2R Control

Referring to equation (3.47), the tuning of R and Q of ratio for the state estimation is an art, while the nature of poor noise disturbance rejection makes the non-threaded state estimation tuning more difficult compared with the threaded R2R control. In Figure 3.14, we proposed a framework of the automatic tuning non-

threaded controller as an extension of the non-threaded R2R research. The process of R and Q ratio tuning is described in Figure 7.2.

The data collection phase ensures the non-threaded R2R controller collects enough data, and the performance function can be evaluated as a mean squared error (MSE),

$$J = \frac{1}{n}\Sigma_{i-1}^{n}(\hat{y}_i - y_i)^2 \tag{7.2}$$

where n is the number of valid record in the moving horizon, $\hat{y}_i$ is the output prediction and y_i is the actual metrology. A new R and Q ratio can be validated through a passive validation phase before it can be applied to the production.

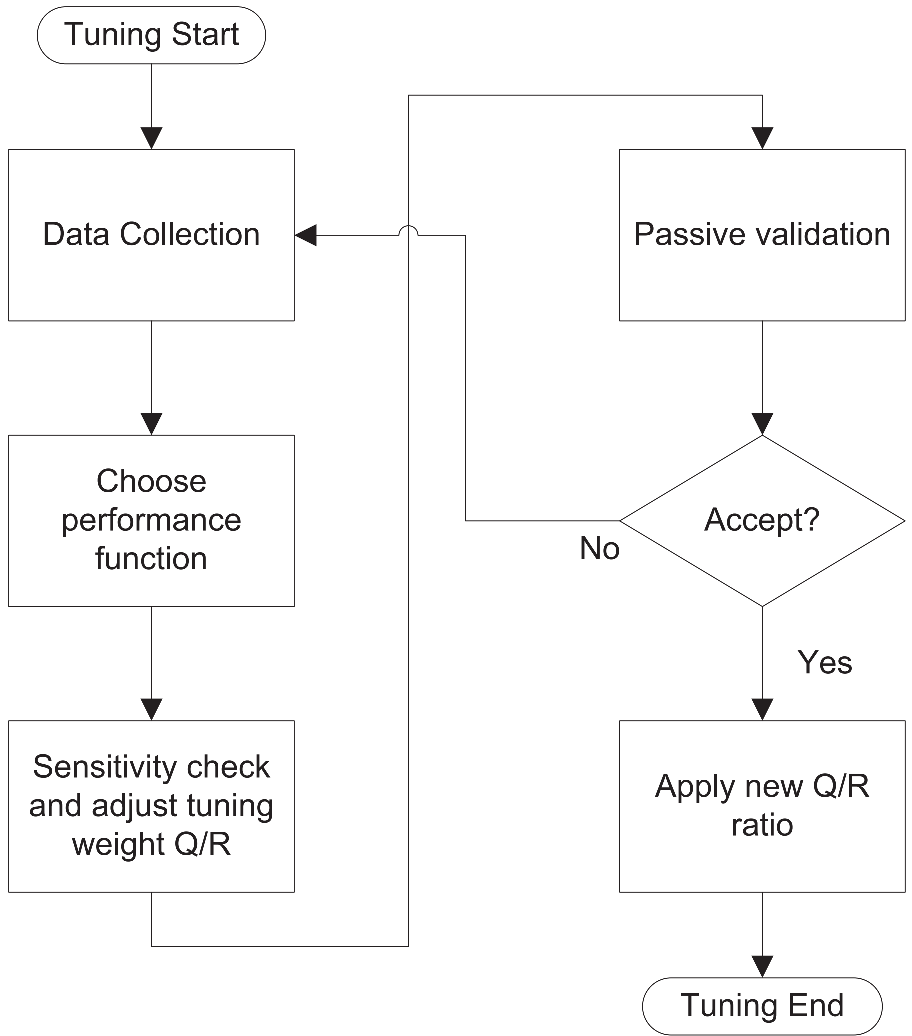


Figure 7.2. The process of auto-tuning state estimation Q and R for a non-threaded R2R control.

If Q is a fixed value, then the "sensitivity" is given by,

$$Sensitivity = \frac{\partial J}{\partial R} \tag{7.3}$$

and the sign of "sensitivity" can be used to determine the tuning direction of R. However, how to come up with "sensitivity" (numerically or analytically) and the magnitude of adjustment are still to be researched in future.

7.2.3 VM and Wafer Level Implant R2R

In Chapter 5, we demonstrated the prediction capability of a poly-silicon profile in Figure 5.16. One of our goals is to realize the benefit of such a diffusion VM system. Dry etch R2R controller application was proposed in literature [127], while we want to extend this methodology to an ion implantation step.

Referring to Figure 7.3, poly-silicon resistors are determined by the factors like sheet resistance, poly-silicon resistor dimension and the contact resistances. Kyuho et al. applied APC in the ion implantation for an analog device [14] and a strong correlation was discovered between poly-silicon resistance and the poly-silicon thickness, as shown in Figure 7.4.

As an extension of the research, combining the thickness profile prediction in Chapter 5 and the R2R algorithm [14] of an ion implant, a wafer level R2R

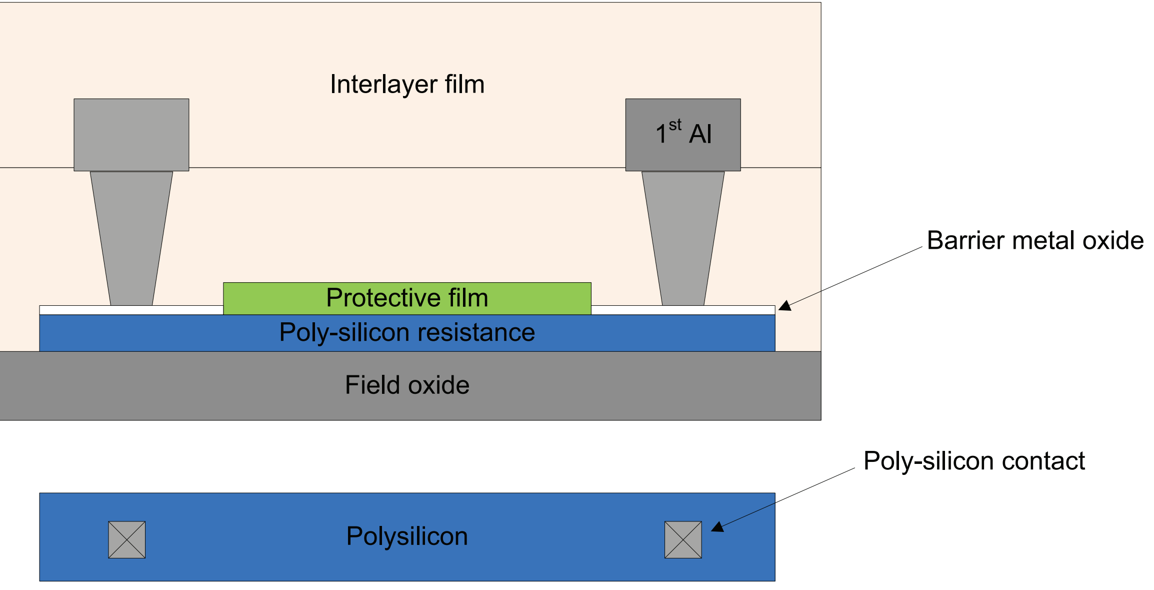


Figure 7.3. Poly-silicon resistor [14]

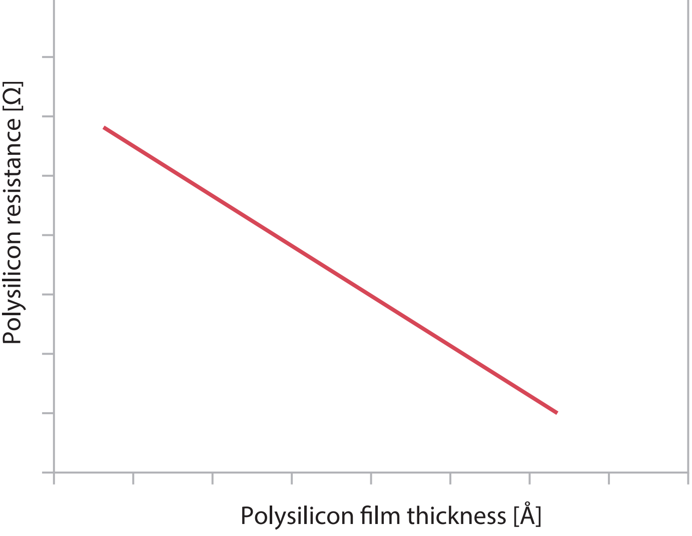


Figure 7.4. The relationship between the poly-silicon resistance and the gain size of the poly-silicon [14]

controller can be implemented at the downstream ion implant step to compensate for the furnace thickness profile.

