Copyright
by
Xiaojing Jiang
2012
The Dissertation Committee for Xiaojing Jiang Certifies that this is the approved version of the following dissertation:

Control Performance Assessment of Run-to-run Control System Used in High-mix Semiconductor Manufacturing

Committee:

Thomas F. Edgar, Supervisor
Isaac C. Sanchez
Glenn Y. Masada
John D. Stuber
Jin Wang
Gyeong S. Hwang
Control Performance Assessment of Run-to-run Control System Used in High-mix Semiconductor Manufacturing

by

Xiaojing Jiang, B.E.; M.E.

Dissertation
Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
August, 2012
Dedicated to my families
Acknowledgements

First of all, I wish to thank Dr. Edgar for providing me an opportunity to pursue my PhD degree at the University of Texas. He has been so nice and patient with me and has kept me on track through the last four years.

I would also like to thank Dr. Stuber, Dr. Wang, Dr. Sanchez, Dr. Masada, and Dr. Hwang for serving as my committee members. In addition, I thank all the professors at the University of Texas who enhanced my knowledge of the process control and chemical engineering area.

This dissertation is finally supported by the National Science Foundation and Texas Instruments under GOALI Grant CBET-0854033. I thank Dr. Stuber and other engineers in Texas Instruments for their help on my research.

My lab mates form Dr. Edgar’s group have been such a great source of knowledge and distractions in academic research and happiness in life. I really enjoyed the time in this warm and multicultural group. I thank all the help from those brilliant people from different countries. I would especially thank Dr. Ricardo Dunia, Ivan, Ben Spivey, Bhalinder Gill, Doug French, Kody Powell, Kriti Kapoor, Ramiro Palma, Wesley Cole, Ahn Nguyen, and Jong Suk Kim. I wish all of them good luck in their future research, careers, and life.

The last four years in Austin has been made so enjoyable by a wonderful roommate Yingying Jiang. I also had a great time with Ye Chen, Wei Xie, Chao Wang, Yunping Fei and many other friends. I will miss you all and wish you all good luck and happy.
Lastly, I would like to thank my family for supporting me all these years. I am grateful to all the love and support from my parents. Thank you, Mum. You are my source of strength and inspiration, role model and friend. My elder sister, Tian, has been great company all these years. The love and help from her has made me feel so happy and warm. I hope all the dreams of my families will come true.
Control Performance Assessment of Run-to-run Control System Used in High-mix Semiconductor Manufacturing

Xiaojing Jiang, Ph.D.
The University of Texas at Austin, 2012

Supervisor: Thomas F. Edgar

Control performance assessment (CPA) is an important tool to realize high performance control systems in manufacturing plants. CPA of both continuous and batch processes have attracted much attention from researchers, but only a few results about semiconductor processes have been proposed previously. This work provides methods for performance assessment and diagnosis of the run-to-run control system used in high-mix semiconductor manufacturing processes.

First, the output error source of the processes with a run-to-run EWMA controller is analyzed and a CPA method (namely CPA I) is proposed based on closed-loop parameter estimation. In CPA I, ARMAX regression is directly applied to the process output error, and the performance index is defined based on the variance of the regression results. The influence of plant model mismatch in the process gain and disturbance model parameter to the control performance in the cases with or without set point change is studied. CPA I method is applied to diagnose the plant model mismatch in the case with set point change.
Second, an advanced CPA method (namely CPA II) is developed to assess the control performance degradation in the case without set point change. An estimated disturbance is generated by a filter, and ARMAX regression method is applied to the estimated disturbance to assess the control performance. The influence of plant model mismatch, improper controller tuning, metrology delay, and high-mix process parameters is studied and the results showed that CPA II method can quickly identify, diagnose and correct the control performance degradation.

The CPA II method is applied to industrial data from a high-mix photolithography process in Texas Instruments and the influence of metrology delay and plant model mismatch is discussed. A control performance optimization (CPO) method based on analysis of estimated disturbance is proposed, and optimal EWMA controller tuning factor is suggested.

Finally, the CPA II method is applied to non-threaded run-to-run controller which is developed based on state estimation and Kalman filter. Overall process control performance and state estimation behavior are assessed. The influence of plant model mismatch and improper selection of different controller variables is studied.
# Table of Contents

List of Tables .......................................................................................................................... xii  
List of Figures ............................................................................................................................ xiii  

## Chapter 1  Introduction ................................................................. 1  
1.1 Run to Run Control ................................................................................. 1  
   1.1.1 Overview of Run to Run Control ....................................................... 1  
   1.1.2 EWMA and Extensions ................................................................. 3  
   1.1.3 Kalman Filter and Extension .......................................................... 6  
   1.1.4 Other algorithms ........................................................................... 8  
1.2 High-mix Semiconductor Manufacturing Processes .................................. 9  
   1.2.1 Threaded Control ........................................................................... 10  
   1.2.2 Non-threaded Control ................................................................. 11  
1.3 Control Performance Assessment (CPA) ............................................... 14  
   1.3.1 Overview of Control Performance Monitoring and Assessment .......... 14  
   1.3.2 Harris’ type Performance Indices and Its Extension; .................... 16  
   1.3.3 CPA/CPM of Semiconductor Manufacturing Processes; .................. 21  
1.4 Overview of The Dissertation .................................................................... 23  

## Chapter 2  Control Performance Assessment (CPA) Based on Parameter Estimation .................. 25  
2.1 EWMA Controller System ...................................................................... 25  
2.2 ERROR Source Analysis ........................................................................... 26  
2.3 Optimal EWMA Controller ..................................................................... 28  
2.4 CPA (I) Based on Closed-loop Parameter Estimation .............................. 29  
2.5 CPA (I) in Threaded EWMA Controller System with Constant Set Point .... 31  
2.6 CPA (I) in Non-threaded EWMA Controller System ............................. 34  
   2.6.1 Varied Set points with Plant Model Mismatch .............................. 34  
   2.6.2 Varied Set points Without Plant Model Mismatch .......................... 38  
   2.6.3 Varied Set Points with Improper Tuning ........................................ 42  
   2.6.4 The Selection of ARMAX Regression Model ................................. 44  
   2.6.5 Disturbance Model Different from IMA ......................................... 47  
2.7 Conclusion and Discussion ...................................................................... 48
Chapter 3 Control Performance Assessment BASED on filter design .......... 50
  3.1 MATHEMATICAL Foundation of CPA II .................................................. 50
  3.1.1 Error Source Analysis ........................................................................ 50
  3.1.2 Design of Q Filter .................................................................................. 52
  3.1.3 Control Performance Assessment Based on Q Filter (CPA II) .......... 54
  3.1.4 Improved ARMAX Regression ......................................................... 56
  3.2 Simulation Results of CPA ..................................................................... 59
    3.2.1: Incorrect Tuning ............................................................................. 60
    3.2.2: Plant Model Mismatch .................................................................. 61
    3.2.3: High-Mix Parameters ................................................................... 62
    3.2.4: Metrology Delay ............................................................................ 63
  3.3 Application of Direct ARMAX Regression ........................................... 64
  3.4 Summary ................................................................................................. 68

Chapter 4 CPA and CPD in Threaded Control System ......................... 70
  4.1 TI data Analysis with CPA II ................................................................. 70
    4.1.1 Model Used in TI ............................................................................ 70
    4.1.2 Analysis of Industrial Data .............................................................. 71
  4.2 Determination of Optimal Range in CPA II ........................................ 76
    4.2.1 Inaccurate Estimation of Metrology Delay .................................... 77
    4.2.2 Varied Metrology Delay .................................................................. 77
    4.2.3 Plant Model Mismatch in The Case with Metrology Delay .......... 81
  4.3 Control Performance Optimization ...................................................... 83
  4.4 Summary and Conclusion: .................................................................... 86

Chapter 5 CPA for non-threaded high-mix control system .................. 88
  5.1 Non-threaded Control of High-mix Semiconductor Manufacturing System ...... 88
  5.2 General Idea about Non-threaded Control ........................................... 90
  5.3 Simulation of CPA II for Non-threaded Control .................................. 92
    5.3.1 Plant Model Mismatch .................................................................. 93
    5.3.2 Improper Selection of θ ................................................................. 95
    5.3.3 Improper Selection of R or Q ......................................................... 97
  5.4 Summary ............................................................................................... 98

Chapter 6 Conclusions and Future Work ............................................. 100
6.1 Summary of Research ........................................................................................................ 100
6.2 Recommendations to Future Work ...................................................................................... 102
  6.2.1 Estimated Disturbance Computation .............................................................................. 102
  6.2.2 Online Adaptive Tuning ................................................................................................ 103
  6.2.3 Control Performance Diagnosis ..................................................................................... 104
Appendix A ................................................................................................................................ 106
Appendix B ................................................................................................................................ 109
Bibliography ............................................................................................................................ 110
Vita ........................................................................................................................................... 125
**List of Tables**

Table 2-1 Parameters of Simulation 2-1 ................................................. 31
Table 2-2. Process parameters used in Simulation 2-2 ......................... 34
Table 2-3 Process parameters used in Simulation 2-3 ......................... 39
Table 2-4. Process parameters used in Simulation 2-4 ......................... 42
Table 3-1 Q filter and estimated disturbance of different cases .......... 53
Table 3-2 Parameters of Simulation 3-1 ................................................. 58
Table 3-3 Parameters used in Simulations 3-6, 7, 8, 9, 10 ..................... 65
Table 4-1 Error Analysis of different reticles in Example 4-3 ............... 76
Table 4-2 Parameters used in Simulation 4-1 ........................................ 77
Table 4-3 Parameters used in Simulation 4-2 ........................................ 78
Table 4-4 Process output variance of Simulation 4-2 and 4-3 .............. 79
Table 4-5 Parameters used in Simulation 4-4 ........................................ 82
Table 4-6 Suggested $\omega$ and new variances in Simulation 4-6 .......... 84
Table 5-1 Parameters set up in Simulation 5-1, 2, 3, 4 .......................... 92
Table 5-2 Parameter change after $k=500$ in simulations 5-1 to 5-4 .... 92
LIST OF FIGURES

Figure 1-1 Diagram of High-mix processes.............................................. 10
Figure 2-1. Block diagram of a RtR controller in IMC structure ............. 27
Figure 2-2 Simulated process data and output variance in Simulation 2-1 ........................................................................................................... 32
Figure 2-3. Estimated gain in simulation 2-1............................................. 33
Figure 2-4. On-line monitoring of Simulation 2-1........................................ 33
Figure 2-5. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-2 ........................................................................................................... 35
Figure 2-6. Estimated $(\beta/b -1)$ in Simulation 2-2............................ 36
Figure 2-7. Estimated $1-\omega*(\beta/b)$ and PI1 in Simulation 2-2 .......... 37
Figure 2-8. Estimated $\theta / [1-\omega*(\beta/b)]$ in Simulation 2-2............. 37
Figure 2-9. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-3 ........................................................................................................... 40
Figure 2-10. Estimated $1-\omega*(\beta/b)$ and PI1 in Simulation 2-3 .......... 40
Figure 2-11. Estimated $\theta / [1-\omega*(\beta/b)]$ in Simulation 2-3 ............. 41
Figure 2-12. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-4........................................................................................................... 43
Figure 2-13. Estimated $1-\omega*(\beta/b)$ (a) and PI1(b) in Simulation 2-4 ...... 43
Figure 2-14. Estimated $(\beta/b-1)(11-a)$ and $\theta / [1-\omega*(\beta/b)] (11-b)$ in Simulation 2-4........................................................................................................... 44
Figure 2-15. Estimated $1-\omega*(\beta/b)$ (a), $PI1(b)$, and $\theta / [1-\omega*(\beta/b)] (c)$ in Simulation 2-5................................................................. 45
Figure 2-16. Estimated $1-\omega*(\beta/b)$ (a) and PI1 (b) in Simulation 2-6 ...... 46
Figure 2-17 Estimated $\beta/b-1$ (a) and $\theta / [1-\omega*(\beta/b)] (b)$ in Simulation 2-6 ................................................................. 46
Figure 2-18 Estimated gain in simulation 2-7............................................ 48
Figure 3-1 Influence of forgetting coefficient in ARMAX regression (Simulation 3-1)......................................................................................... 59
Figure 3-2. ARMAX regression results of output error and estimated disturbance in Simulation 3-2. (a). Result of $-C(2)$ from CPA I; b) Result of M from CPA II ................................................................. 60
Figure 3-3. ARMAX regression results of output error and estimated disturbance in Simulation 3-3. (a). Result of $-C(2)$ from CPA I; b) Result of $M$ from CPA II ................................................................. 61

Figure 3-4. ARMAX regression results of output error and estimated disturbance in Simulation 3-4. (a). Result of $-C(2)$ from CPA I; b) Result of $M$ from CPA II ................................................................. 62

Figure 3-5. ARMAX regression results of output error and estimated disturbance in Simulation 3-5. (a). Result of $-C(2)$ from CPA I; b) Result of $M$ from CPA II ................................................................. 64

Figure 3-6. Variance of process output error vs. Run Number $i$ (Simulation 3-5) ................................................................................................................................. 64

Figure 3-7. Parameter estimation by directly ARMAX regression ((a), Simulation 3-6-a; (b) Simulation 3-6-b) ................................................................. 66

Figure 3-8. Parameter estimation by directly ARMAX regression ((a), Simulation 3-7-a; (b) Simulation 3-7-b) ................................................................. 67

Figure 4-1 Process output error (a) and $M$ calculated from CPA II (b) of Example 4-1 ................................................................................................................................. 72

Figure 4-2 Process output error (a) and $M$ calculated from CPA II (b) of Example 4-2 ................................................................................................................................. 73

Figure 4-3. Process output error (a) and $M$ calculated from CPA II (b) of Example 4-3 ................................................................................................................................. 74

Figure 4-4. Influence of estimated metrology delay (Example 4-3) ....... 75

Figure 4-5. Process output error of Example 4-3 (Solid squares indicate data from Reticle E575-599-001) ................................................................................................................................. 76

Figure 4-6. $M$ of Simulation 4-1 with different estimated metrology delay ................................................................................................................................. 78

Figure 4-7. $M$ of Simulation 4-2 (Varied metrology delay) .............. 79

Figure 4-8. $M$ of Simulation 4-2 ($\omega = 0.24$, (a). CPA II method; (b) Modified CPA II method) ................................................................................................................................. 80

Figure 4-9. $M$ of Simulation 4-3 (d is selected from 0 to 10. (a) $\omega = 0.15$; (b) $\omega = 0.2$) ................................................................................................................................. 81

Figure 4-10. Influence of model plant mismatch to the optimal range of $M$ in CPA II (Simulation 4-4) ................................................................................................................................. 82

Figure 4-11. Variance of estimated disturbance vs. Qfilter parameter adjustments ................................................................................................................................. 84

Figure 5-1 Process output and CPA II analysis of overall process (Simulation 5-1) ................................................................................................................................. 93
Figure 5-2 Actual value, estimated value and CPA II analysis result of every single state variable (Simulation 5-1) .......................................................... 94
Figure 5-3 Process output and CPA II analysis of overall process (Simulation 5-2) ................................................................................................. 95
Figure 5-4 Actual value, estimated value and CPA II analysis result of every single state variable (Simulation 5-2) .................................................. 96
Figure 5-5. Process output and CPA II analysis of overall process (Simulation 5-3) ................................................................................................. 98
Figure 5-6. Process output and CPA II analysis of overall process (Simulation 5-4) ................................................................................................. 98
CHAPTER 1
INTRODUCTION

1.1 RUN TO RUN CONTROL

1.1.1 Overview of Run to Run Control

Run-to-run control is one category of Advanced Process Control (APC). In order to assess control performance, statistical methods can be applied to monitor a process output. A basic method called statistical process control (SPC) uses Western Electric rules[1] and other statistical attributes such as mean and variance parameters to detect excursions from an assumed Gaussian disturbance. However, SPC is only available to determine when such excursion happens, but cannot correct such an event using manipulated variables. The development of APC solved this limitation of SPC by suggesting corrective action mathematical algorithms. This reduces the process variability and the control performance could be improved.

Run-to-run control belongs to one category of APC where during the batch measurements are not available. In most semiconductor manufacturing processes, real-time measurement cannot normally be achieved and only ex-situ measurements of product quality characteristics can be obtained. As a result, the recipe of one batch reaction will be decided at the beginning of the reaction. After the reaction, product quality characteristics, such as etching depth and width, position and so on, can be measured, and then the results will be used in a mathematical algorithm and the recipe of next run can be calculated.

Run-to-run control is based on an assertion that the variation of the process results from a deterministic component which can be compensated using available knowledge and measurement of the process. Recipe adjustment of a run-to-run controller is made by special algorithms, based on mathematical process models and measurements of the process. In run-to-run control, measurements of key quality components are made at the
end of every run rather than real time measurements during the process that is why it is called run-to-run.