The resistance of the poly-silicon can be predicted as,

$$R_1 = a y_{poly} + b \tag{7.4}$$

where R_1 is the prediction of resistance through poly-silicon thickness y_{poly}, and a and b are coefficients.

The distance between the predicted resistance and the target value, R_2, can be computed as,

$$R_2 = R_t - \lambda(R_1 - R_t) \tag{7.5}$$

where R_t is the poly-silicon resistance target and λ is a damping factor.

Finally, the wafer level implant dose, Q_d, can be calculated as following,

$$Q_d = cR_2^2 + dR_2 + e \tag{7.6}$$

where c, d and e are coefficients.

This is a feed-forward wafer level R2R control implementation at the ion implant step. The implant dose is manipulated by the incoming wafer thickness predictions. It would be a very interesting project to realize the VM benefits of a wafer level thickness prediction for the diffusion furnace in the future.

APPENDIX A

CURVE FITTING IN MATLAB

The curve fitting toolbox is used for the diffusion furnace VM project, and this is a quick introduction to the usage of this toolbox. The curve fitting application can be opened by entering "cftool" in the Command Window of Matlab.

A.1 Data Selection

The data selection tab can be accessed by clicking the "data" button, and one can use the drop-down lists in the curve fitting application to select X data and Y data in the workspace. Our furnace data set includes an input vector p and an output vector y, which are the boat slots and their corresponding thickness metrology data. Refer to Figure A.1: select p as "X data" and select y as "Y data."

A.2 Data Fitting

The data fitting tab can be accessed by clicking the "Fitting" button, and "Custom Equations" is selected for the "Type of fit" as shown in Figure A.2. We can choose p as the "Independent variable", and then we can enter equation (5.35) as the "General Equations". The lower and upper constraints are used for better convergence.

A.3 Curve Fitting Results

The detail curve fitting result is the screen shot shown in Figure A.3, and the output includes the five standard deviation terms of all five Gaussian curves, five peak values and one intercept term. The goodness of fit is also part of the output.

The results of curve fitting are plotted in Figure A.4. Please note that five standard deviation terms are assumed to be fixed in the VM model while the five

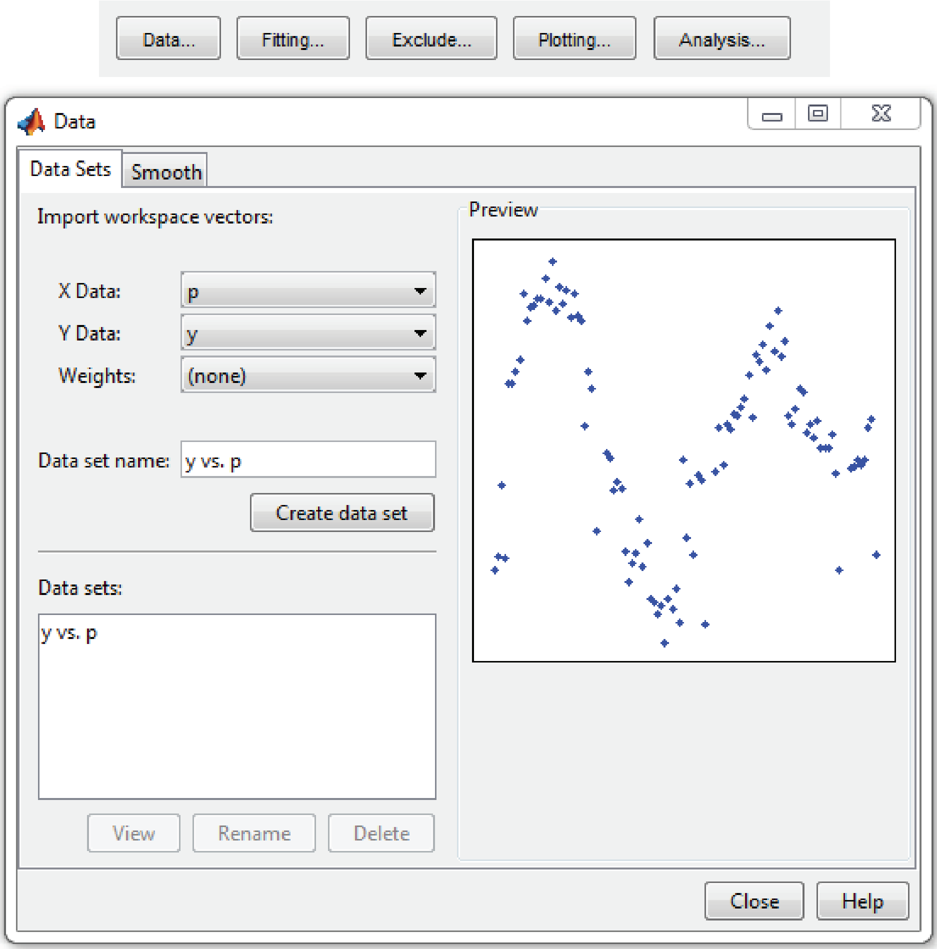


Figure A.1. Data selection for curve fitting application

peak values and one intercept term can be used as the initial states, which can be updated by the actual metrology data.