Run-to-run controller was first developed by Sachs and co-workers[2, 3] with the Exponentially Weighted Moving Average (EWMA) controller. In EWMA controller system, the process disturbance is considered as an integrated moving average (IMA) time series, and as a result, EWMA filter is used to remove the random noise and capture the deterministic trends of process disturbance. If the disturbance is not an IMA model, other filters will be applied. Later, other controllers such as double-EWMA controller, Predictor Corrector Controller (PCC)[4, 5] and application of IMC controller[6-8] were developed based on different disturbance types. Several researchers have concentrated on developing other filtering methods, and Kalman filter[9-11] or other modified filters based on Kalman’s theory were developed, especially for Multiple-Input-Multiple-Output (MIMO) run-to-run processes.

In the following section, different filtering and control methods in run-to-run control system will be reviewed.

Algorithms used in run-to-run control generally consist of three basic elements: the process model, the controller and the observer. The process model is generally a simplified model of the process, and it depends on the knowledge of the process. For example, many semiconductor manufacturing processes can be described by a steady state linear model. The controller is related to the process model. When the process model is linear, the controller is the inverse of the model.

The observer is the most important part of a run-to-run controller. The main difference between different run-to-run controllers is the selection of observer. In the beginning, EWMA filter was used as the filter of run-to-run (RtR) controller, later, different observer or filters were developed. More work about run-to-run control in semiconductor manufacturing processes can be found in Moyne[12].
1.1.2 EWMA and Extensions

The exponentially weighted moving average (EWMA) theory was introduced to quality control in by Roberts[13], and then it was widely used in the field of process monitoring and quality control chart[14-16]. The first EWMA filter was developed by Sachs[2]. It was used in feedback control system of semiconductor manufacturing processes[3]. Different types of EWMA filters are used based on difference process disturbance model.

The basic EWMA filter is developed based on an assertion that the process disturbance can be expressed by an Integrated Moving Average (IMA) model. EWMA theory can be applied to the process disturbance or process gain estimation. Sachs’ run-to-run controller operated in three modes[3]: Gradual Mode (GM), Rapid Mode(RM) and Generalized SPC. GM was used to compensate the gradual process drift which is usually described by IMA model; RM is used to detect the occurrence and magnitude of process shifts and recommend appropriate control action; Generalized SPC is used to distinguish the process shift and drift, and to determine whether RM mode should be applied. The key mathematical part of EWMA filter is GM mode.

In many semiconductor manufacturing processes, the real process can be expressed by a linear model:

\[ y_k = \beta u_k + \alpha_k \]  \hspace{1cm} (1-1)

Here \( y_k \) is the process output, \( u_k \) is the process input. \( \beta \) is the actual process gain, and \( \alpha_k \) is the process disturbance. When GM mode is applied, the process disturbance is considered as IMA time series:

\[ \alpha_k = \alpha_{k-1} + \epsilon_k - \theta \epsilon_{k-1} \]  \hspace{1cm} (1-2)

where \( \epsilon_k \) is white noise, and \( \theta \) is the IMA model parameter. The mathematical model used in run-to-run control algorithm is

\[ \hat{y}_k = b u_k + \alpha_k \]  \hspace{1cm} (1-3)
where $\hat{y}_k$ is the estimated output, $b$ is the nominal process gain, and $a_k$ is the observed process disturbance which is estimated by an EWMA filter:

$$a_{k+1} = \omega(y_k - bu_k) + (1 - \omega)a_k$$

(1-4)

Here $\omega$ is an adjustable filter parameter, which acts as the tuning parameter for the EWMA algorithm. Usually, $0 \leq \omega \leq 1$. The optimal tuning factor is related to the model-plant mismatch, metrology delay and the value of $\theta$. In most semiconductor manufacturing processes where metrology delay cannot be avoided, the optimal value of $\omega$ is usually selected between 0.1 to 0.3. Detailed discussion about the optimal $\omega$ will be presented in Chapter 2 and 3.

The control law used to determine the process input of next run is the inverse of the plant model:

$$u_k = \frac{1}{b}(r_k - a_k)$$

(1-5)

where $r_k$ is the target or set point of the $k^{th}$ run.

As the process variations are attributed to multiple factors in real chemical processes, therefore the EWMA algorithm should be implemented to estimate and update the bias from different factors. In semiconductor manufacturing processes, those biases may be attributed to different layers, different tools, different products and so on. Manufacturing context is used to describe the information about layers, tools, products and so on. EWMA filter is then applied to the runs with same manufacturing contexts, which is considered as the same “thread”. This algorithm is called “threaded control”, and more discussion about this will be shown in the following section about high-mix processes.

When the disturbance is considered as the sum of an IMA time series and some other dominant drift, some corrective method should be applied, and this method is called Predictor Corrector Control (PCC). PCC was first developed by Butler[17] and Stefani[18]. With this methodology, an extra exponential filter is used to estimate the
current process output, and as a result, the trend in the process output can be detected and corrected. As two coupled EWMA controllers are applied in this PCC method, it’s also called “double-EWMA” (d-EWMA) controller[19].

When a process is under linear drift[19], d-EWMA is a stable controller that can make the output converges to the desired process target. In this case, the process can be shown as follows,

The process disturbance can be estimated by double-EWMA controller,

\[ y_k = \beta u_k + c_k \]  \hspace{1cm} (1-6)

The process model is

\[ \hat{y}_k = bu_k + \hat{c}_k \]  \hspace{1cm} (1-7)

The intercept of the model is

\[ \hat{c}_k = \hat{c}_{k-1} + p_{k-1} \]  \hspace{1cm} (1-8)

where

\[ p_k = \omega (y_k - bu_k - \hat{c}_k) + (1 - \omega)p_{k-1} \]  \hspace{1cm} (1-9)

The stability and optimal selection of controller tuning factors of d-EWMA controller has been discussed by Good and Qin[20].

The application of double EWMA controller to MIMO processes has been carried out by Del Castillo and Rajagopal[21]. The stability conditions and the impact of estimation errors of process gain were discussed.

Kazemzadeh et al.[22] have discussed the problem of designing and investigating of EWMA/d-EWMA run-to-run controller with quadratic process model. Their research showed that the quadratic run-to-run control models outperform linear run-to-run controls in the certain manufacturing process.

Ko[23] developed a controller with two separate exponential weighted moving average (EWMA) filters simultaneously to solve the control problem of semiconductor manufacturing processes with different characteristics of the process. By utilizing dual
filters, the influence of the white noise is reduced and the accurate process control is made possible.

EWMA controller and d-EWMA controller can also be represented by an equivalent IMC structure. Standard IMC structure was modified by Adivikolanu and Zafiriou[24], and as a result, the offset subject to inherent measurement delay of run-to-run control can be eliminated. This offset was quantified as a function of modeling error, rate of drift, measurement delay, and controller tuning. A robust performance criterion and robust stability criterion in the presence of measurement delay was researched for Single-Input-Single-Output (SISO) case, and later extended to Multiple-Input-Single-Output (MISO) and over-determined Multiple-Input-Multiple-Output (MIMO)[7] cases. With the robustness criteria for the IMC structure, the effect of modeling uncertainty and measurement delays on controller performance was analyzed. The researchers also examined the tradeoff between IMC control performance, robustness and the controller’s ability to reject measurement noise.

EWMA controller was posed in a model-predictive run-to-run control framework by Campbell[25], and as a result, EWMA controller can be extended to multiple-input-multiple-output (MIMO) cases. However, the more popular algorithm used in MIMO cases are Kalman filter, which will be reviewed in the following subsection.

1.1.3 Kalman Filter and Extension

Kalman filter and its extensions are the most widely used linear filter for state estimation. When the process is expressed by a state-space model, the process disturbance can be expressed as a function of state variables. As a result, once the state variable can be estimated from data of previous runs, the run-to-run controller can be implemented.
Kalman filter was introduced to run-to-run algorithm by Palmer et al.[11] for a wafer spin coating process. In Palmer’s research, process disturbance are accounted for by adding Gaussian white noise sequences to the parameters. Standard Kalman filter equations are used to recursively update the process parameters. The advantages and disadvantages of Kalman filter were reviewed by El Chemali et al.[9], and it was pointed out that Kalman filter is optimal for MIMO cases but it is difficult to determine the optimal tuning factors of Kalman filter method. For the case with dynamic processes and wide range of process disturbances, Kalman filter works better than EWMA controller. However, the design and tuning of Kalman filter need more information about disturbance and noise statistics, which restricts the application of Kalman filter in real industry.

Kalman filter is applied to model-predictive control (MPC) theory by Mullins[26] to dynamically estimate the process states. In the case with drift disturbance, the MPC formulation works better than traditional EWMA because the drift disturbance can be explicitly included in the state-space formulation and therefore considered by the controller.

In high-mix semiconductor manufacturing processes, where different layers, products, tools may be included, Kalman filter and its extension also received a lot of attention from the researchers as a useful and powerful algorithm of non-threaded control. In non-threaded control, different manufacturing contexts may be included in one control loop, and state-estimation method is applied to estimate process parameters from measurements of previous runs.
1.1.4 Other algorithms

Other run-to-run algorithms are applied different filters, and some of them applied different process model algorithms, and there are also some methods applied different determination procedure.

The knowledge-based Interactive Run-to-Run control (KIRC)[27] algorithm is a machine learning algorithm which uses a classification decision tree to determine the control action. With this KIRC algorithm, control action is taken when the process output is out of the target range. Once an excursion from the target range is detected, leaves that match the desired target range and the current output classification will be selected. KIRC works similarly to EWMA controller in linear system, but unable to be applied in non-linear system[28].

There are also some other learning methods, such as Artificial Neural Network (ANN) algorithm, are applied to run-to-run controllers[29]. In such controllers, ANN is used as a process model method, and EWMA filter is applied to estimate the process parameters. This method was first introduced by Wang and Mahajan[30], and later Card et al.[29] tried to apply dynamic ANN to plasma etching processes. Wang and Yu[31] extended this method to the process of chemical-mechanical planarization.

Smith and Boning[32] used ANN method to compute the optimal EWMA controller of MIMO system. The effects of drift, noise, target change, and model error on the optimal EWMA weight were investigated. With ANN method, the EWMA controller tuning factor can be adjusted dynamically, and simulations showed that with this method, the controller performance of a fixed EWMA controller can be for the cases of high output noise, low rate of drift and low noise, high rate of drift. However, it is not stable when large model error exists[33].

Model-uncertainty of run-to-run controller is also an attractive issue for researchers. Model uncertainty was first addressed by Hamby et al.[34], and in Hamby’s where model parameters were treated as random variables with associated statistical
properties, and their joint probability density function was computed. Incorrect prediction of the probability of stability happened as a result of the assumption that model function may be not accurate. A performance metric based on closed-loop system behavior and the probability of closed-loop performance was calculated from Monte Carlo method.

Baras and Patel[35] tried to solve this model uncertainty problem with set-valued run-to-run controller. A set of feasible parameters to be used at the next run is computed based on past outputs and inputs. Then an optimization of a convex cost function is performed to find the least-cost inputs within the feasible output sets.

1.2 HIGH-MIX SEMICONDUCTOR MANUFACTURING PROCESSES

In semiconductor fabs, there are typically hundreds of batch unit operations such as etching, photolithography, and deposition and so on. Run-to-run control algorithm have played critical and growing role in improving the efficiency of semiconductor manufacturing processes.

The analytical capital cost in the semiconductor industry is considerably higher than other industry due to the frequent technology upgrades. Therefore, in order to maximize the revenue, manufactures try to maximize the use of their tools and minimize the down or idle time. As a result, once a product is required by customers, it is necessary to use whichever tool is available. Therefore, different products may be produced in one tool, and different tools may be used for the manufacturing of one product. The processing path through the fab of one lot of a specific product will be quite different than the next lot of that same product. This high-mix process is shown in Fig. 1-1. Products A and B may be produced in any possible tool, I, II or III, and as a result, the processing path of the 1st A and the 2nd A may be quite different.
In industry, products manufacturing are not randomly mixed but grouped with special rules. Similar products with the same processing step will be grouped together and processed on the same group of tools. For example, the first silicon nitride (Si$_3$N$_4$) layer of a CPU chip and the third Si$_3$N$_4$ layer of a memory chip can be removed by the same group of etch tools that are dedicated to remove Si$_3$N$_4$ films.

1.2.1 Threaded Control

In high-mix processes, variations of product quality are not only functions of the product being produced, but also influenced by the fabrication history of this product. The products may behave differently because of the differences of processing path. The sequence of specific tools used and the product information of one step is termed as “manufacturing context”. It may include the tool number, product number, layers, and so on. The steps or runs with the same manufacturing context are considered as “one thread”. In current run-to-run control method of industry, metrology/machine data from batches of the same “thread” are used in one control loop to control/monitor the current processes. This is called “threaded control”.

With threaded control, historical data is partitioned into different threads which are defined by manufacturing context, and the sources of variation is obviously reduced. However, the disadvantage of threaded control of high-mix processes is also quite obvious.
From the definition of manufacturing thread, it can be easily hypothesized that more contexts will lead to more threads. With the increasing of number of products, the number of threaded models will increase dramatically, and model maintenance and parameter estimation will become more difficult.

The criterion of thread definition also leads to the problem that fewer data points could be used in the parameter or state estimation of single thread. However, product behavior is also related to previous runs which may be included in other threads. As a result, the estimation performance is degraded in threaded control, and this scenario is called “Data Poverty”.

The control performance degradation of threaded control is also possible to be caused by long delay between adjacent lots in one thread, or in another word, low running-thread. In this case, some products may not been manufactured frequently, and the next lot of this thread may appear a long time later than the previous run. The long delay may result in a loss of information, and significant process drift or shift may happen during the delay time. As a result, the estimation result of threaded run-to-run control will not be reliable. It also results in a problem about control performance monitoring, because the data set for low running thread is too small to achieve reliable statistical result for control performance monitoring.

Problems caused by threaded control were characterized by Miller[36] and information sharing between different threads became more and more imperative. Several authors have proposed solutions of high-mix semiconductor processes, such as CMP[25, 37], photolithography[38] and overlay[39]. There are also a lot of research studies about non-threaded control, which will be introduced in the next subsection.

1.2.2 Non-threaded Control

In the last few years, non-threaded control methods which are based on state-estimation have drawn considerable interest from both academic and industry. The target
of non-threaded control is sharing information among different contexts. The interaction among different individual states is assumed as linear, and then a linear filter or observer is used to identify the contribution from different variation sources. For example, in high-mix etching processes, variation of products may come from devices, layers, and reticles, and a context matrix is defined to identify the information of manufacturing context. With some linear regression method, such as Kalman filter[40], variation will be attributed to different factors and updated with a linear model structure. The main difficulty of these methods is the loss of observability, because the number of metrology result is usually smaller than the number of state variables in the context matrix. Different approaches have been presented by researchers and each of them tried to make the system observable [41-43].

In non-threaded control system, the process disturbance is usually attributed to a linear sum of individual context states, and a state estimation method, such as Kalman filter [41, 42] and other linear recursive methods, is used to identify the contributions to variations from each individual context items.

The process model used for non-threaded control is usually a linear function as follows.

\[ y_k = bu_k + e_{tot,k} \]

Where \( e_{tot,k} \) is the overall process disturbance and is considered as the sum of variations from different context items.

\[ e_{tot,k} = \sum_{i}^{m} e_{i,k} \]

Here, \( e_{i,k} \) represents error comes from the context \( i \) of the \( k^{th} \) run. The total number of manufacturing contexts is \( m \). A context matrix \( H \) is used to represent which context is included in current run, and detailed representation of non-threaded control with context matrix can be found in Chapter 5.
The basic linear regression method applied in non-threaded control is Kalman filter. Early research about non-threaded control with Kalman filter used a Gauss-Markov model. The state space representation of the process is as follows.

\[ x_{k+1} = x_k \]
\[ z_k = H_k x_k + v_k \]

where \( z_k \) is the overall process disturbance, \( H_k \) is the context matrix, and \( v_k \) is white noise to represent the measurement noise. This model is easily extended to integrated white noise by adding a white noise to the state variable equation.

\[ x_{k+1} = x_k + w_k \]

This Kalman filter with integrated white noise (Kalman-IW) model has been applied to non-threaded state estimation by several researchers[44].

There are also some researchers who represented the Integrated Moving Average (IMA) disturbance in state-space model. Prabhu[45] represented IMA disturbance with state-space model, and applied Kalman filter to missing data estimation in EWMA control system of semiconductor manufacturing processes. Wang et al.[43] applied Kalman filter with IMA model (Kalman-IMA) to non-threaded control and simulation results showed that Kalman-IMA method can reduce the variation of the high-mix processes, and works better than Kalman-IW model.

Similar to Kalman filter, recursive least squares estimation (RLS) method was also used as estimator of non-threaded run-to-run control by Wang[46]. EWMA-type and RLS-type estimates are compared under measurement delay, measurement noise and deterministic drift. Ma et al. [47] developed a state estimation method based on analysis of variance (ANOVA) to estimate the relative states of each product and tool to the grand average performance of this station in the fab, and the ANOVA method is formulated in the form of a recursive state estimation using the Kalman filter.

Just in time adaptive estimation (JADE) algorithm uses recursive least squares parameter estimation to identify the contributions to the variation that are dependent upon
manufacturing context. JADE was first developed by Firth et al.[48] and had several advantages over threaded control. JADE has ability to estimate the separate context-based states at the same time. Along with the improvement in state estimation, JADE is able to indicate exactly which context item has undergone a disturbance. JADE shows less degradation in performance with delayed processes as compared to threaded control.

Several such advantages have been studied and proved through simulations [48]. JADE has several limitations listed in Firth et al.[48] and Wang et al.[46]. JADE shows improved control vs. EWMA but it depends on the assumption of a correctly specified disturbance model. JADE models the disturbance as a linear combination of context based states from each of the contributing context items. If all important context items contributing to likely disturbances to the process are not included in the JADE disturbance model, the algorithm has difficulty rejecting the unknown disturbance. If the disturbance model is nonlinear and a suitable linear form cannot be used within the operating region, then the performance degrades. Also, JADE needs qualification runs to obtain a unique solution. Wang et al.[49] demonstrated degradation in performance of JADE due to resetting the estimate covariance at each run. It also suggests increasing the weighting on the previous estimates to improve the performance of JADE when applied to stationary processes.

All these control algorithms use recursive least squares solution and have demonstrated advantages over the standard threaded EWMA. They have information sharing; and avoid data poverty and thus may be better state estimators than threaded EWMA.

1.3 CONTROL PERFORMANCE ASSESSMENT (CPA)

1.3.1 Overview of Control Performance Monitoring and Assessment

Performance degradation in control system is very common in industry due to various causes, such as inadequate controller tuning and lack of maintenance, changes in
the characteristics of the material/product being used, modifications of operating points/ranges, and changes in the status of the plant equipment (wear, plant modifications). Therefore, control performance monitoring/assessment (CPM/CPA) is imperative to maintain high efficient operation performance of automation system in manufacturing plants. The term CPA means the action of evaluating the control performance at a certain point of time, and the term CPM refers to the action of detecting or monitoring the change of automatic control performance over time. In the literature, CPA and CPM are used somewhat interchangeable.

The main objective of CPA is to provide an online procedure to determine whether current control performance are meeting specified performance targets or response characteristics. With CPA method, plant information related to automatic control is collected and statistical methods are applied to evaluate the control performance.

The field of CPA has attracted growing interest since the ground-breaking work of Harris[50] in 1989. In recent years, many research studies about CPA in both continuous and batch processes have been presented, and several excellent reviews of this field have been published.

The basic procedure for control performance assessment/monitoring and diagnosis include the following stages: (1), determine the capability of the automatic control system; (2), design or select proper benchmark for performance assessment; (3), detect and evaluate the poor control performance. Once suboptimal or degraded of control performance is detected, analysis should be carried out to diagnose the possible underlying cause and try to give suggestions to improve the control performance. The action of diagnosis and modification is the most difficult part. Many performance matrices or indices have been developed to assess the control performance for both single-input-single-output (SISO) and multiple-input-multiple-output (MIMO) cases.
1.3.2. Harris’ type Performance Indices and Its Extension;

Variation of feedback control system can occur due to two reasons: the change of process state or the degradation of control performance. When there is no major process parameter change, CPA should be carried out to verify whether the controller itself is behaving optimally under the given condition.

The first effort towards developing a performance index for monitoring feedback control systems was made by Harris[50]. In his research, the minimum variance control is considered as the best achievable performance by a feedback system. When the variance is larger than the minimum variance, the control performance is considered as “suboptimal”. The basic Harris’ type performance index is defined as follows.

\[
P_{I_{\text{Harris}}} = \frac{\text{Actual Variance}}{\text{Best Achievable Minimum Variance}}
\]

When actual variance is obviously larger than the best achievable minimum variance, the control performance is considered as suboptimal; otherwise, control performance is optimal.

The initial performance index derived by Harris is applicable only to SISO systems and involves fitting a univariate time series to process data collected under routine control, which is then compared to the performance of a minimum variance controller. Subsequent research applied the concept about minimum variance control to both feedback and feed-forward controlled univariate systems[51, 52]. The variance contributions of the inputs and different disturbance that may be present in the system are evaluated, and this can be applied to both the assessment and design of feed-forward/feedback controllers[52].

There are a lot of researches about the application of MVC to single loop. Stanfelj et al. [53] applied MVC criteria to single loop feed-forward-feedback system and developed a method to isolate whether the poor performance is due to feed-forward loop or the feedback loop by analyzing plant time series data using autocorrelation and cross-correlation functions. Lynch and Dumont [54] have used MVC estimators in control loop
performance monitoring, and influence from plant model linearity and time delay were investigated.

Other statistical methods are also applied to control performance assessment. Zhang[55] investigated the problem of control variance performance assessment with dynamic behavior constraint with covariance-assignment theory for multivariable systems. The dynamic performance of the closed-loop system was taken into account by an iterative linear matrix inequality technique (ILMI). Srinivasan[56] developed a new performance metric to assess the control performance of SISO control loop, and the proposed metric is a specific scaling of the generalized Hurst exponent, which was computed via a method of detrended fluctuation analysis (DFA). Tyler and Morari[57] introduced statistical likelihood ratio test to performance monitoring and an extension of Harris index and applied this method to unstable and non-minimum phase systems. Lynch and Dumont[58] used Laguerre networks to control performance assessment.

The Harris type performance index has been extended to MIMO processes. Huang et al[59, 60] have introduced a useful method for monitoring of MIMO processes with feedback control, which is known as Filtering and Correlation (FCOR) analysis. Time delay for a MIMO process is considered and an interactor matrix is calculated. This method is later extended to feedback and feedforward plus feedback control systems[61, 62].

In the area of control performance monitoring of MIMO systems, the interactor matrix played import roles. Harris[63] extended the MVC index to multivariable feedback processes similar to Huang[60], and non-parametric autocorrelation test is used to replace the filtering to evaluate the control performance. In Xia’s research[64], the order to the interactor matrix is determined by an input/output (I/O) delay matrix, and this method is applied to MIMO systems.

In order to simplify the calculation of multivariable performance indices, many researches are focused on eliminating the interactor matrix[65]. Ettaleb[66] firstly tried to
extend SISO CPA to MIMO system where the variables are not highly correlated. Ko and Edgar[67] uses the first few Markov parameters of the plant (up to the delay order of the process) and a set of closed-loop operating data to estimate an expression for the outputs under minimum-variance control, and a result, no prior knowledge of the interactor matrix is required.

In recent years, data-driven algorithms based on subspace approach have also received considerable attentions. McNabb[68, 69] developed a control performance monitoring method based on subspace projections. It was shown that the minimum variance output space can be considered an optimal subspace of the general closed-loop output space, and a control performance monitoring and diagnosis technique was developed based on the analysis of output covariance and generalized eigenvector.

Yu extended this data-driven covariance benchmark to multi-loop control and multivariable MPC system[70], and identified the control loops or controlled variables responsible for the performance degradation or improvement with two multivariate contribution methods[71]. One is a loading based contribution chart, and the other is to examine the angle between each individual loop/variable and the worse/better performance subspace.

Huang[72] developed a control performance method as an alternative method for CPM of MIMO control systems. The multi-step optimal prediction error variance is calculated based on subspace algorithms, and the scenarios with or without prior knowledge of time delay and model-plant mismatch were studied.

Silva and Salgado[73] computed performance bounds for MIMO systems with non-minimum phase zeros and arbitrary delay structure, and the optimal controller was obtained in Youla-parameterized form.

More literature about recent development in the field of control performance monitoring of MIMO system can be found in the review of Qin[74].
Control performance assessment and monitoring of the cases with set point change is also an attractive research topic. Ko and Edgar[75] have estimated the minimum achievable variance in a cascade control loop with set-point change in the inner loop. Seppala et al.[76] demonstrated the benefit of separating the output error into set point and disturbance, and discussed the influence of set point changes on Harris index. Thornhill et al.[77] examined factors that influence the MV performance measure of a SISO control loop and discussed the reasons why performance during set point changes differs from the regulatory performance during operation at a constant set point. Recent work about control performance assessment in the case with set point change was carried out by Wang and et al. [78]. Salsbury[79] has developed statistical procedure for the change detection of processes with random load changes. A normalized index was used and it is similar to the damping ratio in a second order process.

The influence of sampling rate and measurement delay or jitter to control performance assessment is also an attractive research topic. Horch [80] proposed a method which allows the evaluation of Harris’ performance index at sampling rates faster than the used for data collection. Yu[81] investigated how the stochastic sampling jitter affects the performance of control loops and developed a method to analyze the effect of sampling jitter when the jitter is unknown and not directly measurable.

Other statistical method was also applied to control performance assessment of non-linear system [82, 83]. Harris and Yu[84] have extended minimum variance techniques to nonlinear systems which can be identified by polynomial models. Yu et al.[82] have proposed a CPA performance index for general non-linear models based on an ANOVA-like variance decomposition method, and simulation results illustrated the efficiency and accuracy of this method. Yu et al.[83] have also proposed strategies for accurately assessing the quality of the control performance for loops with considering the influence of nonlinearities caused by valve stiction. Harris index was applied, and the bias inherent in the standard CPA calculations can be avoided by considering the valve
stiction. The effect of uncertainties and non-linearity in an IMC framework based system was quantified by Patwardhan and Shah[85], and process model, delay and disturbance model uncertainties were used to determine bounds on the performance index of the system.

The great majority of practical controllers are PID type, which is quite different from MVC. The PID controller performance is influenced by different factors, such as process and controller order, structure and action. Therefore, the calculation of the realistic performance indicators is different from Harris performance index. When restricting the controller type to PID only, the lower bound of the variance can be calculated for different disturbance models.

Ko and Edgar[86] have made second contributions in the field of control performance of PID controller. When the PI controlled process is perturbed by stochastic load disturbance, the achievable PI control performance is calculated and a minimum variance (MV) performance benchmark is used. Later, this method is extended to multivariable feedback control by Ko[67, 87] using a finite horizon MV benchmark with specified horizon length. As indicated in the CPA of MIMO systems, in this method, prior knowledge of the interactor matrix is not necessary, and only the first few Markov parameters are required. Ko and Edgar have proposed a method to calculate the best achievable PID control performance bound[88]. The controller parameters are optimized by an iterative algorithm, and a confidence interval for the performance index is derived from the optimized controller parameters. The performance assessment can be carried out for stochastic disturbance regulation processes as well as deterministic setpoint tracking. Later, this iterative algorithm is used by Prabhu[89] in the control performance assessment of EWMA run-to-run controller. The best achievable minimum variance of the EWMA run-to-run controller is calculated, and a Harris type performance index is used in Prabhu’s research.
Ma and Zhu[90] used a least-squares fit of the desired closed-loop dynamic characteristic to calculate the optimal PID setting for the control performance assessment of PID controller. Swanda and Seborg[91] have suggested a set-point response approach to monitor the performance of PI controller. Huang and Jeng[92] have developed a method to determine the performance of PI/PID controllers with an IAE index. Set point tracking is also used to obtain the step response of the system. Huang[60] has suggested a pragmatic approach toward control performance of PI/PID controllers, where an optimal LQG control law was developed to provide more realistic benchmarks for the system.

Chen et al.[93] proposed a performance assessment method for batch control system when the iterative learning control is applied. The changes from both disturbance change and set point change were considered, and the minimum variance bounds and the achievable variance bounds for each time point were calculated. Kendra and Cinar[94] have developed frequency domain techniques to CPA/CPM. The sensitivity function was applied to determine whether the control performance has degenerated and the bandwidth and peak magnitude of the sensitivity function of actual and designed systems is compared.

There are too many articles about the research in the field of CPA in the past 23 years, and only a few of them have been listed here. More comprehensive list of those methods and applications in this field can be found in the excellent reviews done by Qin[95], Harris et al. [96] and Jelali [97].

1.3.3 CPA/CPM of Semiconductor Manufacturing Processes;

As reviewed in section 1.2, most of the major processes involved in semiconductor manufacturing are batch processes, and when situ-measurements are not available, the batch recipe is usually controlled with a run-to-run controller[98]. Different filter algorithms are adopted in these semiconductor manufacturing processes, and EWMA filter is most widely used. Miller[99] et al. and Tanzer et al.[100] have expressed
that it is necessary and imperative to develop standardized benchmarks for run-to-run controllers in semiconductor manufacturing processes.

Bode et al.[101] dealt with performance assessment of run-to-run linear model predictive controllers used in semiconductor manufacturing with a minimum variance approach. A best achievable PID control performance benchmark was proposed by Ko and Edgar [88]. This was an iterative algorithm which optimized the controller parameters. Using the theoretical equivalence of EWMA controllers with discrete integral controllers, this iterative algorithm can be used for performance monitoring of run-to-run EWMA[101] controllers, commonly used in semiconductor manufacturing.

Prabhu[89, 102] derived an iterative solution method for the calculation of best achievable performance of a run-to-run EWMA controller, then a normalized performance index is defined based on the best achievable performance. The size of the moving window used during iterative analysis was optimized and the effect of model-plant mismatch, metrology delay and nonlinearity was discussed.

Ma et al.[103] have investigated the problem of EWMA controller tuning and performance evaluation in a mixed product system, and it was found that when product frequency changed, the tuning guidelines of a threaded EWMA controller were different for different types of disturbances.

Chen[104] developed the analytical expression of minimum variance performance (MVP) and the best achievable performance (BAP) of run-to-run control in semiconductor manufacturing. In his research, closed-loop identification was applied to estimate the noise dynamics via routine operating data, and numerical optimization is employed to calculate the BVP bounds of the run-to-run control loops.

Wu et al.[105] analyzed the influences of metrology delay on both the transient and asymptotic properties of the product quality for the case when a linear system with an initial bias and a stochastic autoregressive moving average (ARMA) disturbance is under an exponentially weighted moving average (EWMA) run-to-run control. A virtual
metrology technique was presented as a solution to tackle the problem of metrology delay.

Wang[106] developed a method of control performance assessment of threaded EWMA run-to-run controller bases on state estimation. And this method will be introduced in detail in Chapter 2.

Good and Qin[107]analyzed the performance of process under EWMA run-to-run control by considering time-delayed closed-loop processes with model-plant mismatch. The performance bounds were calculated based on modified stability analysis tools, and simulation results illustrated that it is important to consider metrology delay and model-plant mismatch in such EWMA run-to-run control system.

1.4 OVERVIEW OF THE DISSERTATION

This dissertation details a new algorithm developed to address the issue about control performance assessment of run-to-run control in high-mix semiconductor manufacturing processes.

The initial research of this algorithm, called Control Performance Assessment (CPA) I, is outlined in Chapter 2, and simulations are carried out to demonstrate both the advantages and disadvantages of CPA I.

In Chapter 3, a time series called “estimated disturbance” is generated from the process output error via a filter named Q filter. Then ARMAX regression is applied to the estimated disturbance and a new improved CPA algorithm named “CPA II” is developed based on CPA I.

In Chapter 4, CPA II method is applied to industrial data with threaded EWMA controller, and more discussions about the application of CPA II to real industry are carried out in this chapter.
In Chapter 5, a new non-threaded control method which is based on state estimation is introduced and CPA II method is applied to this non-threaded control system.

In Chapter 6, conclusions are drawn about the utility of CPA II method for run-to-run control problems in high-mix semiconductor manufacturing. Future extensions and other possible applications for CPA II are also explored in Chapter 6.
CHAPTER 2
CONTROL PERFORMANCE ASSESSMENT (CPA) BASED ON PARAMETER ESTIMATION

In current high-mix semiconductor manufacturing processes, various controllers are applied in different processes with different disturbance models. The EWMA controller is one of the typical controllers applied in those processes. In this chapter, the output error source of the processes with an EWMA controller will be analyzed and a CPA method will be proposed which is based on closed-loop parameter estimation. Simulations of different cases were carried out to check the feasibility and advantages/disadvantages of this method.