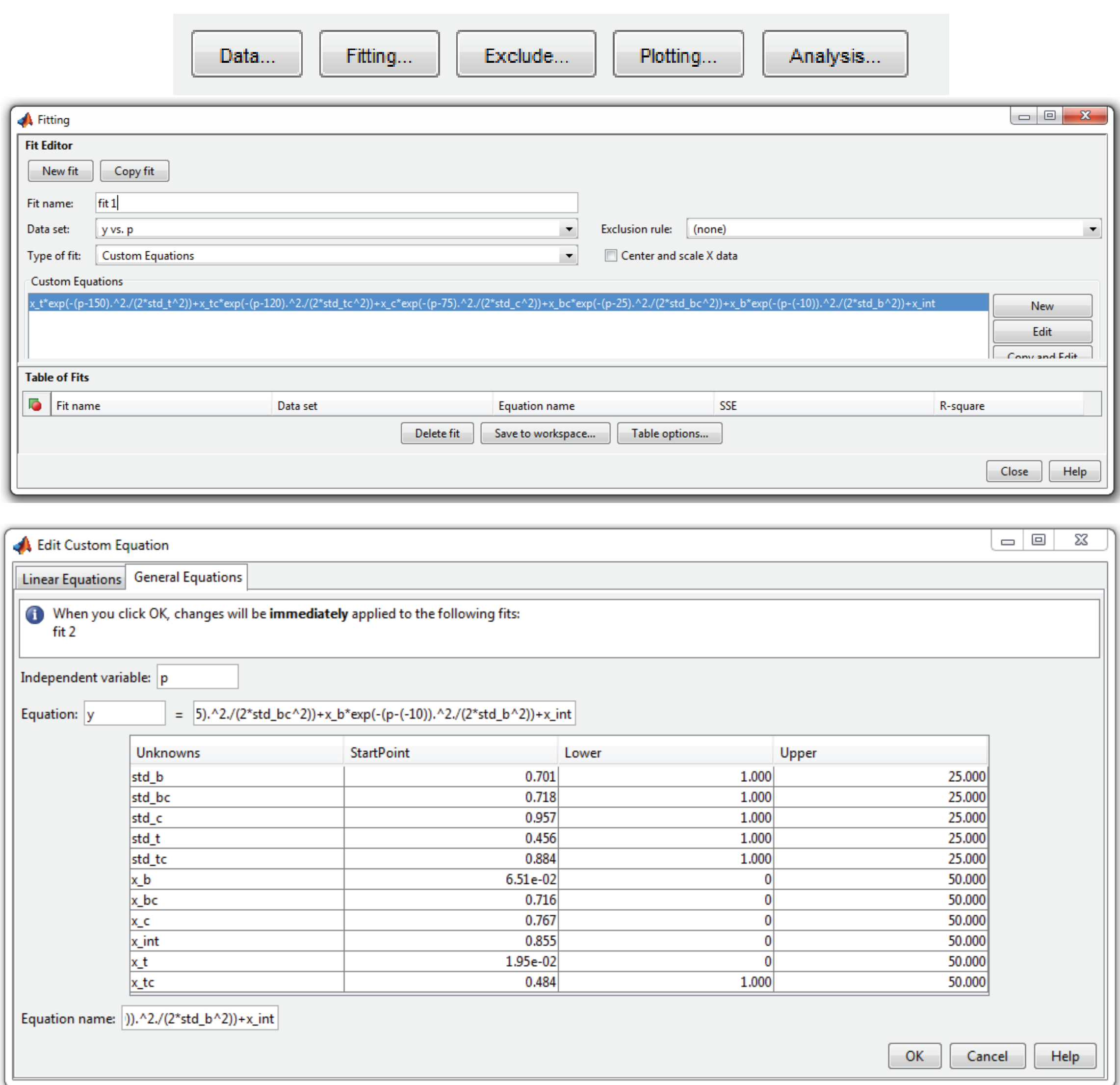


Figure A.2. Data fitting in the curve fitting application

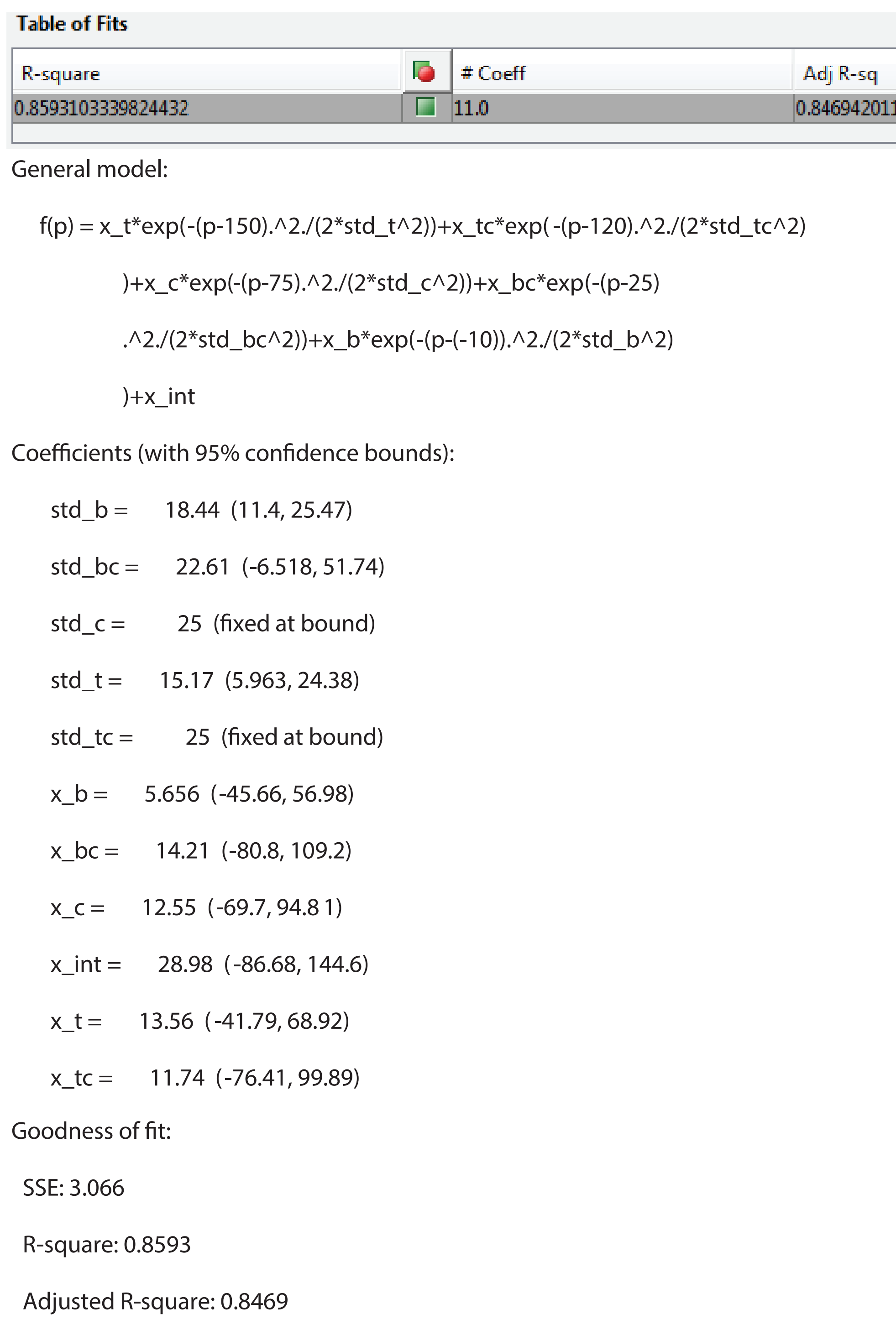


Table of Fits

R-square		# Coeff	Adj R-sq
0.8593103339824432		11.0	0.8469420116951853

General model:

f(p) = x_t*exp(-(p-150).^2./(2*std_t^2))+x_tc*exp(-(p-120).^2./(2*std_tc^2)

)+x_c*exp(-(p-75).^2./(2*std_c^2))+x_bc*exp(-(p-25)

.^2./(2*std_bc^2))+x_b*exp(-(p-(-10)).^2./(2*std_b^2)

)+x_int

Coefficients (with 95% confidence bounds):

std_b = 18.44 (11.4, 25.47)

std_bc = 22.61 (-6.518, 51.74)

std_c = 25 (fixed at bound)

std_t = 15.17 (5.963, 24.38)

std_tc = 25 (fixed at bound)

x_b = 5.656 (-45.66, 56.98)

x_bc = 14.21 (-80.8, 109.2)

x_c = 12.55 (-69.7, 94.8 1)

x_int = 28.98 (-86.68, 144.6)

x_t = 13.56 (-41.79, 68.92)

x_tc = 11.74 (-76.41, 99.89)

Goodness of fit:

SSE: 3.066

R-square: 0.8593

Adjusted R-square: 0.8469

RMSE: 0.1836

Figure A.3. Curve fitting results

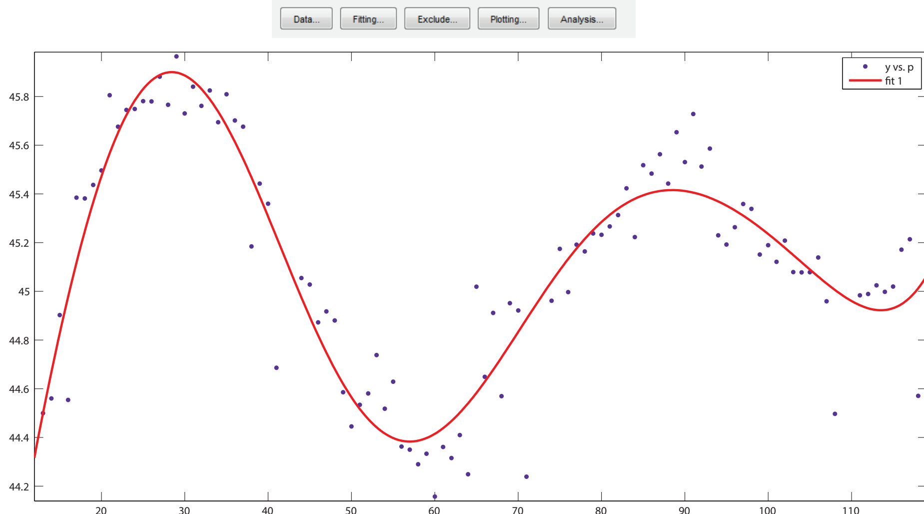


Figure A.4. The plot of the curve fitting output

APPENDIX B

DATA COLLECTION OF ETCH RATE AND HF WEIGHT PERCENT

Table B.1. 1000:1/T2 HF etch rate and weight percent

Date and Time	LotID	Sample Wt%	ER (Å/min)	Predicted Wt%
6/21/2013 10:28:00 AM	V556252.002	0.053	9.775	NA
6/27/2013 9:24:00 AM	V670782.002	0.053	10.432	NA
7/6/2013 5:22:00 PM	V839482.002	0.052	10.257	NA
7/10/2013 10:20:00 AM	W903622.002	0.054	10.369	NA
7/13/2013 4:17:00 PM	W960442.002	0.052	10.221	0.046
7/23/2013 8:15:00 AM	X122062.002	0.053	10.244	0.046
7/24/2013 5:22:00 PM	X155312.002	0.054	10.325	0.047
8/1/2013 12:29:00 PM	W289142.002	0.055	11.147	0.046
8/2/2013 8:52:00 AM	W307782.002	0.053	10.803	0.046
8/3/2013 9:23:00 AM	W326502.002	0.053	10.838	0.046
8/3/2013 2:49:00 PM	W334102.002	0.054	10.851	0.046
8/7/2013 11:25:00 AM	W390932.002	0.052	10.401	NA

Table B.2. 500:1/T2 HF etch rate and weight percent

Date and Time	LotID	Sample Wt%	ER (Å/min)	Predicted Wt%
6/21/2013 9:17:00 AM	V567022.002	0.104	2.270	NA
6/27/2013 9:01:00 AM	V666592.002	0.1	2.190	NA
7/6/2013 5:02:00 PM	V825732.002	0.101	2.163	NA
7/10/2013 10:08:00 AM	W898522.002	0.102	2.256	NA
7/13/2013 3:55:00 PM	W960452.002	0.101	2.299	0.094
7/23/2013 7:44:00 AM	X131292.002	0.101	2.246	0.093
7/24/2013 5:02:00 PM	X100182.002	0.104	2.316	0.094
8/1/2013 12:09:00 PM	W296322.002	0.103	2.424	0.093
8/2/2013 9:03:00 AM	W312522.002	0.102	2.430	0.093
8/3/2013 9:46:00 AM	X312572.002	0.101	2.201	0.093
8/3/2013 5:10:00 PM	W334162.002	0.1	2.227	0.092
8/7/2013 10:47:00 AM	W402762.002	0.104	2.273	NA

Table B.3. 500:1/T4 HF etch rate and weight percent

Date and Time	LotID	Sample Wt%	ER (Å/min)	Predicted Wt%
6/22/2013 2:44:00 PM	V589252.002	0.11	2.360	NA
6/29/2013 12:03:00 PM	V710242.002	0.11	2.270	NA
7/6/2013 3:17:00 PM	V832402.002	0.107	2.293	NA
7/10/2013 10:32:00 AM	W896002.002	0.113	2.316	NA
7/13/2013 3:05:00 PM	W960432.002	0.107	2.354	0.108
7/23/2013 6:59:00 AM	X126302.002	0.107	2.295	0.107
7/24/2013 4:30:00 PM	X154062.002	0.108	2.300	0.107
8/1/2013 11:04:00 AM	X296312.002	0.107	2.361	0.108
8/2/2013 10:20:00 AM	W312782.002	0.108	2.327	0.108
8/3/2013 7:05:00 AM	W312522.002	0.108	2.285	0.108
8/3/2013 2:05:00 PM	W334152.002	0.107	2.305	0.107
8/7/2013 10:35:00 AM	W402742.002	0.108	2.342	NA

Table B.4. 100:1/T4 HF etch rate and weight percent

Date and Time	LotID	Sample Wt%	ER (Å/min)	Predicted Wt%
6/22/2013 2:25:00 PM	V589222.002	0.486	24.950	NA
6/29/2013 11:23:00 AM	V711272.002	0.478	24.500	NA
7/6/2013 2:50:00 PM	V837162.002	0.482	24.842	NA
7/10/2013 10:52:00 AM	W894602.002	0.477	24.243	NA
7/13/2013 3:25:00 PM	W955982.002	0.477	24.353	0.476
7/23/2013 6:15:00 AM	X125102.002	0.483	24.452	0.477
7/24/2013 4:11:00 PM	W150602.002	0.482	24.431	0.478
8/1/2013 1:00:00 PM	W282402.002	0.487	24.432	0.479
8/2/2013 10:01:00 AM	W264442.002	0.482	24.736	0.478
8/3/2013 7:38:00 AM	X296312.002	0.477	24.213	0.475
8/3/2013 1:45:00 PM	W334122.002	0.474	24.656	NA
8/7/2013 10:06:00 AM	W394242.002	0.48	24.207	NA

APPENDIX C

DESIGN OF EXPERIMENT FOR DESCUM

C.1 DOE Table and Interpretation

Although the process knobs or recipe setpoints can be pushed to their extremes, some constraints have to be considered in this process. For example, the total gas flow has to be above 5000 sccm and the minimum chamber pressure is 710 mTorr. The DOE design and results are listed in Table C.1.

Figure 6.18 and Figures C.1 to C.4 are the etch rate response of selected process knobs, and equations (C.1) to (C.5) are the linear fits of etch rate responses with the selected process knobs.

$$ER = 1035.58(\frac{H_2N_2}{O_2}) + 193.45 \tag{C.1}$$

$$ER = -0.0217(O_2) + 425.95 \tag{C.2}$$

$$ER = 0.1954(H_2N_2) + 194.34 \tag{C.3}$$

$$ER = 0.0512(ForwardPower) + 137.35 \tag{C.4}$$

$$ER = 0.3824(ChamberPressure) + 24.14 \tag{C.5}$$

Table C.1. DOE of O_2 plasma descum

No.	O2GasFlow (Sccm)	H2N2GasFlow (Sccm)	ForwardPower (Watt)	ChamberPressure (mTorr)	Etch Rate (Å/min)
1	5280	600	3300	750	300.07
2	4458	600	3300	750	321.47
3	6480	600	3300	750	291.06
4	5280	600	3300	750	307.40
5	5280	500	3300	750	309.68
6	5280	700	3300	750	319.39
7	5280	600	3000	750	282.16
8	5280	600	3600	750	312.80
9	4680	600	3300	750	336.02
10	5280	600	3300	710	288.32
11	5280	600	3300	790	325.44
12	5080	600	3300	750	319.86
13	5500	600	3300	750	307.14
14	5280	600	3300	750	317.77
15	5280	400	3300	750	267.23
16	5280	800	3300	750	360.08
17	5280	600	3300	750	316.03
18	5280	600	3000	750	291.57
19	5280	600	3600	750	322.37
20	6050	600	3300	750	287.16
21	5280	600	3300	710	306.5
22	5280	600	3300	790	330.57
23	5280	600	3300	750	306.6