2.1 EWMA CONTROLLER SYSTEM

For most semiconductor processes, the deterministic trends can be adequately modeled as a static process with a constant gain, and the dynamic characteristic is contributed solely by the time-varying process disturbance. The mathematical model for a semiconductor process is given in Eq. (2-1):

\[ y_k = \alpha_k + \beta u_k \]  

(2-1)

where \( k \) is the run number, \( y_k \) is the process output, usually a product property (e.g., in an etching process, it is the film thickness); \( u_k \) is the manipulated process input (such as etching time in an etching process); \( \beta \) is a model parameter representing the process gain; and \( \alpha_k \) denotes process disturbance. The IMA (Integrative moving average) model is one of the most widely used disturbance models in semiconductor manufacturing processes. In this case, the disturbance \( \alpha_k \) can be expressed as:

\[ \alpha_k = \alpha_{k-1} + \varepsilon_k - \theta \varepsilon_{k-1} \]  

(2-2)

where \( \varepsilon_k \) is Gaussian white noise and \( \theta \) is the IMA model parameter, usually \( 0 \leq \theta \leq 1 \).
When the disturbance can be modeled by an IMA model, the EWMA (Exponential Weighting Moving Average) controller is applied to give optimal control performance. In an EWMA controlled process, the process is estimated by Eq. (2-3).

\[ a_{k+1} = (1 - \omega)a_k + \omega(y_k - bu_k) \]  

(2-3)

where \( a_{k+1} \) is the on-line estimate of the process disturbance, and \( \omega \) is the EWMA weighting.

The manipulated process input \( (u_k) \) can be determined by a deadbeat control law as follows:

\[ u_k = \frac{r_k - a_k}{b} \]  

(2-4)

where \( r_k \) is the set point of the process and \( b \) is the off-line estimate of the process gain \( \beta \).

From Equations (2-2) to (2-4), we can see that for semiconductor processes, the control performance is determined by the accuracy of the off-line estimate \( b \) and the on-line estimate \( a_{k+1} \). As a result of the process changes caused by equipment wearing, part replacement and other maintenance events, the process gain changes over time. Therefore, even if \( b \) was estimated accurately initially, the real process gain may become quite different from the estimated \( b \) after a while. If an optimal tuning of the EWMA weighting is made at the beginning of a series of process steps, it will be suboptimal once the parameters change.

2.2 ERROR SOURCE ANALYSIS

The EWMA controller is equivalent to an internal model control (IMC) structure as shown in Fig.2-1.

In this structure, \( G_c(q^{-1}) \) denotes the controller used in this RtR control system, and \( G_p(q^{-1}) \) and \( \hat{G}_p(q^{-1}) \) denote the plant and its nominal model used by the controller. \( q^{-1} \) is the backward shift operator. \( G_E(q^{-1}) \) shows the controller transfer function, such as using EWMA control theory. The time delay or metrology delay is also shown in this figure, and it is represented by \( D = q^{-d} \), where \( d \) is the time delay. For simplicity, the
argument \((q^{-1})\) is omitted in the following discussion. In the process using the EWMA controller as described in section 2.1,

\[
\begin{align*}
G_p &= \beta \\ 
\hat{G}_p &= b \\ 
\alpha_k &= \frac{1}{1 - \theta q^{-1}} \varepsilon_k \\ 
G_c &= \frac{1}{b} \\ 
G_E &= \frac{\omega q^{-1}}{1 - (1 - \omega)q^{-1}}
\end{align*}
\]

Using these equations, the contributions from model-plant mismatch and controller tuning to the overall control performance can be explicitly accounted for.

In the control block diagram shown in Fig. 2-1, the following closed-loop relationships for the output error \((e_k)\) and the prediction error \((\hat{y}_k)\) can be derived as follows.

\[
e_k = r_k D - y_k = \frac{1 + G_p G_c (DG_E - 1) - \hat{G}_p G_c G_E}{1 + (G_p D - \hat{G}_p)G_c G_E} Dr_k - \frac{1 - \hat{G}_p G_c G_E}{1 + (G_p D - \hat{G}_p)G_c G_E} D\alpha_k
\]

\[
\hat{y}_k = y_k - \hat{y}_k = (G_p D - \hat{G}_p)G_c \hat{e}_k + D\alpha_k
\]

\[
\hat{e}_k = r_k - G_E \hat{y}_k
\]

Here \(\hat{e}_k\) is the input to the controller. As all of the signals, \(e_k\), \(r_k\), \(\hat{y}_k\) and \(\hat{e}_k\) can be obtained from the operating data, then it is possible to estimate both the model-plant mismatch \((G_p D - \hat{G}_p)\) and the disturbance \((\alpha_k)\).
For a semiconductor manufacturing processes with EWMA controller, as described in section 2.1, applying Eqs. (2-5 to 2-9) to Eq. (2-10), then the output error can be calculated as:

\[
e_k = \left\{ \frac{\beta}{b} \left( q^{-d} - q^{-d-1} \right) + \left( 1 - q^{-1} \right) \right\} r_k + \left\{ \frac{1 - \theta q^{-1}}{1 - q^{-1} + \omega \frac{\beta}{b} q^{-1-d}} \right\} \varepsilon_{k-d} \tag{2-13}
\]

As shown in Eq. (2-13), the output error \( e_k \) is comprised of two parts. One part is related to the set point \( r_k \), and this part can be named as “deterministic error”; the second part is a function of white noise \( \varepsilon_k \), and this part can be called “stochastic error”.

When the process set point \( r_k \) is constant, the output error is

\[
e_k = \left\{ \frac{1 - \theta q^{-1}}{1 - q^{-1} + \omega \frac{\beta}{b} q^{-1-d}} \right\} \varepsilon_{k-d} \tag{2-14}
\]

In this case, the output error \( e_k \) is the stochastic error.

In most semiconductor manufacturing processes, the process set point is kept constant in the control loop. In the following discussion, the case with varied set points will also be discussed, but in fact, when \( r_k \) is not constant, the conventional EWMA controller may be not a good choice because the difference between each \( r_k \) will lead to serious disturbance that cannot be compensated by tuning the controller parameter.

### 2.3 Optimal EWMA Controller

Box and Kramer[108] have shown that for the case without model-plant mismatch, the optimal EWMA tuning for an IMA disturbance can be derived. For an EWMA controller, the mean squared error of the forecast is minimized if the disturbance is an integrated moving average (IMA) time series of the form:

\[
\alpha_k = \alpha_{k-1} + \varepsilon_k - \theta \varepsilon_{k-1} \tag{2-2}
\]

Here \( \varepsilon_k \) is a white noise sequence. The relationship between the optimal EWMA controller weighting and the IMA model is:

\[
\omega_{opt, \omega} = 1 - \theta \tag{2-15}
\]
The best tuning result of the weighting parameter $\omega$ can also be adjusted when the model gain is different from the real process gain. However, in reality it is seldom to have a control model that fits the plant perfectly. Wang et al [106, 109] have derived the optimal control tuning for IMA disturbance with model-plant mismatch.

\[
\omega_{opt,\omega} = (1 - \theta) \frac{b}{\hat{\beta}}
\]  

(2-16)

For the case with time delay, the analysis of the optimal control parameters can be found from the optimal $d$-step-ahead prediction. The optimal control tuning for IMA disturbance with model plant mismatch and time delay can be derived as follows. Details about derivation can be found in Appendix A.

\[
\omega_{opt} = \frac{1 - \theta}{1 + (1 - \theta)d \hat{\beta}} \frac{b}{\hat{\beta}}
\]  

(2-17)

2.4 CPA (I) Based on Closed-Loop Parameter Estimation

Various forms of control performance indices have been defined and applied to different processes. Most of the existing indices are defined as the ratio of the achieved control performance compared to a benchmark performance, such as the minimum variance control performance (the theoretical lower bound) or the best achievable control performance (the statistical lower bound).

As discussed before, in high-mix processes, the high cost of process equipment drives manufacturers to maximize the use of their tools, having as little down or idle time as possible. Therefore, different products with different targets may be produced with the same tool. Suppose for a specific tool, the process gain $\beta$ does not change much even though the targets are quite different. Therefore, one natural choice of CPA/CPD for high-mix process is to estimate the process gain (the estimated process gain is $\hat{\beta}$) and compare the optimal weighting to the actual weighting parameters. If the EWMA controller is well-tuned initially, the value of $|b/\hat{\beta} - 1|$ could mean the controller is close to its optimal performance. Otherwise, the control performance is suboptimal and
can be improved by either proper tuning (change the EWMA weighting $\omega$) or control model updating (change the estimated gain $b$ used in the controller)\[106].

From Eq. (2-13), parameters $\beta$ and $\theta$ can be estimated from the output error $e_k$ by applying ARMAX regression.

A general ARMAX model is:

$$A(q^{-1})e(t) = B(q^{-1})r(t) + C(q^{-1})\varepsilon(t)$$

When the set point is constant or the influence from set point change can be eliminated, the output error $e_k$ can be expressed by Eq. (2-14), then $B = 0$ in Eq. (2-18), and as a result, the ARMAX model should be:

$$A(q^{-1})e(t) = C(q^{-1})\varepsilon(t)$$

Details about ARMAX model and ARMAX regression are shown in Appendix B.

However, a closer examination shows that the value of $|b/\hat{\beta} - 1|$ may not be reliable indices in several cases. The first unreliable case is that when the closed-loop performance is close to its optimum. When the process set point $r_k$ is constant and $\omega$ is equal or close to $\omega_{opt}$, the closed-loop system defined by Eq. (2-13) will reduce to a white noise sequence and lose its observability due to the pole-zero cancellation, leading a biased estimation. As a result, the variance of $\hat{\beta}$ (Var($\hat{\beta}$)) will be very large. On the other hand, when the controller performance is suboptimal, the estimation is not biased and Var($\hat{\beta}$) will not be so large. However, if the variances are small, it may indicate that the estimation is accurate and the estimated parameters can be used in CPD. Based on this fact, an intuitive performance index can be proposed as follows.

$$PI_1 = \frac{1}{\sqrt{\text{Var}(\hat{\beta}/b)}}$$

Control performance assessment method based on Eq. (2-20) is called “CPA (I)” in this dissertation. In addition, the traditional performance index can also be calculated as follows.
\[ P_{I_2} = \frac{\text{var}(y)}{\text{var}(\epsilon)} \]  

(2-21)

\( \text{var}(\epsilon) \) can be obtained from the closed-loop identification or other methods. In following discussion, simulations of EWMA controller for different cases are made to check the advantages and disadvantages of CPA (I).

2.5 CPA (I) IN THREADED EWMA CONTROLLER SYSTEM WITH CONSTANT SET POINT

In this section, the window approach is applied for on-line performance monitoring. The process set point and process gain are kept constant in one control loop to simulate threaded control processes.

To monitor the controller performance on-line, a moving window approach is implemented to perform the closed-loop identification. The details of the on-line algorithm are given below, where \( n \) denotes the window size.

1) At run \( k \), set up data matrix for the closed-loop identification, use the data from the run \( k-n+1 \) to the run \( k \).

2) Perform system identification to obtain estimates of \( \beta_k, \theta_k \).

3) Once \( m \) new measurements become available, \( k=k+m \) and return to step 1.

In the simulation, we choose the window size \( n=200 \), and update the estimation every 20 runs \( (m=20) \). Parameters used in the simulation are shown in Table 2-1. \( \omega=\omega_{opt}=0.375 \), and metrology delay \( d = 0 \).

<table>
<thead>
<tr>
<th>( \beta )</th>
<th>( \theta )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>0.7</td>
<td>1</td>
</tr>
<tr>
<td>5.0</td>
<td>6</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Table 2-1 Parameters of Simulation 2-1

In simulation 2-1, a total of 6000 runs are simulated, and a ramp disturbance was introduced to the process gain \( \beta \) at run 2000. The ramp has duration of 200, and a slope of -0.01. Then the true value of \( \beta \) from run 2200 to run 6000 is 2.0. To simulate equipment maintenance events, after every 50 runs, the disturbance \( \alpha \) is reset to its initial value, and
the corresponding $\varepsilon$ reset to zero to start a new IMA sequence with the same parameter $\theta$. The simulated process data and the output variance are shown in Fig. 2-2. It is difficult to determine whether the control performance is optimal or not from $\text{Var}(y)$, because it did not show obvious trends here.

![Simulated process data and output variance in Simulation 2-1](image)

Figure 2-2 Simulated process data and output variance in Simulation 2-1

The estimated value of $\beta$ is shown in Fig. 2-3. Before the ramp disturbance occurs in $\beta$ ($i = 2000$), the controller performance is optimal, and as shown in Fig. 2-3, the estimated value of the process gain before the ramp clearly shifted and $\text{var}(\hat{\beta})$ is quite large. For the original weighting, $\omega$ is optimal for $\beta=4.0$; after the ramp disturbance, $\beta=2.0$, and the weighting $\omega=0.375$ is not the optimal choice now. As we can see from Fig.2-3, after the ramp disturbance, variance of the estimated value of $\beta$ ($\hat{\beta}$) is much smaller.

This simulation shows that when the true process gain is kept constant, a large variance of the estimated $\beta$ denotes the control performance is near the optimum, while a small variance of estimated $\beta$ represents suboptimal control.
Figure 2-3. Estimated gain in simulation 2-1

The performance index of this simulation is shown in Fig. 2-4

Figure 2-4. On-line monitoring of Simulation 2-1

From Fig. 2-4, we find that when $PI_1$ is near 1, it means the process is near the optimal control, but when it is much larger than 1, it may indicate a suboptimal
performance. The result of simulation 2-1 indicates the feasibility of CPA method (I) for threaded Run-to-run EWMA controller when the process disturbance can be expressed as IMA model.

2.6 CPA (I) in NON-THREADED EWMA CONTROLLER SYSTEM

   In this section, high-mix case with multiple-targets and multiple-process gains in one control loop is considered. All simulations were made in MATLAB. For all simulations, the total number of runs is \( N=6000 \). The moving window size for system identification and parameter estimation is \( wn=200 \). The estimation is updated every 20 runs. For all simulations, the target \( r_k \) was randomly selected from the collection \( R=\{-50, -40, -30, -20, -10, 0, 10, 20, 30, 40, 50\} \) to reflect changes in the product being produced on the tool. White noise \( \epsilon_k \) was generated by normally distributed pseudorandom numbers, with the initial seed=0. For every 50 runs, the value of white noise is reset to the initial value, which indicates a process maintenance step. For all simulations in this chapter, the metrology delay \( d = 0 \).

2.6.1 Varied Set points with Plant Model Mismatch

   Parameters used in Simulation 2-2 are shown in Table 2-2.

   Table 2-2. Process parameters used in Simulation 2-2

   \[
   \begin{array}{ccc}
   \alpha_0=0.6 & \theta=0.7 & \beta \text{ varies} \\
   \alpha_0=0.55 & \omega=0.3 & b=5 \\
   \end{array}
   \]

   In simulation 2-2, the influence of \( \beta \) on the output error and CPA results was studied. From \( i=1 \) to \( i=2000 \), \( \beta=b=5 \). As indicated by Eq. (2-16), the control performance in this region was “optimal”. From \( i=2001 \) to \( i=2200 \), a ramp disturbance in the actual process gain with \( slope=-0.005 \) was introduced to \( \beta \). As a result, when \( i=2200 \), \( \beta=4 \). From \( i=2201 \) to \( i=6000 \), \( \beta \) was fixed at a value of 4, and control performance in this
region was “suboptimal”. The ARMAX model shown in Eq. (2-18) was applied. The output error ($e_k = y_k - r_k$) and the variance of $e(t)$ for this simulation is shown in Fig. 2-5.

![Graphs showing process errors and variance](image)

Figure 2-5. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-2

When $\beta=b$, the variance of output error is relatively small, and when $\beta \neq b$, the variance of the output error was obviously large. It is because in this case, when $\beta \neq b$, the deterministic error related to set point change contributes a large part to the process output error; when there is no model-plant mismatch, and no time delay (as simulation 2-2), from Eq. (2-13), it can be found that the process output error only includes the stochastic error, which is relatively small.

Three parameters can be estimated by the ARMAX model shown in Eq. (2-18). They are $\beta/b -1$ (estimated from matrix $B$), $1-\omega^*(\beta/b)$ (estimated from matrix $A$) and $\theta$ (estimated from matrix $C$). Note that the value of $\beta/b$ can be estimated from both $A$ and $B$, but only the value estimated from $A$ is the one applied to the performance index of CPA (I).

The estimated ($\beta/b -1$) from $A$ is shown in Fig. 2-6. When the set point varies from run to run, the ARMAX regression is highly excited and precise estimation of
matrix A in ARMAX model can be expected. As shown in Fig. 2-6, for both “optimal” and “suboptimal” cases, we can always obtain precise estimation of $(\beta/b -1)$. In other words, once an obvious change in the estimated $(\beta/b -1)$ was observed, it indicates that the nominal process gain $b$ used in the model is not equal to the real process gain $\beta$, and some adjustments to the process model should be made.

![Figure 2-6. Estimated $(\beta/b -1)$ in Simulation 2-2](image)

From the system identification, the value of $1-\omega*(\beta/b)$ and $\theta$ can also be estimated from polynomial $B$ and $C$ (which are represented by matrix in MATLAB) of ARMAX model, as shown in Fig. 2-7. For the optimal case with $\beta = b$ and $\theta = 1 - \omega(\beta/b)$, precise estimation of $1-\omega*(\beta/b)$ cannot be obtained, because $e(t)$ is white noise in this case. From Fig. 2-7, we can find that in the “optimal” control region (Run Number $i = 1$ to 1000), the oscillations of both the estimated $1-\omega*(\beta/b)$ and $\theta$ were very severe. When the control performance is not optimal ($i = 2000$ to 6000), the oscillations of the estimated value of both $1-\omega*(\beta/b)$ was not so apparent, and a relatively precise estimation of the
two parameters can be obtained. The value of the estimated $\theta / [1 - \omega^* (\beta / b)]$ was also calculated in Fig. 2-8, but in this case, it did not supply any useful information.

Figure 2-7. Estimated $1 - \omega^* (\beta / b)$ and $PI_1$ in Simulation 2-2

Figure 2-8. Estimated $\theta / [1 - \omega^* (\beta / b)]$ in Simulation 2-2

From these figures, it can be found that:
a) When the process set points varied from run to run, for both optimal and suboptimal control performance, precise estimation of $\beta/b$ can be achieved from ARMAX regression.

b) When there is no model plant mismatch ($\beta=b$), the output error is small, and the main error comes from stochastic disturbance; when there is model plant mismatch ($\beta\neq b$), the output error is very large, because deterministic error comes from set point change and cannot be eliminated by tuning the controller.

c) For the optimal case with $\beta=b$ and $\theta = 1 - \omega(\beta/b)$, the output error is close to white noise and as a result, the estimation variance of $1-\omega*(\beta/b)$ or $\theta$ is very large. For the suboptimal case, estimation of the two parameters with much smaller variance can be achieved.

In this simulation (2-2), when the control performance was “suboptimal”, neither $\beta=b$ nor $\theta = 1 - \omega(\beta/b)$ were satisfied. Although from ARMAX regression result, the value of $\beta/b$ can be identified in this case, it still not clear whether the suboptimal performance is only caused by this model plant mismatch. In other words, the deterministic error related to the set point change can be diagnosed by ARMAX regression of the value $\beta/b$, but stochastic error related to process disturbance is difficult to identify in this case. Further discussion about the diagnosis of stochastic error will be shown in chapter 3.

### 2.6.2 Varied Set points Without Plant Model Mismatch

As indicated in Simulation 2-2, in high mix RtR EWMA controlled processes where the set point varies from run to run, it is possible to identify the model-plant mismatch with ARMAX analysis of the process output error. Once model-plant mismatch is found, this mismatch can be fixed by changing the nominal process model. Then in the following simulation, it is assumed that model-plant mismatch dose not exist and $\beta=b$ for all runs.
Parameters used in Simulation 2-3 are shown in Table 2-3. In this simulation, a ramp disturbance was introduced to $\theta$ to simulate suboptimal behavior results from improper controller tuning. At the beginning, $\theta=1-\omega=0.7$, which indicated an “optimal” case. From $i=2001$ to $i=2200$, a ramp disturbance with slope $=-0.002$ was introduced to $\theta$. From $i=2201$ to the end, $\theta=0.3$, and the control performance was not optimal.

Table 2-3 Process parameters used in Simulation 2-3

<table>
<thead>
<tr>
<th>$a_0$</th>
<th>$\theta$ varies</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>$\theta$ varies</td>
<td>5</td>
</tr>
<tr>
<td>0.55</td>
<td>$\omega=0.3$</td>
<td>$b=5$</td>
</tr>
</tbody>
</table>

Note that for both cases, $\beta=b$, and from Eq. (2-13), the output error signal $e(t)$ was only a function of white noise. So in this case, ARMAX model expressed by Eq. (2-19) should be used, and only the values of $1-\omega*(\beta/b)$ and $\theta$ can be estimated.

The output error ($e_k=y_k-r_k$) and the variance of $e(t)$ in Simulation 2-3 are shown in Fig.2-9. Because $\beta=b$ for all runs, although $\theta$ changed considerably, the output error $e(t)$ did not change so much (Fig.2-9 (a)). The variance of $e(t)$ of the suboptimal case ($i=2000$ to 6000) is slightly larger than the optimal case ($i=1$ to 1000), but the difference is not very obvious (see Fig. 2-9(b)).
Figure 2-9. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-3

The values of estimated $1-\omega*(\beta/b)$ and $\theta$ estimated in Simulation 2-3 were shown in Fig. 2-10. Similar to the result of Simulation 2-2, when control performance was optimal, the estimated values of both parameters varied a lot; and when control performance was not optimal, a relatively precise estimation can be achieved.

Figure 2-10. Estimated $1-\omega*(\beta/b)$ and $P_l$ in Simulation 2-3
We also calculated the ratio of these two parameters, and the value of $\theta / [1 - \omega^*(\beta/b)]$ is shown in Fig.2-11. From Fig.2-11, it can be easily recognized that for the suboptimal case with $\theta$ changed, the value of $\theta / [1 - \omega^*(\beta/b)]$ clearly deviated from 1. Based on this result, we know that in this suboptimal case, $\theta < 1 - \omega^*(\beta/b)$. In order to return to the optimal control, we need to increase the value of $\omega$. An approximate ratio of $\theta / [1 - \omega^*(\beta/b)]$ can also be obtained from Fig.2-11, which can be helpful to determine the optimal $\omega$.

Conclusions for Simulation 2-3:

a) When $\beta = b$, the change of $\theta$ can influence the output error variance, but not in an obvious way.

b) For the optimal case with $\beta = b$ and $\theta = 1 - \omega(\beta/b)$, we cannot obtain precise estimation of $1 - \omega^*(\beta/b)$ or $\theta$ by the ARMAX model (Eq. (2-14)), and the variability of the estimated results was fairly large. For the suboptimal case with $\beta = b$ but
\[
\theta \neq 1 - \omega (\beta/b),
\]
we can get estimation results of the two parameters with much smaller variations by the ARMAX model in Eq. (2-19).

The most important change in Simulation 2-3 vs. Simulation 2-2 is that we used a different ARMAX model. In Simulation 2-3, there is no model plant mismatch, so the output error is only the stochastic error, and as a result, the polynomial \( B \) in the ARMAX model should be 0 (Eq. 2-19); in Simulation 2-2, \( B \) is not 0 (Eq. 2-18). The influence of improper selection of regression model will be discussed in simulation 2-4.

2.6.3. Varied Set Points with Improper Tuning

The parameters used in Simulation 2-4 are shown in Table 2-4. The change of \( \beta \) is the same as that in Simulation 2-2. The only difference between Simulation 2-4 and Simulation 2-2 is the value of \( \omega \). Here, \( \omega = 0.375 = (1 - \theta) \times (5/4) \). As a result, in the first region \( (i = 1 \text{ to } 2000) \), \( \beta = b \) but \( \theta \neq 1 - \omega (\beta/b) \), and the main output error is “stochastic error”; in the second region \( (i = 2001 \text{ to } 6000) \), \( \beta \neq b \) but \( \theta = 1 - \omega (\beta/b) \), and the main output error is “deterministic error”. Neither case is really “optimal”. As \( \beta \neq b \) in the second region, \( B \) in the ARMAX model is not 0.

<table>
<thead>
<tr>
<th>Table 2-4. Process parameters used in Simulation 2-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_0 = 0.6 )</td>
</tr>
<tr>
<td>( \alpha_0 = 0.55 )</td>
</tr>
</tbody>
</table>

The output error \( e(t) \) and the variance of \( e(t) \) in Simulation 2-4 are shown in Fig. 2-12. This result is similar to Simulation 2-2 in Fig. 2-5. The performance of the estimated \( 1 - \omega (\beta/b) \) and \( \theta \) (Fig. 11) are also the same as that in Simulation 2-2 (Fig. 2-5). Remember that from \( i = 1 \) to 2000, \( \theta \neq 1 - \omega (\beta/b) \). So in fact the control performance in this region was not “optimal” due to a suboptimal value of \( \omega \). But the significant variation of the estimated \( 1 - \omega (\beta/b) \) in Fig. 2-13 seems to indicate an “optimal” control
performance. The estimated $\beta/b$ and $\theta / [1-\omega*(\beta/b)]$ in Simulation 2-4 are shown in Fig. 2-14.

Comparing results of simulation 2-2 and 2-4, it can be found that when an imprecise regression model is applied, the variance of estimated $1-\omega*(\beta/b)$ is not sensitive to the change of the actual $\theta$. Suboptimal behavior caused from stochastic error may be buried by the deterministic error, and incorrect assessment may be concluded from the method CPA (I).

Figure 2-12. High-mix process errors $e(t)$ and variance of $e(t)$ in Simulation 2-4

Figure 2-13. Estimated $1-\omega*(\beta/b)$ (a) and $PL_1$ (b) in Simulation 2-4
2.6.4 The Selection of ARMAX Regression Model

The process in Simulation 2-5 is the same as that in Simulation 2-2; and the process in Simulation 2-6 is the same as that in Simulation 2-3. However, a different ARMAX model was applied. In Simulation 2-5 ($\beta/b$ changed from 1 to 0.8), Eq. (2-19) was used, and in Simulation 2-6 ($\beta/b=1$ but $\theta$ changes from 0.7 to 0.3), Eq. (2-18) was applied.

The results of Simulation 2-5 are shown in Fig.2-15. Even for the “suboptimal” case (i =2001 to 6000), the variances of the estimated $1-\omega*(\beta/b)$ is still somewhat large due to the wrong model. Also the value of $\theta / [1-\omega*(\beta/b)]$ did not provide any useful information.

Figure 2-14. Estimated ($\beta/b-1$)(11-a) and $\theta / [1-\omega*(\beta/b)]$ (11-b) in Simulation 2-4
The results of Simulation 2-5 are shown in Figs. 2-16 and 2-17, where the same problem happens. In Fig. 2-16 large variations of estimated $1-\omega*(\beta/b)$ can still be observed for the “suboptimal” region ($i = 2001$ to 6000) with $\theta \neq 1-\omega(\beta/b)$. Then it is difficult to determine whether the control performance is optimal or not. The value of the estimated $\beta/b$ in Fig.2-17 shows that no change occurs in $\beta/b$, but the estimated $\theta / [1-\omega*(\beta/b)]$ varied so much that it was not very reliable to determine how to adjust the value of $\omega$. 

Figure 2-15. Estimated $1-\omega*(\beta/b)$ (a), $PL_1(b)$, and $\theta / [1-\omega*(\beta/b)]$ (c) in Simulation 2-5
Figure 2-16. Estimated $1-\omega*(\beta/b)$ (a) and PI$_1$ (b) in Simulation 2-6

Figure 2-17 Estimated $\beta/b$-1 (a) and $\theta / [1-\omega*(\beta/b)]$ (b) in Simulation 2-6

Conclusions for Simulation 2-5 and 2-6:

a) The estimated result from the ARMAX model expressed by Eq. (2-18) ($B(q)$ not assumed to be 0) is sensitive to the change of $\beta$, but not so sensitive to the change of $\theta$. It can always provide precise estimation of $\beta/b$, so it can be used to determine whether the
nominal process gain \( b \) is optimal or not. If \( \beta \neq b \), this estimation of \( \beta/b \) can also be used to adjust the value of \( b \).

b) The estimated results from the ARMAX model expressed by Eq. (2-19) \((B(q) \text { is assumed to be 0})\) are sensitive to the change of \( \theta \), but not so reliable when \( \beta \) changes. When \( \beta = b \), the value of \( \theta / [1-\omega*(\beta/b)] \) provided by this model is very useful for controller tuning.

In order to achieve a valid CPA result in process with varied set points, ARMAX regression with \( B \) should be applied first, to determine whether there is model plant mismatch or not. Once it is determined that there’s no model plant mismatch, ARMAX model with \( B=0 \) should be applied to investigate the influence of controller tuning factor.

2.6.5 Disturbance Model Different from IMA

Simulations 2-1 to 2-6 showed the feasibility of CPA method (I) in cases with single/varied set points, with/without model plant mismatch, proper/improper controller tuning. However, all those simulations are based on the assumption that the process disturbance can be described by an IMA model. In real industry, this assumption does not always work. Simulation (2-7) is carried out to study the influence of an improper disturbance model.

All parameters used in Simulation 2-7 are the same as Simulation 2-1, and the only difference is a periodical disturbance is introduced to the process gain \( \beta \), which may be possible in some industrial processes. The estimated \( \beta \) is shown in Fig.2-18. The figure shows obvious oscillation, which may be considered to be an indication of optimal control performance. As a result, variance from oscillating process parameters is often considered as a symbol of “optimal control”, which in fact is a wrong conclusion.
2.7 CONCLUSION AND DISCUSSION

In this chapter, the error source of the Run-to-run EWMA control system is analyzed, then a Control Performance Assessment method (CPA I) is generated based on the error source analysis. In CPA I, ARMAX regression is directly applied to the process output error, and the performance index is defined based on the variance of the regression results.

Simulations of both threaded (single target, single process gain) and non-threaded (multiple set points) are carried out and the results demonstrated that ARMAX regression of process output can provide much information about control performance assessment. In the cases with multiple set points, it is also possible to diagnose the model-plant mismatch and controller tuning problem.

However, simulations also revealed the disadvantage of CPA I. It does not work for some cases such as varied process gains which cannot be described by IMA model. Metrology delay is also not considered in those simulations, while it is really important in
real high-mix semiconductor manufacturing processes. Discussion about cases with metrology delay will be covered in Chapter 3.

Furthermore, in CPA I, in order to calculate the performance index, variance of the ARMAX regression results of different windows should be calculated. As a result, data from many runs is required to calculate one variance result, which is not acceptable in real industry.

The study of CPA method (I) indicated that it is possible to investigate the control performance from ARMAX regression of process output error. It is not only possible to determine whether the control performance is good or not, but also possible to offer diagnosis information of suboptimal behavior. However, CPA method (I) is too sensitive to the disturbance model, and the problem of varied process parameters and metrology delay is solved in Chapter 3, where a new method (CPA II) will be developed.
In Chapter 2, simulations have shown that when the set point changed from run to run, which means the reference signal of this closed-loop control system is not zero, it is possible to estimate the ratio of real process gain to the nominal process gain ($\hat{\beta}/b$).

Once the value of $\hat{\beta}/b = 1$ is not zero, it requires special attention because the plant model mismatch will cause severe suboptimal control behavior in this case.

Simulation also can show that when there is no reference signal, which means the set points always remains constant, and is difficult to estimate the plant model mismatch. In addition, it is difficult to estimate the process gain from closed loop identification when the set point does not change. In current semiconductor manufacturing processes, the set point is always constant in one control loop, and in this case, it is difficult to determine whether there is plant model mismatch or not. So how can we investigate the control performance with a constant set point?

In this chapter, the closed loop case without reference signal (set point remains constant) will be discussed and a new CPA method is developed called CPA II. With the new CPA II method, the cases with plant model mismatch, improper controller tuning, metrology delay and high-mix process parameters are studied and the results showed that the new CPA II method can quickly identify control performance degradation and it is also possible to diagnose the suboptimal behavior and correct it.

3.1 MATHEMATICAL FOUNDATION OF CPA II

3.1.1 Error Source Analysis

In Fig. 2.1, the IMC structure of the run-to-run EWMA controller has been shown as a proportional controller with EWMA filter, and the process model is represented by a transfer function. If we represent the process with a state space frame[110], as follows:
\[
x_{k+1} = Ax_k + B_k
\]  
(3-1)

with

\[
A = \begin{bmatrix}
1 & 0 \\
\omega & 1 - \omega
\end{bmatrix}
\]  
(3-2)

and the process output can be expressed as:

\[
y_k = Cx_k + v_k
\]  
(3-3)

The state variable is:

\[
x_k = \begin{bmatrix}
\alpha_k \\
\alpha_k
\end{bmatrix}
\]  
(3-4)

and

\[
B_k = \begin{bmatrix}
\Delta\alpha_{k|k-1} \\
0
\end{bmatrix}
\]  
(3-5)

More detailed description of the state-space representation of this system can be found in Prabhu’s work[110].

It can be found when the process disturbance \(\alpha_k\) is considered as the state variable, this state variable cannot be measured directly, so an observer should be applied in the controller system. In this system, the EWMA filter is used as an observer, and then the online estimate \(\alpha_k\) can be considered as the observed value of the state variable.