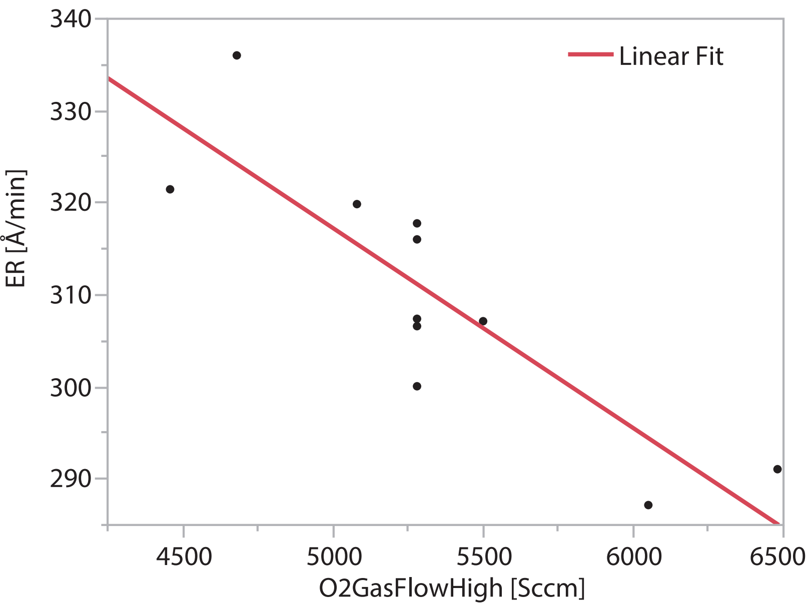


Figure C.1. Etch rate response with the O_2 gas factor

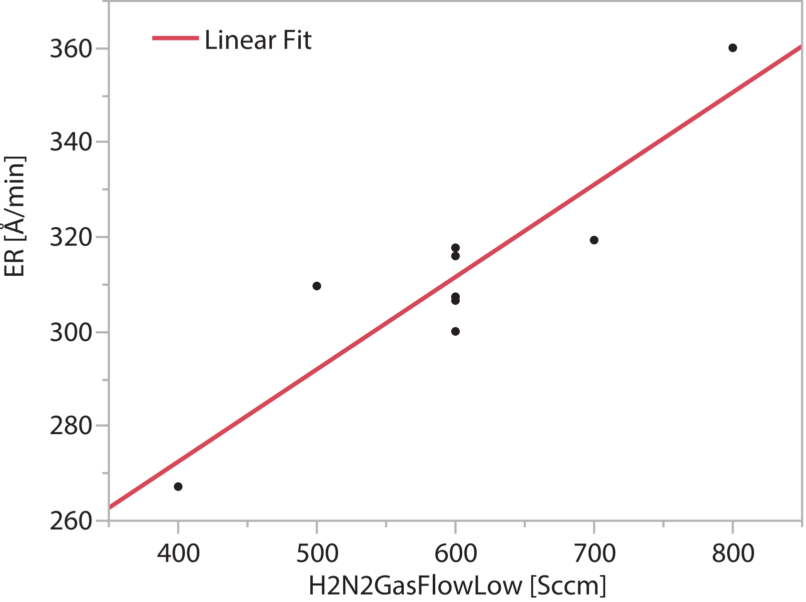


Figure C.2. Etch rate response with the H_2N_2 gas factor

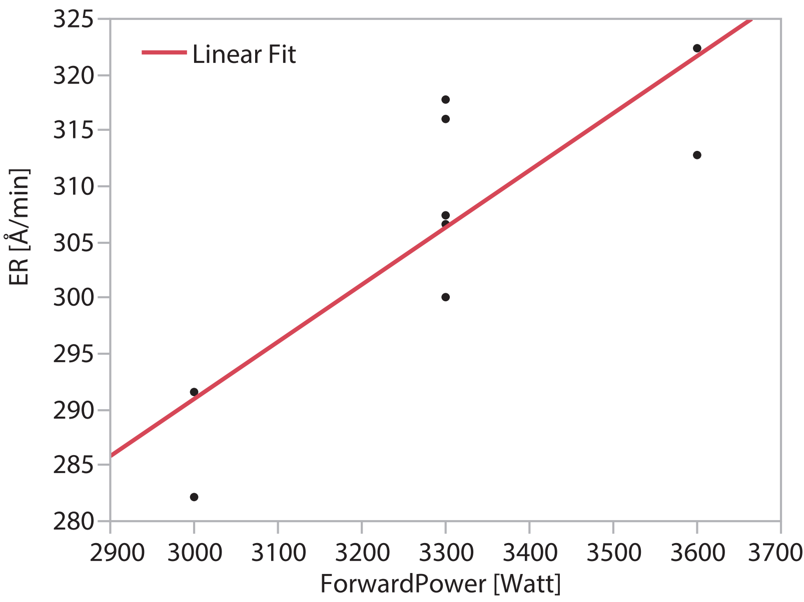


Figure C.3. Etch rate response with the RF forward power factor

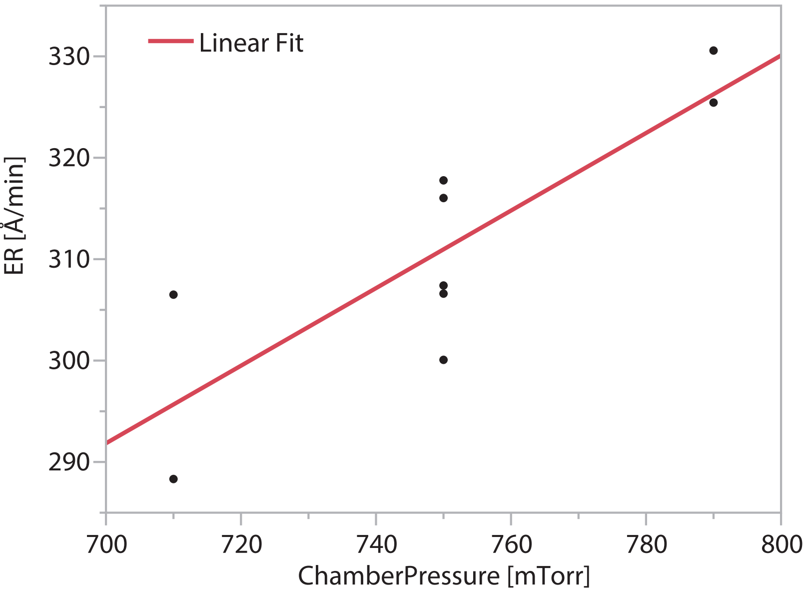


Figure C.4. Etch rate response with the chamber pressure factor

REFERENCES

[1] S. J. Qin, G. Cherry, R. Good, J. Wang, and C. A. Harrison, "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *Journal of Process Control*, vol. 16, no. 3, pp. 179–191, 2006.

[2] A. A. Khan, J. R. Moyne, and D. M. Tilbury, "Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares," *Journal of Process Control*, vol. 18, no. 10, pp. 961–974, 2008.

[3] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta*, vol. 185, pp. 1–17, 1986.

[4] S. Wold, A. Ruhe, H. Wold, and W. Dunn, III, "The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[5] M.-H. Hung, T.-H. Lin, F.-T. Cheng, R.-C. Lin *et al.*, "A novel virtual metrology scheme for predicting cvd thickness in semiconductor manufacturing," *IEEE/ASME Transactions on Mechatronics*, 2007.

[6] S. Verhaverbeke, I. Teerlinck, C. Vinckier, G. Stevens, R. Cartuyvels, and M. Heyns, "The etching mechanisms of sio2 in hydrofluoric acid," *Journal of The Electrochemical Society*, vol. 141, no. 10, pp. 2852–2857, 1994.

[7] M. Quirk and J. Serda, *Semiconductor manufacturing technology*. Prentice Hall Upper Saddle River, NJ, 2001, vol. 1.

[8] D. M. Manos and D. L. Flamm, *Plasma etching: an introduction*. Elsevier, 1989.

[9] J. E. Spencer, R. L. Jackson, and A. Hoff, "New directions in dry processing using the flowing afterglow of a microwave discharge," in *Proceedings of the Symposium for Plasma Processing*, vol. 87, no. 86, 1987, pp. 186–200.

[10] E. Degenkolb, C. Mogab, M. Goldrick, and J. Griffiths, "Spectroscopic study of radiofrequency oxygen plasma stripping of negative photoresists. i. ultraviolet spectrum," *Applied Spectroscopy*, vol. 30, no. 5, pp. 520–527, 1976.

[11] S. Dzioba, G. Este, and H. Naguib, "Decapsulation and photoresist stripping in oxygen microwave plasmas," *Journal of The Electrochemical Society*, vol. 129, no. 11, pp. 2537–2541, 1982.

[12] S. Fujimura, K. Shinagawa, M. T. Suzuki, and M. Nakamura, "Resist stripping in an o2+ h2o plasma downstream," *Journal of Vacuum Science & Technology B*, vol. 9, no. 2, pp. 357–361, 1991.

[13] V. Premachandran, "Etch rate enhancement of photoresist in nitrogen-containing plasmas," *Applied Physics Letters*, vol. 55, no. 24, pp. 2488–2490, 1989.

[14] T. Kyuho, T. Tsukihara, Q. Wang, M. Yamaoka, T. Motosue, and K. Kimura, "Device-level apc in ion implantation for analog device," in *Semiconductor Manufacturing, 2006. ISSM 2006. IEEE International Symposium on*. IEEE, 2006, pp. 110–113.

[15] J. S. Oakland, *Statistical process control*. Routledge, 2007.

[16] J. B. Keats and D. C. Montgomery, *Statistical process control in manufacturing*. Dekker, 1991.

[17] D. C. Montgomery, *Introduction to statistical quality control*. John Wiley & Sons, 2007.

[18] B. E. Goodlin, D. S. Boning, H. H. Sawin, and B. M. Wise, "Simultaneous fault detection and classification for semiconductor manufacturing tools," *Journal of the Electrochemical Society*, vol. 150, no. 12, pp. G778–G784, 2003.