In order to minimize the process output variance, an optimal controller can be designed based on this model according to minimum variance control theory[111].

The process can be represented as:

\[
A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})\varepsilon(t)
\]  
(3-6)

where \(\varepsilon(t)\) is zero-mean white noise with variance \(\sigma_e^2\). In the process shown in Fig. 3.1, we can have \(A = 1 - q^{-1}, B = \beta(1 - q^{-1}), \) and \(C = 1 - \theta q^{-1}\).

Then a minimum variance controller can be designed as:

\[
u = -\frac{G}{BF}y
\]  
(3-7)

Where \(G\) and \(F\) must satisfy the equation

\[
C = AF + q^{-d}G
\]  
(3-8)
Once $A, C$ and $d$ are determined, $F$ and $G$ can be calculated. In the case without a reference signal ($r_k = 0$), the EWMA controller can be represented by two equations:

(P controller) 
$$u = -\frac{1}{b}a$$ (3-9)

(EWMA observer) 
$$a = \frac{\omega q^{-1}}{1 - q^{-1}}y$$ (3-10)

That means

$$u = -\frac{\omega q^{-1}}{b(1 - q^{-1})}y$$ (3-11)

Then the optimal controller is

$$-\frac{\omega q^{-1}}{b(1 - q^{-1})} = -\frac{G}{\beta(1 - q^{-1})F}$$ (3-12)

So

$$\omega = \frac{bG}{\beta F}$$ (3-13)

As a result, in the case without a reference signal (set point $r_k$ remains constant), the influence of plant model mismatch to the control performance can be compensated by selecting a proper value of $\omega$. This conclusion is also supported by the error source analysis. As shown in Eq. (2-14), when the value of $\beta/b$ changes, the output error $e_k$ can be kept constant by changing the value of $\omega$. As a result, it is not necessary to know the exact model plant mismatch if we only want to assess and diagnose the control performance.

If one assumes there is no model plant mismatch ($\beta = b$), the optimal controller indicates the optimal observer (predictor) $a_k[k|k-d]$, which can minimize the variance of observer error $\alpha_k - a_k[k|k-d]$. In another words, in order to determine whether the control performance is optimal or not, we can check whether the predictor is optimal or not.

3.1.2 Design of Q Filter

When the set point stays constant, the output error $e_k$ can be expressed as a function of the disturbance $\alpha_k$. When $\alpha_k$ is a white noise driven time series, such as IMA(1,1) series, the output error can also be expressed as a function of white noise.
Table 1 shows expressions of output error of different cases with set point staying constant. Substitute the expression of output error to the expression of $\hat{e}_k$, then $\hat{e}_k$ can also be expressed as a function of white noise $\varepsilon_k$.

Because $\theta$ and $\beta$ are unknown, they should be represented in terms of known variables such as $\omega$ and $d$. Here $\delta$ is a factor that varies from 0.1 to 10, and it has no special physical meaning.

When $\delta=1$, the estimated noise is just the noise that was not predicted by the controller. If this noise is white noise, then the controller is optimal; otherwise, the controller performance is not optimal. Either ARMAX regression or Maximum Likelihood method can be applied to check whether it is white noise or not. Here, we use the ARMAX regression in the form of $\hat{e}_k = (1 + C(2)q^{-1})\varepsilon_k$. When $-C(2)$ is near 0, $\hat{e}_k$ is white noise and the control performance is optimal. When $-C(2)$ is obviously different from 0, it may indicate suboptimal behavior.

Discussion about the cases with $\delta\neq1$ will be shown in Section 4. It is related to the calculation of best or better achievable performance variance and suggestion of controller parameter tuning. Additional information about Table 3-1 can be found below.

Table 3-1 Q filter and estimated disturbance of different cases

<table>
<thead>
<tr>
<th>$\beta = b$, $d = 0$</th>
<th>$e_k$</th>
<th>$\omega$</th>
<th>$\hat{e}_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta \neq b$, $d = 0$</td>
<td>$1 - \theta q^{-1}$</td>
<td>$\omega$</td>
<td>$1 - (1 - \hat{\omega})q^{-1}e_k$</td>
</tr>
<tr>
<td>$\beta = b$, $d \neq 0$</td>
<td>$1 - \theta q^{-1}$</td>
<td>$\omega$</td>
<td>$1 - (1 - \hat{\omega})q^{-1}e_k$</td>
</tr>
<tr>
<td>$\beta \neq b$, $d \neq 0$</td>
<td>$1 - \theta q^{-1}$</td>
<td>$\omega$</td>
<td>$1 - (1 - \hat{\omega})q^{-1}e_k$</td>
</tr>
</tbody>
</table>

(a. For all cases, the output target $r_k$ stays constant;
(b) Known variables in Table 3-1: \( b \) is the nominal process gain in the model; \( \omega \) is the EWMA weighting factor; \( d \) is the metrology delay, \( d = D_{\text{eff}} - 1 \), where \( D_{\text{eff}} \) means the effective time delay. This definition comes from the work of Su[112];

(c) Unknown variables in Table 3-1: \( \beta \) is the real process gain; \( \theta \) is the IMA parameter of the real disturbance \( \alpha_k \cdot \alpha_{k+1} = \alpha_k + \varepsilon_{k+1} - \theta \varepsilon_k \).

(d) Filter adjustment factor: \( \delta \). \( \delta \) is set to be 1 when the estimated disturbance \( \hat{\varepsilon} \) is used in control performance assessment (CPA II). The value of \( \delta \) can also be changed to find the optimal value of \( \omega \) in control performance optimization (CPO) (which will be presented in Chapter 4.)

3.1.3 Control Performance Assessment Based on Q Filter (CPA II)

In Section 3.1.1, it has been shown that in the case without set point change or metrology delay, if one assumes there is no model plant mismatch, then the optimal controller yields the optimal observer, and in other words, it means the estimated disturbance which indicates the observer error should be white noise. When time delay is considered, the estimated disturbance is not exactly equal to the observer error, but it is still required to be white noise in order to ensure optimal control performance.

When the filter Q is designed as shown in Table 3.1, the general mathematical expression of the estimated disturbance can be expressed as:

\[
\dot{\varepsilon}_k = \left( \frac{1 - q^{-1} + \omega q^{-1-d}}{1 - q^{-1} + \omega \frac{B}{D} q^{-1-d}} \right) \left( \frac{1 - \theta q^{-1}}{1 - \left( 1 - \hat{\omega} \right) q^{-1}} \right) \varepsilon_k
\]  

\[
\approx \left[ \frac{1 - \theta q^{-1}}{1 - \left( 1 - \hat{\omega} \right) q^{-1}} \right] \varepsilon_k = \left[ 1 - \frac{\left( \hat{\omega} - (1 - \theta) \right) q^{-1}}{1 - \left( 1 - \hat{\omega} \right) q^{-1}} \right] \varepsilon_k
\]

\[
\dot{\varepsilon}_k \approx \left[ 1 - \left( \left( \hat{\omega} - (1 - \theta) \right) q^{-1} \right) \right] \varepsilon_k
\]

When \( \hat{\omega} = (1 - \theta) \), \( \dot{\varepsilon}_k \approx \varepsilon_k \). From Table 3.1, it can also be found that \( \hat{\omega} = (1 - \theta) \) leads to \( \omega = \omega_{\text{opt}} \), which means the controller is optimal. Then it can be concluded that
when the estimated disturbance $\hat{e}_k$ is close to white noise, the control performance is optimal.

In Chapter 2, it has been proved that when the output error is not white noise (suboptimal control), estimation of value $\theta$ is possible to be achieved although it is not precise enough to help in controller design. Discussion in Chapter 2 also shown that when the control performance is optimal, output error is white noise. As a result, the polynomials A and C in ARMAX model will be the same and this pole-zero cancellation will lead to inaccurate estimation.

However, when a time series is near to white noise, the ARMAX regression method can still be applied to check the structure of the time series.

As analyzed in Eq. (3-15),

$$\hat{e}_k \approx \left[1 - \left(\hat{\omega} - (1 - \theta)\right)q^{-1}\right]e_k$$

(3-15)

Then we can apply ARMAX regression model with $B=0$, and $A=1,C = 1 + C_2q^{-1}$, the only value need to be estimated is $M = -C_2 \approx \hat{\omega} - (1 - \theta)$. When $M$ approaches to 0, the estimated disturbance $\hat{e}_k$ is nearly white noise , which indicate the control performance is optimal; otherwise, when $M$ is far away from 0, the control performance is not optimal.

In this Chapter, we select an range of $M$ from -0.1 to 0.1 as the optimal control range, and when $M$ is outside this range, the control performance is considered as suboptimal. When $-0.1 \leq M \leq 0.1$, $|\hat{\omega} - (1 - \theta)| \leq 0.1$, the difference between real $\omega$ and optimal $\omega$ is about $(1 - d\omega)^2 \times 0.1$, which is smaller than 0.1. When the difference between real $\omega$ and optimal $\omega$ is very small, it is not a good idea to change the current controller tuning factor. This range is only a theoretical value for the simulation, and when applying this concept to industry, it should be adjusted to the special characteristics of process. More discussion of this optimal range will be presented in Chapter 4.
3.1.4 Improved ARMAX Regression

In CPA I, in order to check the variance of the estimated value of A, B, and C, ARMAX regression is applied to different windows and for one special window, only one estimation result will be achieved. Usually the window size should no less than 50. However, in real industry, 50 data points mean too many runs and it is not acceptable to wait for 50 runs for only one analysis step. In CPA II, the value of ARMAX regression result rather than the variance is needed. As a result, it is not necessary to separate the original time series into different pieces and calculate the regression result of every piece. A recursive regression method can be applied to estimate the model parameter and the estimation result can be updated every run.

ARMAX model can be expressed as [113]

\[ A(q)y(t) = B(q)u(t) + C(q)\epsilon(t) \]  

(3-16)

Because the noise \( C(q)\epsilon(t) \) is not white noise, linear regression cannot be used for model parameters estimation. Then recursive method is applied to estimate the model parameters.

As

\[ C(q) = 1 + C(2)q^{-1} + C(3)q^{-2} + \cdots + C(n_c)q^{-n_c+1} \]  

(3-17)

where \( q \) is the shift operator, and \( q^{-1}\epsilon(t) = \epsilon(t - 1) \).

Then Eq.(1) can be rewritten as

\[ A(q)y(t) = B(q)u(t) + (C(q) - 1)\epsilon(t) + \epsilon(t) \]  

(3-18)

Which means

\[ y(t) + A(2)y(t - 1) + \cdots A(n_a)y(t - n_a + 1) = B(1)u(t) + B(2)u(t - 1) + \cdots + B(n_b)u(t - n_b + 1) + C(2)\epsilon(t - 1) + \cdots + C(n_c)\epsilon(t - n_c + 1) + \epsilon(t) \]

Let \( t = [u(t), u(t - 1), \ldots, u(t - n_b + 1), \epsilon(t - 1), \ldots, \epsilon(t - n_c + 1)] \) , \( Z(q) = [B(1), B(2), \ldots B(n_b), C(2), \ldots, C(n_c)]^T \), then Eq. (3-18) can be rewritten as
Then a recursive method can be applied to estimate the value of $Z$.

Let $\epsilon(t) = \hat{\epsilon}(t)$, then we have

$$x(t) = [u(t), u(t - 1), ..., u(t - n_b + 1), \epsilon(t - 1), ..., \epsilon(t - n_c + 1)]^T$$

(3-20)

$$\epsilon(t) = y(t) - x^T(t)\hat{Z}(t - 1)$$

(3-21)

$$P(t) = \frac{1}{\varphi} \left\{ P(t - 1) - \frac{P(t - 1)x(t)x^T(t)P(t - 1)}{\varphi + x^T(t)P(t - 1)x(t)} \right\}$$

(3-22)

$$K(t) = P(t)x(t)$$

(3-23)

$$\hat{Z}(t) = \hat{Z}(t - 1) + K(t)\epsilon(t)$$

(3-24)

Here, $\varphi \in (0, 1)$ is the **forgetting factor coefficient**. If $Z$ is constant, $\varphi$ is selected to be 1, and information from previous data is used as much as possible. If $Z$ varies, $\varphi$ should be a value smaller than 1, and as a result, the estimation can give a quick response to the change of $Z$.

In CPA II method, ARMAX regression shown in this section is applied to the estimated disturbance, and the model is assumed as

$$\hat{\epsilon}(t) = \epsilon(t) + M\epsilon(t - 1)$$

(3-25)

If $M$ is close to 0, the control performance is considered as optimal; otherwise, the control performance is not optimal. As the value of $M$ is not considered as constant, a proper value of $\varphi$ should be selected.

As shown in Eq. (3-22), the value of $\varphi$ ($0 \leq \varphi \leq 1$) can change the convergence speed. The larger the $\varphi$, the more information of previous data can be used. As a result, the response speed will be slower. Otherwise, if the value of $\varphi$ is smaller, less information of previous runs can be used, then the variance of the estimated results will be very large and erratic. The purpose of CPA is try to detect a change in the control
performance, so we need to select a proper value of \( \varphi \) to make sure that the response speed is quick enough and the variance is small enough.

Parameters used in Simulation 3-1 are shown in Table 3-2. A step disturbance was introduced to \( \theta \) at \( i=2001 \). As a result, the optimal control performance \( (1 \leq i \leq 2000) \) became suboptimal \( (2001 \leq i \leq 6000) \). Six different values of \( \varphi \) are selected and the results are shown in Fig. 3-2.

<table>
<thead>
<tr>
<th>Run Number i</th>
<th>( \beta )</th>
<th>( b )</th>
<th>( \theta )</th>
<th>( \omega )</th>
<th>( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 2000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>2001 to 6000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

When \( \varphi = 1 \), the oscillation of the estimated \( M \) is smallest, but the step change cannot be detected. When \( \varphi \) is smaller, then change can be detected, but the oscillation of the estimated result is too large. As shown in Fig. 3-1, when \( \varphi = 0.998 \), the estimated result of \( M \) can give a quick and obvious response to the control performance degradation, and the oscillation of the estimated result is also small enough. In the following research, \( \varphi \) is selected as 0.998.
Figure 3-1 Influence of forgetting coefficient in ARMAX regression (Simulation 3-1)

3.2 Simulation Results of CPA

Four cases are simulated to evaluate the advantages of the new CPA II method. Output targets $r_k$ of all four studied cases are the same, $r_k = 0$. For every case, the total number of runs is $N=6000$. In Simulations 3-2, 3, and 4, time delay is 0, and in Simulation 3-5, the time delay is 1. For $1 \leq i \leq 2000$, the control performance is set to be “optimal”. Starting at $i=2001$, different disturbances are added to the process. In Simulation 3-2, 4 and 5, step disturbance is added to the IMA model parameter $\theta$, and as a result, the tuning factor becomes incorrect in different cases. In Simulation 3-3, step disturbance is introduced to the real process gain, and then plant model mismatch happens. Details of the disturbances are shown in following subsections.

$\hat{\epsilon}_k$ and $M$ are calculated for every run and the result of $M$ from CPA II are shown in the figures. The estimated results of $-C(2)=1-\omega^*(\beta/b)$ which is calculated by CPA I method (as shown in Chapter 2) are also shown in the figures and the two methods (CPA
I and CPA II) will be compared. In the CPA I method, the window size for ARMAX regression is 100 and the ARMAX regression is updated every 50 runs.

3.2.1: Incorrect Tuning

In Simulation 3-2, $\beta = b = 0.05$ for all runs. At $i=2001$, a step change is added to the value of $\theta$, changing $\theta=0.4$ for $2001 \leq i \leq 6000$ to indicate incorrect tuning. The value of $-C(2)$ in CPA I (Fig. 3-2a) and the value of $M$ in CPA II (Fig. 3-2b) are shown below. In Fig. 3-2a, as in previous research, before $i=2000$, a larger variance for the estimated value of $\theta$ is observed, which indicates optimal control performance. Although the large variance is apparent, it is not so easy to interpret. It is also difficult to define the optimal case, because the value of this large variance may vary from case to case.

![Figure 3-2. ARMAX regression results of output error and estimated disturbance in Simulation 3-2. (a) Result of $-C(2)$ from CPA I; b) Result of $M$ from CPA II](image)

Both problems are handled well by the new method. In Fig. 3-2b, at $i=2001$, the estimated parameter (which is expressed as $M$ in the ARMAX model in MATLAB) is far away from 0. As discussed in Section 3.1, when the control performance is optimal, the value of $M$ should be near 0; $M$ different from 0 may indicate that the control performance is not optimal. The difference between “optimal” and “suboptimal” in Fig. 3-2b is more obvious than that in Fig. 3-2a, and it is also very easy to define a benchmark.
for the “optimal”. When the deviation of tuning factor from optimal is smaller than 0.1, it is difficult to be detected and thus not necessary to improve the controller.

### 3.2.2: Plant Model Mismatch

In Simulation 3-3, \( \theta = 1 - \omega = 0.8 \) in all runs. At \( i=2001 \), a step change is added to the value of \( \beta \), and as a result \( \beta=0.1 \) for \( 2001 \leq i \leq 6000 \), indicating plant model mismatch. The value of \(-C(2)\) in CPA I (Fig. 3-3a) and the value of \(M\) in CPA II (Fig. 3-3b) are shown below. When the set point \( r \) is constant, even if \( \beta \neq b \), the new method still works satisfactorily. Also, the performance of the new method can be visually identified from the figures, thus it is very easy to distinguish “optimal” from “suboptimal” control performance. From Fig. 3-3, it can be found that, compared to CPA I presented in Chapter 2, CPA II is more sensitive to the performance change caused by plant model mismatch.

![Diagram](image)

**Figure 3-3.** ARMAX regression results of output error and estimated disturbance in Simulation 3-3. (a). Result of \(-C(2)\) from CPA I; b) Result of \(M\) from CPA II
3.2.3: High-Mix Parameters

In Simulation 3-4, $\beta = b = 0.05$ for all runs, but $\theta$ and $\omega$ change from run to run, which is indicative of a high-mix situation with changing product or tool models.

For $i=1:2000$

$$\theta(t) = 0.8 - 0.1\sin(t/50) \quad (3-26-a)$$

$$\omega(t) = 1 - \theta(t) = 0.2 + 0.1\sin(t/50) \quad (3-26-b)$$

For $i=2001:6000$

$$\theta(t) = 0.6 - 0.1\sin(t/50) \quad (3-27-a)$$

$$\omega(t) = 0.2 + 0.1\sin(t/50) \quad (3-27-b)$$

So the control performance is “optimal” from $i=1$ to $i=2000$, but it is not optimal from $i=2001$ to $i=5000$.

From Fig. 3-4a, it can be observed that when the value of $\theta$ and $\omega$ vary from run to run, the ARMAX regression result of the output error $e$ in CPA I method does not offer useful or reliable information. From Fig. 3-4b, the new method can still distinguish the optimal and suboptimal performance.

Figure 3-4. ARMAX regression results of output error and estimated disturbance in Simulation 3-4. (a). Result of $-C(2)$ from CPA I; b) Result of $M$ from CPA II
Most of the suboptimal behavior can be detected and shown in red points. In high-mix processes, parameters such as $\theta$ and $\beta$ may change from run to run, because different runs may be performed in different reticles. Thus CPA II in this Chapter is much better than CPA I (Chapter 2) for high-mix processes.

### 3.2.4: Metrology Delay

In Simulation 3-5, $\beta = b = 0.05$ for all runs and time delay $d = 2$. For $1 \leq t \leq 2000$, $\theta = 0.4$ and $\omega = 0.3750$, which is the optimal choice of $\omega$ in this case. At $i=2001$, a step change is added to the value of $\theta$ and as a result $\theta = 0.9$ from $i=2001$ to $i=5000$, and thus the control performance becomes suboptimal. As shown in Table 3-1, when the effective time delay is not 0 and there is no plant-model mismatch, $\hat{e}_k$ can be calculated as $\hat{e}_k = \frac{1-q^{-1}+\omega q^{-1-d}}{1-(1-\hat{\theta})q^{-1}}q^d e_k$. The results of both CPA I and CPA II are shown in Fig. 3-6. From Fig. 3-6a, when the time delay is not 0, the ARMAX regression result of output error $e$ does not offer correct and reliable information. From Fig. 3-6b, the ARMAX regression of $\hat{e}_k$ can still distinguish the optimal and suboptimal cases. In Fig. 3-5b, before $i=2000$, $M$ is near 0 and the control performance is optimal; after $i=2001$, $M$ is obviously larger than 0, which indicates that the control performance after $i=2001$ is suboptimal, and the value of $\omega$ should be reduced to make the control performance better.

The variance of $e_k$ in Simulation 3-5 is shown in Fig. 3-6. It is interesting that $\text{var}(e_k)$ after $i=2001$ is even smaller than $\text{var}(e_k)$ before $i=2001$. It is because the value of $\theta$ after $i=2001$ is larger than that before $i=2001$. When the time delay $d$ is not 0, a larger value of $\theta$ can lead to better control performance with smaller variance. This is because when the values of $\theta$ are different, the best achievable minimum variance is different. The relationship between the $\min(\text{var}(e))$ and $\theta$ is

$$
\min(\text{var}(e)) = (1 + d \cdot (1 - \theta)^2)\text{var}(e)
$$

(3-23)
As result, when $\theta$ is larger, the best possible performance is better than that with smaller $\theta$. So when the time delay is not zero, the change of variance is not valid indication of the degradation of control performance.

Figure 3-5. ARMAX regression results of output error and estimated disturbance in Simulation 3-5. (a). Result of $-C(2)$ from CPA I; b) Result of M from CPA II

Figure 3-6. Variance of process output error vs. Run Number $i$ (Simulation 3-5)

3.3 Application of Direct ARMAX Regression

Because the ARMAX regression method has been modified in Section 3.1, in this section, we will check how the direct ARMAX regression method works in different cases. The forgetting factor is $\varphi = 0.998$ as indicated before. Parameters used in the
simulations are shown in Table 3-3. For all simulations, the control performance before
$i = 3000$ is optimal. After $i = 3000$, step change was introduced to either $\beta$ or $\theta$, and
control performance became suboptimal.

<table>
<thead>
<tr>
<th>Simulation 3-6-a</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 3-6-b</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.10</td>
<td>0.05</td>
<td>0.6</td>
<td>0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 3-7-a</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.285</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 3-7-b</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.10</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 3-8-a</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Varied $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.285</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation 3-8-b</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Varied $r_k$)</td>
<td>1 ≤ $i$ ≤ 3000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3001 ≤ $i$ ≤ 6000</td>
<td>0.10</td>
<td>0.05</td>
<td>0.6</td>
<td>0.285</td>
<td>1</td>
</tr>
</tbody>
</table>

ARMAX regression method was applied to calculate two parameters of the
process: $\beta/b - 1$ and $\theta$. The value of $\beta/b - 1$ is related to the plant model mismatch,
and $\theta$ is related to the disturbance model. Results of all simulations are shown below.

In Simulation 3-6, there is no metrology delay or set point change, and as shown
in Fig. 3-7, the result of direct ARMAX regression is ineffective to detect the control
performance change or process parameter change. It is the same as shown in Chapter 2.
In Simulation 3-7, the set point is still kept constant, but metrology delay is considered not to be 0. When the metrology delay \( d \) is not 0, even when the control performance is optimal, the process output error is not white noise. As a result, in this case, process parameters can be estimated in the situation with both optimal and suboptimal control. But the estimated result of \( \theta \) is not precise enough. As shown in Fig. 3-8(a), when the parameter \( \theta \) has been changed in the simulation, the estimated result did not shown very obvious change. However, in Fig. 3-8(b), it is gratifying to see that in the case with metrology delay, direct ARMAX regression is possible to detect the plant model mismatch even if the set point always kept constant. In CPA II method presented in sections 3.1 and 3.2, plant model mismatch is not considered. If the ARMAX regression is also applied to the process output error, then it is possible to estimate the value of \( \beta/b \) and modify the Q filter in CPA II method. There is also another problem. From Fig. 3-8-a, a small step disturbance can be found in the figure of the estimated value of \( \beta/b \), but in fact, only the value of \( \theta \) has been changed. As a result, the estimated result of \( \beta/b \) in this case is not accurate.
In simulation 3-8, the metrology delay is not 0, and the set points varied from run to run. Again, it can be found that when ARMAX regression method is applied to the process output error directly, it is possible to detect the plant model match, but it is not very sensitive to the change of $\theta$. The direct ARMAX regression result cannot give direct and solid conclusion about whether the control performance is good or not, but parameters estimated from this method can help to modify the CPA II method by offering information about plant model mismatch.
Figure 3-9. Parameter estimation by directly ARMAX regression. ((a), Simulation 3-8-a; (b) Simulation 3-8-b)

From the simulation results, it can be found that when there is no time delay and control performance is optimal, it is impossible to get precise estimation of the ARMAX model parameters. For both cases with constant set points, when $\theta$ changes, the estimated value of $\beta/b$ may also change. As a result, it is difficult to determine whether the control performance change is from the change of $\beta/b$ or $\theta$. In conclusion, if one applies ARMAX regression directly to the process output error rather than the estimated disturbance, it is not possible to achieve definitive results about control performance assessment. However, when the metrology delay is not zero, this regression is able to estimate some process parameters (although maybe not accurately), especially with plant model mismatch. This is a good complement to the CPA II method.

3.4 SUMMARY

In this chapter, the error source of this RtR EWMA controller system is reconsidered as a combination of observer error and plant model mismatch error. In the case when there is no set point change, a time series called “estimated disturbance” is generated from the Q filter, and this disturbance signal is considered as the observer
error. When the estimated disturbance is white noise, the control performance is optimal. Then the CPA II method is developed on the ARMAX regression result of the estimated disturbance. A recursive regression method is applied for this estimation and simulations shown that this new CPA II method is sensitive to the performance degradation caused by improper controller tuning or model plant mismatch in the case without set point change. Simulations also proved that the CPA II method in this chapter works very well in the case with metrology delay and high-mix process parameters, which indicates that this method is possible to be applied to industry data of high-mix semiconductor manufacturing processes.

In the following chapter, this method will be applied to industry data and the results will be discussed.

Once suboptimal behavior is detected, the next work for the control engineer should be diagnosing the suboptimal performance, trying to determine the reason and fixing the problem. This will also be discussed in the next chapter.
CHAPTER 4
CPA AND CPD IN THREADED CONTROL SYSTEM

In this chapter, the new CPA II method is applied to industrial data from a high-mix photolithography process in Texas Instruments. Preliminary analysis of industrial data will be presented in Section 4.1, and the results indicate that the theoretical optimal range of $M$ (from -0.1 to 0.1) is not always valid in real applications. In Section 4.2, the influence of metrology delay and model plant mismatch to the determination of the optimal range of $M$ will be discussed. Later in Section 4.3, a method named CPO (control performance optimization) will be presented to help determine the optimal range of $M$. With the CPO method, the best achievable minimum variance can also be calculated and a value of the controller tuning factor $\omega$ which can improve the control performance can be suggested.

4.1 TI DATA ANALYSIS WITH CPA II

4.1.1 Model Used in TI

In the high-mix photolithography process, the controlled variable is the printed image critical dimension (CD), and this variable can be expressed as a linear function of the exposure energy.

$$CD = \text{Slope} \times \text{Exposure} + B_{\text{base}} + B_{\text{reticle}} \quad (4-1)$$

Where

- $CD$ is the line (or hole or space) width (micron, measured by SEM)
- $\text{Slope}$ is the process gain, and it can be determined by experiment. (J/m²/µm)
- $\text{Exposure}$ is the manipulated variable, which indicates the exposure energy used for the photolithography process.
- $B_{\text{base}}$ is the machine/technology/tool offset (micron), and $B_{\text{reticle}}$ is the offset for a special reticle (micron).
Different from the basic model shown in Chapters 2 and 3, in this process, there are two different offsets or disturbance signals. In semiconductor manufacturing applications, the offset $B_{\text{base}}$ is usually considered as an IMA time series, and an EWMA tuning strategy is used to modify the current value of $B_{\text{base}}$.

$$B_{\text{base}}(k + 1) = B_{\text{base}}(k) + \text{Gain}_{\text{base}} \times (CD(k) - \text{Predicted CD}(k)) \quad (4-2)$$

Where $\text{Predicted CD}$ is usually equal to the set point, and $\text{Gain}_{\text{base}}$ is the EWMA tuning factor, which usually ranges from 0.1 to 0.3 in real industry.

The offset from special individual reticle can also be updated by EWMA theory as

$$B_{\text{reticle}}(j + 1) = B_{\text{reticle}}(j) + \text{Gain}_{\text{reticle}} \times (CD(j) - \text{Predicted CD}(j)) \quad (4-3)$$

Although same EWMA tuning strategy is used for both $B_{\text{base}}$ and $B_{\text{reticle}}$, the tuning factors $\text{Gain}_{\text{base}}$ and $\text{Gain}_{\text{reticle}}$ are quite different from each other. Usually the value of $\text{Gain}_{\text{base}}$ will stay constant, while $\text{Gain}_{\text{reticle}}$ changes a lot. Details about the control strategy for this process can be found in the work of Stuber[114].

It should be noticed that the update of $B_{\text{base}}$ and $B_{\text{reticle}}$ belongs to different control loops. $B_{\text{base}}$ belongs to a control loop with the same set point, same tool and some other special contexts, and in this control loop, different reticles may be included, while $B_{\text{reticle}}$ is only updated in a control loop with the same reticle and some other contexts. As a result, the values of $k$ and $j$ are not the same for the data of a special run.

### 4.1.2 Analysis of Industrial Data

As shown in Section 4.1.1, the offset result from a special reticle usually tends to be nearly constant after 50 runs. In some processes, the reticle gain is considered to approach to zero after 50 runs, and in some other control methods, this reticle bias is updated as a “dynamic gain”. As this dynamic gain is not constant, also because the reticle bias usually does not change a lot in one control loop, in this section, only the base bias from machine will be used for the CPA analysis.
For a control loop related to the update of base bias, the estimated disturbance is calculated based on the base bias model. The controller tuning factor is $Gain_{\text{base}}$, and the process gain is $Slope$. The only problem is the metrology delay. In real industry, the metrology delay may vary from 0 to 10 runs. As the metrology delay information is not recorded in the given dataset, it is considered as 0 initially, and the influence of the metrology delay to the CPA result will be discussed later.

Three different control loops are analyzed and both the output variance and the M value in the CPA II will be shown.

**Example 4-1:**

In Figure 4-1(a), the output error of this example is shown. In this figure, the $x$ axis is the scanner count number in the dataset, which is indicated by the value $k$ in Eq. 4-2. As the process is high-mix, the total number for one special control loop is not very large. It seems that after the scanner count $i=1430$, most of the process output error is above 0, which looks like suboptimal behavior, but it is too early to conclude that it is suboptimal and it is not impossible to tell why this happens.

Figure 4-1 Process output error (a) and M calculated from CPA II (b) of Example 4-1
In Fig. 4-1(b), the value of $M$ in CPA II is calculated and shown. As the RLS method needs at least 20 data to make a relatively stable estimation, the first 20 points are not accurate and should not be considered during the assessment of the control performance. Again, we select the range from -0.1 to 0.1 as the optimal control range, and from Fig. 4-2, it can be found that after $i=1430$, an obvious change in $M$ can be detected. As discussed in Chapter 2, the value of $M$ is below the optimal range which indicates the current controller tuning factor is smaller than the theoretical optimal value.

**Example 4-2:**

Process error and $M$ of Example 4-2 are shown in Fig. 4-2(a) and Fig. 4-2(b). In this control loop, no special change can be detected from either the process error behavior or the behavior of $M$. All the points of $M$ in this example are located inside the theoretical optimal range from -0.1 and 0.1. But the interesting thing is, from $i=3820$, an obvious decrease of $M$ can be detected again, and the time of this decrease is near to the decrease in Example 2 (Feb, 26, 2012).

**Example 4-3**
The process error of Example 4-3 is shown in Fig. 4-3(a), and it seems that the variance of the process output error became larger after \(i=2220\) in this example. Usually it is related to some degradation of the control performance. Checking the \(M\) value of CPA II (Fig. 4-3(b)), it can be found that the value of \(M\) decreased obviously after \(i=2220\), and this can be considered as a sign of the control performance change. But unfortunately, in this example, the value of \(M\) after \(i=2220\) is located inside the range from -0.1 to 0.1, which is considered to be optimal in our theory.

Figure 4-3. Process output error (a) and \(M\) calculated from CPA II (b) of Example 4-3

As discussed in Chapter 3, the results from output error analysis and the CPA II method may conflict with each other in the case with metrology delay. As the exact metrology delay is not recorded in this dataset, it is difficult to check the real influence, then we assume the metrology delay to be 0, 1 or 2, and calculate the value of \(M\) again. The result is shown in Fig. 4-4. It can be found that for the same dataset, when the assumed metrology delay is different, the value of \(M\) will be different. Although all three lines are inside the same range, the line of \(M\) with smaller metrology delay has more obvious change. When metrology delay \(d\) is assumed to be 0, no special change of \(M\) can be concluded in this example, but if we assume \(d=0\), the CPA II analysis result may
indicate an obvious control performance change. In order to achieve more accurate CPA conclusion, it is necessary to record the metrology delay for every run.