[19] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *Semiconductor manufacturing, IEEE transactions on*, vol. 20, no. 4, pp. 345–354, 2007.

[20] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. Balakrishnan, and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica*, vol. 36, no. 11, pp. 1567–1603, 2000.

[21] J. Ebert, G. Crispieri, A. Emami, D. de Roover, and L. Porter, "Using model based fingerprinting to characterize process variations in an rf etch system," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[22] S. Kotz and N. L. Johnson, *Process capability indices*. CRC Press, 1993.

[23] S. Wang and T. Miller, "The generic r2r control strategy and controller performance monitoring," in *ISMI AEC/APC Symposium*, Austin, Texas, 2010.

[24] S. Wang, T. Miller, C. Thompson, and J. Zou, "A r2r control and preventive maintenance strategy for batch lpcvd furnace," in *APC Conference XXI*, Ann Arbor, Michigan, 2009.

[25] H. Purwins, B. Barak, A. Nagi, R. Engel, U. Hockele, A. Kyek, S. Cherla, B. Lenz, G. Pfeifer, and K. Weinzierl, "Regression methods for virtual metrology of layer thickness in chemical vapor deposition," *Mechatronics, IEEE/ASME Transactions on*, vol. 19, no. 1, pp. 1–8, 2014.

[26] J. Moyne, E. Del Castillo, and A. M. Hurwitz, *Run-to-run control in semiconductor manufacturing*. CRC press, 2000.

[27] W. J. Campbell, S. K. Firth, A. J. Toprac, and T. F. Edgar, "A comparison of run-to-run control algorithms," in *American Control Conference, 2002. Proceedings of the 2002*, vol. 3. IEEE, 2002, pp. 2150–2155.

[28] E. Sachs, A. Hu, A. Ingolfsson, and P. H. Langer, "Modeling and control of an epitaxial silicon deposition process with step disturbances," in *Advanced Semiconductor Manufacturing Conference and Workshop, 1991. ASMC 91 Proceedings. IEEE/SEMI 1991*. IEEE, 1991, pp. 104–107.

[29] S. R. A. Fisher, S. Genetiker, R. A. Fisher, S. Genetician, R. A. Fisher, and S. Généticien, *The design of experiments*. Oliver and Boyd Edinburgh, 1960, vol. 12, no. 6.

[30] S. J. Qin and T. A. Badgwell, "An overview of industrial model predictive control technology," in *AIChE Symposium Series*, vol. 93, no. 316. New York, NY: American Institute of Chemical Engineers, 1971-c2002., 1997, pp. 232–256.

[31] K. R. Muske and J. B. Rawlings, "Model predictive control with linear models," *AIChE Journal*, vol. 39, no. 2, pp. 262–287, 1993.

[32] C. V. Rao, J. B. Rawlings, and J. H. Lee, "Constrained linear state estimationa moving horizon approach," *Automatica*, vol. 37, no. 10, pp. 1619–1628, 2001.

[33] P. K. Findeisen, "Moving horizon state estimation of discrete time systems," Master's thesis, University of Wisconsin, Madison, 1997.

[34] J. Zou, J. A. Mullins, and K. A. Edwards, "Semiconductor run-to-run control system with state and model parameter estimation," Jun. 8 2004, uS Patent 6,748,280.

[35] X. Wang, S. Wu, and K. Wang, "A run-to-run profile control algorithm for improving the flatness of nano-scale products," *Automation Science and Engineering, IEEE Transactions on*, vol. 12, no. 1, pp. 192–203, 2015.

[36] S. Buso, L. Malesani, and P. Mattavelli, "Comparison of current control techniques for active filter applications," *Industrial Electronics, IEEE Transactions on*, vol. 45, no. 5, pp. 722–729, 1998.

[37] T. F. Coleman and Y. Li, "A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1040–1058, 1996.

[38] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Fluids Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[39] K. S. Miller and D. M. Leskiw, *An introduction to Kalman filtering with applications*. Krieger Publishing Company, 1987.

[40] J. Wang, Q. Peter He, and T. F. Edgar, "State estimation in high-mix semiconductor manufacturing," *Journal of Process Control*, vol. 19, no. 3, pp. 443–456, 2009.

[41] C. K. Hanish *et al.*, "Run-to-run state estimation in systems with unobservable states," in *Proceedings of AEC/APC Symposium XVII*, 2005.

[42] S. K. Firth, W. J. Campbell, A. Toprac, and T. F. Edgar, "Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 19, no. 3, pp. 298–315, 2006.

[43] N. S. Patel, "Model regularization for high-mix control," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 23, no. 2, pp. 151–158, 2010.

[44] S. Wang and M. Skliar, "A Novel Implementation of Non-Threaded Run to Run Controller at IM Flash Technologies," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[45] B. C. Kuo, *Digital control systems*. Holt, Rinehart and Winston, 1980.

[46] H. Sasano, W. Liu, D. Mui, K. Yoo, and J. Yamartino, "Advanced gate process critical dimension control in semiconductor manufacturing," in *Semiconductor Manufacturing, 2003 IEEE International Symposium on*. IEEE, 2003, pp. 382–385.

[47] S. Wang and M. Skliar, "Incorporation of multiphysics models into semiconductor virtual metrology," in *APC Conference XXVI*, Ann Arbor, Michigan, 2014.

[48] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and intelligent laboratory systems*, vol. 58, no. 2, pp. 109–130, 2001.

[49] S. Joe Qin, "Statistical process monitoring: basics and beyond," *Journal of chemometrics*, vol. 17, no. 8-9, pp. 480–502, 2003.

[50] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1, pp. 37–52, 1987.

[51] A. A. Khan, J. R. Moyne, and D. M. Tilbury, "An approach for factory-wide control utilizing virtual metrology," *IEEE Transactions on semiconductor Manufacturing*, vol. 20, no. 4, pp. 364–375, 2007.

[52] S. Lynn, J. Ringwood, E. Ragnoli, S. McLoone, and N. MacGearailty, "Virtual metrology for plasma etch using tool variables," in *Advanced Semiconductor Manufacturing Conference, 2009. ASMC'09. IEEE/SEMI*. IEEE, 2009, pp. 143–148.

[53] L. Leng, T. Zhang, L. Kleinman, and W. Zhu, "Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science," in *Journal of Physics: Conference Series*, vol. 78, no. 1. IOP Publishing, 2007, p. 012084.

[54] W. W. Chin, "The partial least squares approach to structural equation modeling," *Modern methods for business research*, vol. 295, no. 2, pp. 295–336, 1998.

[55] J.-B. Lohmöller, *Latent variable path modeling with partial least squares*. Springer Science & Business Media, 2013.

[56] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*. Springer, 2006, pp. 34–51.

[57] Y.-C. Su, F.-t. Cheng, G.-W. Huang, M.-H. Hung, and T. Yang, "A quality prognostics scheme for semiconductor and tft-lcd manufacturing processes," in *Industrial Electronics Society, 2004. IECON 2004. 30th Annual Conference of IEEE*, vol. 2. IEEE, 2004, pp. 1972–1977.

[58] C. Y.-C. Jonathan and F.-T. Cheng, "Application development of virtual metrology in semiconductor industry," in *Industrial Electronics Society, 2005. IECON 2005. 31st Annual Conference of IEEE*. IEEE, 2005, pp. 6–pp.

[59] F.-T. Cheng, H.-C. Huang, and C.-A. Kao, "Dual-phase virtual metrology scheme," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 20, no. 4, pp. 566–571, 2007.

[60] Y.-J. Chang, Y. Kang, C.-L. Hsu, C.-T. Chang, and T. Y. Chan, "Virtual metrology technique for semiconductor manufacturing," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 5289–5293.

[61] T.-H. Lin, M.-H. Hung, R.-C. Lin, and F.-T. Cheng, "A virtual metrology scheme for predicting cvd thickness in semiconductor manufacturing," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 1054–1059.

[62] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of artificial neural networks*. MIT press, 1997.

[63] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*. IEEE, 1989, pp. 593–605.

[64] D. Kriesel, "A brief introduction to neural networks," *Retrieved August*, vol. 15, p. 2011, 2007.

[65] M. Chester, *Neural networks: a tutorial*. Prentice-Hall, Inc., 1993.

[66] F. Harirchi, A. Subramanian, K. Poolla, and StirtonBroc, "Implementation of Nonthreaded Estimation for Run-to-Run Control of High Mix Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, no. 4, pp. 516–528, 2013.

[67] D. Zeng and C. J. Spanos, "Virtual metrology modeling for plasma etch operations," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 22, no. 4, pp. 419–431, 2009.

[68] K. Wang and J. Lin, "A run-to-run control algorithm based on timely and delayed mixed-resolution information," *International Journal of Production Research*, vol. 51, no. 15, pp. 4704–4717, 2013.

[69] L. Chen, M. Ma, S.-S. Jang, D. S.-H. Wang, and S. Wang, "Performance assessment of run-to-run control in semiconductor manufacturing based on imc framework," *International Journal of Production Research*, vol. 47, no. 15, pp. 4173–4199, 2009.

[70] S. Adivikolanu and E. Zafiriou, "Extensions and performance/robustness tradeoffs of the ewma run-to-run controller by using the internal model control structure," *Electronics Packaging Manufacturing, IEEE Transactions On*, vol. 23, no. 1, pp. 56–68, 2000.

[71] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: combining spc and feedback control," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 8, no. 1, pp. 26–43, 1995.

[72] K. Hui and J. Mou, "The devil in foundry apc," *Future Fab*, 2010.

[73] T. F. Edgar, S. Firth, C. Bode, and V. Martinenz, "Multi-product run-to-run control for high-mix fabs," *AEC/APC Asia, Hsinchu, Taiwan*, 2004.

[74] A. J. Pasadyn and T. F. Edgar, "Observability and state estimation for multiple product control in semiconductor manufacturing," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 18, no. 4, pp. 592–604, 2005.

[75] J. Wang, Q. P. He, and T. F. Edgar, "A general framework for state estimation in high-mix semiconductor manufacturing," in *American Control Conference, 2007. ACC'07*. IEEE, 2007, pp. 3636–3641.

[76] C. A. Bode, J. Wang, Q. He, and T. F. Edgar, "Run-to-run control and state estimation in high-mix semiconductor manufacturing," *Annual Reviews in Control*, vol. 31, no. 2, pp. 241–253, 2007.

[77] M.-D. Ma, C.-C. Chang, S.-S. Jang, and D. S.-H. Wong, "Mixed product run-to-run process control–an anova model with arima disturbance approach," *Journal of Process Control*, vol. 19, no. 4, pp. 604–614, 2009.

[78] A. V. Prabhu and T. F. Edgar, "A new state estimation method for high-mix semiconductor manufacturing processes," *Journal of Process Control*, vol. 19, no. 7, pp. 1149–1161, 2009.

[79] A. V. Prabhu, *Performance monitoring of run-to-run control systems used in semiconductor manufacturing*. ProQuest, 2008.

[80] Y. Zheng, Q.-H. Lin, D. S.-H. Wang, S.-S. Jang, and K. Hui, "Stability and performance analysis of mixed product run-to-run control," *Journal of Process Control*, vol. 16, no. 5, pp. 431–443, 2006.

[81] F. W. Fairman, *Linear control theory: the state space approach*. John Wiley & Sons, 1998.

[82] G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.

[83] A. Prado and Y. Feng, "Method and System for Estimating and Sharing State Offsets for Run to Run Control in Semiconductor," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[84] J. Stuber, "A Damascene Trench CD Control System for Plasma Etch," in *APC Conference XXVI*, Ann Arbor, Michigan, 2014.

[85] X. Wang, S. Wu, K. Wang, X. Deng, L. Liu, and Q. Cai, "A spatial calibration model for nanotube film quality prediction," *IEEE Transactions on Automation Science and Engineering*, vol. PP, no. 99, pp. 1–15, 2015.

[86] S. Gatzemeier, "25nm NAND : Run to Run Advanced Process Control," in *SEMICON Singapore*, Singapore, 2011.

[87] J. Zou, G. Tsai, S.-k. Neo, Z.-n. Ma, and K. Khu, "Minimizing pilot runs with non-threaded control technology," in *APC Conference XXVI*, Ann, 2014.

[88] J. Moyne, "International Technology Roadmap for Semiconductors (ITRS): Factory Integration (FI) Revisions for 2013; Prediction, Big Data and Control Systems," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[89] M. Kano, K. Miyazaki, S. Hasebe, and I. Hashimoto, "Inferential control system of distillation compositions using dynamic partial least squares regression," *Journal of Process Control*, vol. 10, no. 2, pp. 157–166, 2000.

[90] K. R. Beebe, R. J. Pell, and M. B. Seasholtz, "Chemometrics: a practical guide," *Chemometrics*, 1998.

[91] E. R. Malinowski, *Factor analysis in chemistry*. Wiley, 2002.

[92] G. ElMasry, D.-W. Sun, and P. Allen, "Near-infrared hyperspectral imaging for predicting colour, ph and tenderness of fresh beef," *Journal of Food Engineering*, vol. 110, no. 1, pp. 127–140, 2012.

[93] F.-T. Cheng, J.-C. Chang, H.-C. Huang, C.-A. Kao, Y.-L. Chen, and J.-L. Peng, "Benefit model of virtual metrology and integrating avm into mes," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 24, no. 2, pp. 261–272, 2011.

[94] P. Chen, S. Wu, J. Lin, F. Ko, H. Lo, J. Wang, C. Yu, and M. Liang, "Virtual metrology: A solution for wafer to wafer advanced process control," in *Semiconductor Manufacturing, 2005. ISSM 2005, IEEE International Symposium on*. IEEE, 2005, pp. 155–157.

[95] V. C. Klema and A. J. Laub, "The singular value decomposition: Its computation and some applications," *Automatic Control, IEEE Transactions on*, vol. 25, no. 2, pp. 164–176, 1980.

[96] T. Byrne, "Implementing virtual metrology from the ground up," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[97] B. Lu, J. Stuber, and T. F. Edgar, "Integrated virtual metrology and fault detection in plasma etch tools," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.

[98] Y.-T. Huang, F.-T. Cheng, and Y.-T. Chen, "Importance of data quality in virtual metrology," in *IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics*. IEEE, 2006, pp. 3727–3732.

[99] Y.-C. Su, T.-H. Lin, F.-T. Cheng, and W.-M. Wu, "Implementation considerations of various virtual metrology algorithms," in *Automation Science and Engineering, 2007. CASE 2007. IEEE International Conference on*. IEEE, 2007, pp. 276–281.

[100] J. Moyne, "Keeping up with the latest advancements in apc : A tutorial ch3: The prediction landscape," in *APC Conference XXV 2013*, Ann Arbor, Michigan, 2013.

[101] A. Ferreira, A. Roussy, and L. Condé, "Virtual metrology models for predicting physical measurement in semiconductor manufacturing," in *Advanced Semiconductor Manufacturing Conference, 2009. ASMC'09. IEEE/SEMI*. IEEE, 2009, pp. 149–154.

[102] G. A. Susto, A. Beghi, and C. De Luca, "A virtual metrology system for predicting cvd thickness with equipment variables and qualitative clustering," in *Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on*. IEEE, 2011, pp. 1–4.

[103] B. Gill, T. F. Edgar, and J. Stuber, "A novel approach to virtual metrology using kalman filtering," in *Proceedings of SEMATECH Advanced Equipment Control/Advanced Process Control Symposium XXII*, 2010, pp. 267–291.

[104] P. Kang, D. Kim, H.-j. Lee, S. Doh, and S. Cho, "Virtual metrology for run-to-run control in semiconductor manufacturing," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2508–2522, 2011.

[105] T.-H. Pan, B.-Q. Sheng, D. S.-H. Wong, and S.-S. Jang, "A virtual metrology model based on recursive canonical variate analysis with applications to sputtering process," *Journal of Process Control*, vol. 21, no. 6, pp. 830–839, 2011.

[106] G. Li, D. Zhou, and S. J. Qin, "Output-relevant fault reconstruction based on total pls," in *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*. IEEE, 2010, pp. 1718–1722.

[107] F.-T. Cheng, Y.-T. Chen, Y.-C. Su, and D.-L. Zeng, “Method for evaluating reliance level of a virtual metrology system,” in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1590–1596.

[108] J. S. Judge, “A study of the dissolution of sio2 in acidic fluoride solutions,” *Journal of the Electrochemical Society*, vol. 118, no. 11, pp. 1772–1775, 1971.

[109] E. Högfeldt, *Stability constants of metal-ion complexes: part A: inorganic ligands*. Pergamon Pr, 1982, vol. 21.

[110] D. A. Skoog, D. M. West, F. J. Holler *et al.*, *Analytical chemistry*. Holt, Rinehart and Winston New York, NY, 1979.

[111] F.-T. Cheng, Y.-T. Chen, Y.-C. Su, and D.-L. Zeng, “Evaluating reliance level of a virtual metrology system,” *Semiconductor Manufacturing, IEEE Transactions on*, vol. 21, no. 1, pp. 92–103, 2008.

[112] C.-A. Kao, F.-T. Cheng, W.-M. Wu, F.-W. Kong, and H.-H. Huang, “Run-to-run control utilizing virtual metrology with reliance index,” *Semiconductor Manufacturing, IEEE Transactions on*, vol. 26, no. 1, pp. 69–81, 2013.

[113] J. Besnard and A. Toprac, “Wafer-to-wafer virtual metrology applied to run-to-run control,” in *Proceedings of the 3rd ISMI symposium on manufacturing effectiveness, USA*, 2006.

[114] P. Smeys, *Local Oxidation of Silicon for Insulation*. Stanford University, 1996.

[115] A. Bhattacherjee, W. Koutny, R. Shrivastava, and T. J. Rodgers, “Rapid thermal nitridized oxide locos process,” Aug. 16 1988, uS Patent 4,764,248.

[116] W. Jost, *Diffusion in solids, liquids, gases*. Academic press, 1960.

[117] P. Packan and J. Plummer, “Transient diffusion of low-concentration b in si due to 29si implantation damage,” *Applied physics letters*, vol. 56, no. 18, pp. 1787–1789, 1990.

[118] J. D. Plummer, *Silicon VLSI technology: fundamentals, practice, and modeling*. Pearson Education India, 2000.

[119] B. G. Streetman and S. Banerjee, *Solid state electronic devices*. Prentice Hall New Jersey, 2000, vol. 4.

[120] B. E. Deal, “The oxidation of silicon in dry oxygen, wet oxygen, and steam,” *Journal of The Electrochemical Society*, vol. 110, no. 6, pp. 527–533, 1963.

[121] R. Levy, *Microelectronic materials and processes*. Springer Science & Business Media, 2012.

[122] I. Asahi, A. Ito, S. Yamada, A. Abe, K. Furuta, and M. Yamakita, “Diffusion furnace control using a learning control method,” in *Industrial Electronics, Control and Instrumentation, 1991. Proceedings. IECON’91., 1991 International Conference on*. IEEE, 1991, pp. 1941–1945.

[123] B. Lin, N. S. Patel, and J. Boone, “Poly silicon deposition process improvement on 300 mm wafers (pc23),” in *Semiconductor Manufacturing, 2003 IEEE International Symposium on*. IEEE, 2003, pp. 119–122.

[124] S. Shinde, A. Sonar, and Y. Sun, “Advanced process control for furnace systems in semiconductor manufacturing,” in *Advanced Semiconductor Manufacturing Conference (ASMC), 2013 24th Annual SEMI*. IEEE, 2013, pp. 275–279.

[125] J. B. Rawlings, “Tutorial: Model predictive control technology,” in *American Control Conference, 1999. Proceedings of the 1999*, vol. 1. IEEE, 1999, pp. 662–676.

[126] B. Van Schravendijk, W. De Koning, and W. Nuijen, “Modeling and control of the wafer temperatures in a diffusion furnace,” *Journal of applied physics*, vol. 61, no. 4, pp. 1620–1627, 1987.

[127] C. A. Bode and A. J. Toprac, “Method and apparatus for modeling thickness profiles and controlling subsequent etch process,” Jun. 25 2002, uS Patent 6,410,351.

[128] S. Hirasawa, S. Kieda, T. Watanabe, T. Torii, T. Takagaki, and T. Uchino, “Temperature distribution in semiconductor wafers heated in a vertical diffusion furnace,” *Semiconductor Manufacturing, IEEE Transactions on*, vol. 6, no. 3, pp. 226–232, 1993.

[129] J. R. Howell, R. Siegel, and M. P. Menguc, *Thermal radiation heat transfer*. CRC press, 2010.

[130] M. A. Lieberman and A. J. Lichtenberg, “Principles of plasma discharges and materials processing,” *MRS Bulletin*, vol. 30, pp. 899–901, 1994.

[131] P. Williams, *Plasma processing of semiconductors*. Springer Science & Business Media, 2013, vol. 336.

[132] C. G. Willson, R. R. Dammel, and A. Reiser, “Photoresist materials: a historical perspective,” in *Microlithography'97*. International Society for Optics and Photonics, 1997, pp. 28–41.

[133] J. Coburn and H. F. Winters, “Ion-and electron-assisted gas-surface chemistryan important effect in plasma etching,” *Journal of Applied physics*, vol. 50, no. 5, pp. 3189–3196, 1979.

[134] D. L. Flamm, V. M. Donnelly, and J. A. Mucha, “The reaction of fluorine atoms with silicon,” *Journal of Applied Physics*, vol. 52, no. 5, pp. 3633–3639, 1981.

[135] B. A. Heath, “Selective reactive ion beam etching of sio2 over polycrystalline si,” *Journal of The Electrochemical Society*, vol. 129, no. 2, pp. 396–402, 1982.

[136] V. M. Donnelly, D. L. Flamm, W. Dautremont-Smith, and D. Werder, “Anisotropic etching of sio2 in low-frequency cf4/o2 and nf3/ar plasmas,” *Journal of applied physics*, vol. 55, no. 1, pp. 242–252, 1984.

[137] R. Petri, D. Henry, and N. Sadeghi, "Tungsten etching mechanisms in low-pressure sf6 plasma," *Journal of applied physics*, vol. 72, no. 7, pp. 2644–2651, 1992.

[138] T. Mayer and R. Barker, "Simulation of plasma-assisted etching processes by ion-beam techniques," *Journal of Vacuum Science & Technology*, vol. 21, no. 3, pp. 757–763, 1982.

[139] J. Coburn, "Surface-science aspects of plasma-assisted etching," *Applied Physics A Solids and Surfaces*, vol. 59, no. 5, pp. 451–458, 1994.

[140] D. C. Gray, I. Tepermeister, and H. H. Sawin, "Phenomenological modeling of ion-enhanced surface kinetics in fluorine-based plasma etching," *Journal of Vacuum Science & Technology B*, vol. 11, no. 4, pp. 1243–1257, 1993.

[141] J. P. McVittie, J. C. Rey, A. Bariya, M. IslamRaja, L. Cheng, S. Ravi, and K. C. Saraswat, "Speedie: A profile simulator for etching and deposition," in *Santa Cl-DL tentative*. International Society for Optics and Photonics, 1991, pp. 126–138.

[142] C. Chang, J. McVittie, K. Saraswat, and K. Lin, "Backscattered deposition in ar sputter etch of silicon dioxide," *Applied physics letters*, vol. 63, no. 16, pp. 2294–2296, 1993.

[143] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.

[144] R. B. Rao, G. Fung, and R. Rosales, "On the dangers of cross-validation. an experimental evaluation." in *SDM*. SIAM, 2008, pp. 588–596.

[145] S. J. Qin, "Recursive pls algorithms for adaptive data modeling," *Computers & Chemical Engineering*, vol. 22, no. 4, pp. 503–514, 1998.

[146] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[147] G. Turban and M. Rapeaux, "Dry etching of polyimide in o 2-cf 4 and o 2-sf 6 plasmas," *Journal of the Electrochemical Society*, vol. 130, no. 11, pp. 2231–2236, 1983.

[148] K. R. Williams and R. S. Muller, "Etch rates for micromachining processing," *Microelectromechanical Systems, Journal of*, vol. 5, no. 4, pp. 256–269, 1996.

[149] J. F. Battey, "The effects of geometry on diffusion-controlled chemical reaction rates in a plasma," *Journal of The Electrochemical Society*, vol. 124, no. 3, pp. 437–441, 1977.

[150] G. Agrawal, M. Yelverton, and T. Miller, "Dynamic run-to-run state estimation," in *APC Conference XXV*, Ann Arbor, Michigan, 2013.