**Analysis of Bias from Reticle**

The analysis shown before is only concerned about the whether the controller tuning factor $G_{Base}$ is optimal or not. As data for one specific control loop with the same reticle is not sufficient to generate a valid ARMAX regression result, the bias from different reticles was not analyzed with the CPA II method.

![Figure 4-4. Influence of estimated metrology delay (Example 4-3)](image)

Once suboptimal behavior was detected by CPA II method, this “suboptimal” behavior may result from improper selection of either $G_{Base}$ or $G_{Reticle}$. The simplest way is to calculate the ratio of the variance of runs of specific reticles to the variance of the runs of the whole control loop. If the ratio is obviously larger than 1, it may indicate that this reticle needs special attention. The variance ratios of Example 3 are calculated and shown in Table 4-1. It can be found that the reticle “E575-599-001” had a major contribution to the variance of the whole control loop. Marking the data points from this reticle with solid squares in the figure of process output error (Fig. 4-5), it can be easily found that most points far away from the mean value came from this reticle. Then special
attention should be paid to this reticle. More practical experience about this process is
needed to diagnose and improve this problem.

Figure 4-5. Process output error of Example 4-3 (Solid squares indicate data from Reticle
E575-599-001)

Table 4-1 Error Analysis of different reticles in Example 4-3

<table>
<thead>
<tr>
<th>Reticle</th>
<th>Number of samples</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>E551-512-001</td>
<td>1</td>
<td>0.265</td>
</tr>
<tr>
<td>E554-086-001</td>
<td>29</td>
<td>0.458</td>
</tr>
<tr>
<td>E575-599-001</td>
<td>53</td>
<td>1.477</td>
</tr>
<tr>
<td>E587-105-001</td>
<td>2</td>
<td>1.224</td>
</tr>
<tr>
<td>E594-627-001</td>
<td>9</td>
<td>0.241</td>
</tr>
<tr>
<td>E624-902-001</td>
<td>1</td>
<td>4.371</td>
</tr>
</tbody>
</table>

4.2 Determination of Optimal Range in CPA II

In Example 4-3, when $M$ is inside the optimal range, the variance of the process
output is larger than the runs with $M$ outside the optimal range. As discussed before, this
may be a result of varied metrology delay. However, it also indicates that the theoretical
optimal range shown in Chapter 3 needs some improvement when it is applied to real
industrial data. In this section, possible factors which may affect the CPA optimal range
will be discussed, including plant model mismatch, metrology delay and inaccurate
disturbance model.
4.2.1 Inaccurate Estimation of Metrology Delay

In this simulation 4-1, $\beta = b$ and the real metrology delay $d = 1$. Other parameters are shown in Table 4-2.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$\omega_{opt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 2000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.4</td>
<td>0.375</td>
<td>0.375</td>
</tr>
<tr>
<td>2001 to 6000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.6</td>
<td>0.375</td>
<td>0.285</td>
</tr>
</tbody>
</table>

The estimated metrology delay used for the calculation of estimated disturbance $\hat{\epsilon}$ is named $td$, and $td$ is selected as 0, 1 and 2 in this simulation. Different values of $M$ are shown in the figure. It can be found that when the metrology delay used for the CPA method is not accurate, the basic shape of the figure of $M$ will not change a lot, but the value will be quite different. In this example, from $i=1$ to 2001, the control performance is optimal because $\omega = \omega_{opt}$. As shown in Fig. 4-6, only when the estimated metrology delay $td = d$, the points of $M$ are near 0 and indicate optimal control performance. When $td$ is quite different from the real metrology delay, the change of the control performance can still be detected from the shape change of the figure of $M$, but the optimal range is not from -0.1 to 0.1 as discussed in Chapter 3. In the industrial data analysis Example 3, the runs with smaller variance had a value of $M$ far away from 0, and the inaccurate estimation of metrology delay may be one reason of this confliction.

4.2.2 Varied Metrology Delay

In the data set offered by TI, the metrology delay is not listed. However, it is possible to record the accurate metrology delay when adjusting the recipe of a specific run. In semiconductor manufacturing processes, the metrology delay may vary in a very large range, e.g., 0 to 10. In this high-mix batch process, it is not so difficult to know the exact metrology delay for every run. Assume that the accurate metrology delay for every run has been recorded, then the filter Q can be adjusted by selecting the accurate value.
of $\omega$ and $d$ for every run. As a result, the varied metrology delay is not a big problem for the CPA II method shown in Chapter 3.

![Figure 4-6. $M$ of Simulation 4-1 with different estimated metrology delay](image)

In simulation 4-2, the metrology delay is randomly selected from 0 to 4, and the value of $\theta$ changes after $i=2000$. The value of $\omega$ is selected as 0.15. With the adjusted Q filter, the value of $M$ can be found in Fig. 4-7.

| Table 4-3 Parameters used in Simulation 4-2 |
|-----------------|----------------|---------|---------|---------|---------|
| $i$             | $\beta$       | $b$     | $\theta$| $\omega$| $d$     |
| 1 to 2000       | 0.05           | 0.05    | 0.4     | 0.15    | 0, 1, 2, 3, 4 |
| 2001 to 6000    | 0.05           | 0.05    | 0.6     | 0.15    | 0, 1, 2, 3, 4 |
It is significant that the control performance change resulting from the change of \( \theta \) can still be clearly detected from the behavior of \( M \). As shown in Fig 4-7(a), the value of \( M(i) \) is always below 0 which indicate the controller tuning factor \( \omega \) is a little bit smaller than the optimal one. Changing \( \omega \) from 0.15 to 0.2, and re-doing the simulation, the variance is shown in Table 4-4. It can be found that by increasing \( \omega \) to 0.2, the output variance for both ranges has been decreased, and the control performance is improved. The behavior of \( M \) with \( \omega = 0.2 \) for this case is shown in Fig. 4-7(b). Now it can be found that for the first range \( (i=0 \) to 2000), the value of \( \omega \) should be enlarged to approach better performance, while for the second range \( (i=2001 \) to 6000), \( M \) is near 0 and it is better not to change \( \omega \). If we change \( \omega \) to 0.24, and re-do the simulation, the variance of the first range is decreased and the variance of the second range is increased, which means our prediction from Fig. 4-7(b) is just right.

**Table 4-4 Process output variance of Simulation 4-2 and 4-3**

<table>
<thead>
<tr>
<th></th>
<th>( d ) varies from 0 to 4</th>
<th>( d ) varies from 0 to 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \omega = 0.15 )</td>
<td>( \omega = 0.2 )</td>
</tr>
<tr>
<td>( i=1 ) to 2000</td>
<td>2.3222</td>
<td>2.1341</td>
</tr>
<tr>
<td>( i=2001 ) to 6000</td>
<td>1.4563</td>
<td>1.3834</td>
</tr>
</tbody>
</table>
However, we have another problem. When \( \hat{\omega} > 1 \), \( \hat{\epsilon} \) will be unstable. As a result, when \( \omega \geq 1/d \), CPA II method cannot give acceptable results. Even if when \( \omega \) is near to be \( 1/d \), analysis of \( M \) is not very good. In this example, when \( d=4 \), \( 1/d=0.25 \) which is quite near to be 0.24. As a result, the regression of \( M \) is very unstable as shown in Fig.4-8(a). In order to solve this problem, during the calculation of \( \hat{\omega} \), if \( d \geq 1/\omega \), then select \( d = \lfloor 1/\omega \rfloor \). By applying this modified CPA II method, the \( M \) can be recalculated and the result is shown in Fig. 4-8(b).

![Figure 4-8. M of Simulation 4-2 (\( \omega = 0.24 \), (a). CPA II method; (b) Modified CPA II method)](image)

The results for \( M \) in Fig 4-8(b) indicates that \( \omega = 0.24 \) is a better choice than 0.2 for \( 0 \leq i \leq 2000 \), and the variance shown in Table 4-4 also agrees with this conclusion. In Simulation 4-3, we select the metrology delay \( d \) from 0 to 10 as for industrial data, and make \( \omega = 0.15 \). All other parameters are the same as Simulation 4-2. Applying the modified CPA II method, the result of \( M \) is shown in Fig. 4-9(a). From this figure, the value of 0.15 seems a little bit smaller than the optimal value. Then we increase \( \omega \) to be 0.2 and re-do the simulation, and the value of \( M(i) \) is shown in Fig. 4-9(b), which indicates that the control performance has been improved by increasing \( \omega \). Variance for both ranges in the two simulations can also be found in Table. Output error variance with
\( \omega = 0.2 \) is smaller than the case with \( \omega = 0.15 \), which is also an evidence that the control performance has been improved.

### 4.2.3 Plant Model Mismatch in The Case with Metrology Delay

Although the metrology delay can be achieved from the manufacturing data, it is still difficult to determine whether there is model plant mismatch or not. In the design of Q filter, model-plant mismatch is not considered. This works OK when the delay is 0, but when the metrology delay is not known, it is necessary to study the influence of model plant mismatch to the result of CPA II.

![Figure 4-9. M of Simulation 4-3 (d is selected from 0 to 10. (a) \( \omega = 0.15 \); (b) \( \omega = 0.2 \)](image)

In Simulation 4-4, \( b=0.05 \) for all runs, a step disturbance is introduced to \( \beta \) at \( i=2001 \) and \( i=4001 \). As a result, the value of \( b/\beta \) varies in the three regions. \( \theta = 0.6 \) and the time delay \( d = 1 \). Three simulations with different \( \omega \) are carried out. According to Eq. (2-17) and the variance shown in Table 4-5., it can be found that, for Simulation 4-4-a, the first region (\( i \) from 0 to 2000) is optimal; for Simulation 4-4-b, the second region (\( i \) from 2001 to 4000) is optimal; for Simulation 4-4-c, the third region (\( i \) from 4001 to 6000) is optimal. \( M \) calculated from CPA II for all three cases are shown in Fig. 4-10. For all three cases, \( M \) can show the change of the control performance, but unfortunately, the
range from -0.1 to 0.1 is not a solid evidence of optimal control performance when \( b/\beta \neq 1 \). For example, in Simulation 4-4-a, the first region \( (1 \leq i \leq 2000) \) is optimal, but the value of \( M \) is far away from 0. In order to apply the CPA II method to real industry data, more research about the determination of the optimal range (or in another word, the “benchmark”) is required. In the next section, a method based on the calculation of estimated disturbance \( \hat{e} \) will be developed to help determine the optimal range.

![Graph showing the influence of model plant mismatch on the optimal range of M in CPA II](image)

**Figure 4-10. Influence of model plant mismatch to the optimal range of \( M \) in CPA II (Simulation 4-4)**

**Table 4-5 Parameters used in Simulation 4-4**

<table>
<thead>
<tr>
<th>Simulation 4-4</th>
<th>( \omega )</th>
<th>( 1 \leq i \leq 2000 )</th>
<th>( 2001 \leq i \leq 4000 )</th>
<th>( 4001 \leq i \leq 6000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 4-4-a</td>
<td>0.428</td>
<td>1.1834</td>
<td>1.1905</td>
<td>1.4574</td>
</tr>
<tr>
<td>Simulation 4-4-b</td>
<td>0.285</td>
<td>1.2618</td>
<td>1.1407</td>
<td>1.2263</td>
</tr>
<tr>
<td>Simulation 4-4-c</td>
<td>0.228</td>
<td>1.348</td>
<td>1.1632</td>
<td>1.2051</td>
</tr>
</tbody>
</table>
4.3 Control Performance Optimization

The calculation of best achievable minimum variance is one of the most important underpinnings in control performance assessment. Performance index which is defined based on Harris’ theory (1989) can be expressed as follows.

\[
\text{Performance Index} = \frac{\text{Actual Output Variance}}{\text{Best Achievable Minimum Variance}}
\]

Here, the actual output variance can be calculated easily from the output data. Much work has been done to calculate the best achievable minimum variance of different controllers. Prabhu (2007) proposed a method to calculate the best achievable minimum variance for run-to-run control system. In his research, an iterative expression of the run-to-run control system was developed and the output \( y \) is expressed as

\[
Y = (I + Sk)^{-1} C
\]

Here \( S \) and \( C \) can be calculated from controller structure information and impulse response coefficients. Then an optimal value of controller parameter \( k_i \) that minimizes the variance of \( Y \) can be calculated using Newton’s method. Then optimal output value \( Y_{opt} \) can be expressed as a function of output \( Y \) and optimal \( k_i \). Then the best achievable variance is variance of \( Y_{opt} \). Theoretically, it is a sound method. But in this method, the problems of plant-model mismatch and time-delay cannot be well handled. Moreover, the calculation of \( S \) and \( C \) is not so direct, more information is needed.

In this section, we develop a new method to calculate the best achievable variance.

As shown in Eq. (3-9) and Fig. 4-10,

\[
\hat{\varepsilon}_k = \left( \frac{1 - q^{-1} + \omega q^{-1-d}}{1 - q^{-1} + \omega \beta q^{-1-d}} \right) \left( \frac{1 - \theta q^{-1}}{1 - (1 - \omega \delta)q^{-1}} \right) \varepsilon_k
\]

By changing the value of \( \delta \), the variance of \( \hat{\varepsilon}_k \) can also be changed. The value of \( \delta \) that leads to the minimum variance of \( \hat{\varepsilon}_k \) can be represented as \( \delta_{opt} \), and the minimum
variance of $\hat{e}_k$ is $\min(var(\hat{e}_k))$. $\delta_{opt}$ can be obtained by Newton’s method. When $\beta = b$ and the metrology delay $d = 0$, $\omega \cdot \delta_{opt}$ is the best value of $\omega$, and $\min(var(\hat{e}_k)) = \text{var}(\varepsilon_k)$ is the best achievable minimum variance. If $\beta \neq b$ or $d \neq 0$, $\min(var(\hat{e}_k)) \neq \text{var}(\varepsilon_k)$, but $\omega \cdot \delta_{opt}$ may yield better performance. If $\delta_{opt} = 1$, no improvement is necessary, and the control performance can be considered as optimal. This method is named CPO (control performance optimization) in this thesis.

![Figure 4-11. Variance of estimated disturbance vs. Q filter parameter adjustments](image)

**Table 4-6 Suggested $\omega$ and new variances in Simulation 4-6**

<table>
<thead>
<tr>
<th>simulations</th>
<th>$\beta$</th>
<th>$b$</th>
<th>$\theta$</th>
<th>$\omega$</th>
<th>$d$</th>
<th>Suggested $\omega$</th>
<th>Current $\text{var}(\varepsilon)$</th>
<th>Optimal $\text{var}(\varepsilon)$</th>
<th>New $\text{var}(\varepsilon)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-6-1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.4</td>
<td>0</td>
<td>0.2</td>
<td>1.1321</td>
<td>1.0445</td>
<td>1.0562</td>
</tr>
<tr>
<td>4-6-2</td>
<td>0.05</td>
<td>0.1</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
<td>0.2860</td>
<td>1.1013</td>
<td>1.0445</td>
<td>1.0661</td>
</tr>
<tr>
<td>4-6-3</td>
<td>0.05</td>
<td>0.05</td>
<td>0.8</td>
<td>0.4</td>
<td>1</td>
<td>0.1667</td>
<td>1.1311</td>
<td>1.0502</td>
<td>1.0428</td>
</tr>
</tbody>
</table>

In control performance assessment, the best achievable performance can be calculated and then applied to the Harris-type performance index. When suboptimal behavior is detected by a Harris-type performance index or the new method (which is sometimes more sensitive to the variance change), the suboptimal process data should be analyzed by changing $\delta$ and an improved $\omega$ will be suggested as $\omega \cdot \delta_{opt}$. Remember, $\omega \cdot \delta_{opt}$ is not the same as the optimal $\omega$. Sometimes, for example, when there is model-
plant mismatch, high-mix process parameters, metrology delay, and so on, it is simply a
better choice than the current one.

Simulations are performed to demonstrate the method. Three different suboptimal
control simulations (4-6-a, 4-6-b, and 4-6-c) are analyzed and improved values of $\omega$ are
calculated. Results of the process with new $\omega$ are also shown in Table 4-6. It can be
observed that for all three cases, the recommended $\omega$ can lead to better performance.

When $\omega \geq 1/d$, it is impossible to obtain the value of improved $\omega$. It is because
in this case, $\hat{\omega} = \frac{\omega}{1-d\omega} < 0$. As a result, $(1 - \hat{\omega} \cdot \delta) > 1$. As shown in Eq. (3-9), one
pole of the expression of $\hat{\epsilon}_k$ is outside the unit circle, so $\hat{\epsilon}_k$ in this case is unstable for all
values of $\delta$, and it is impossible to obtain an improved $\omega$ via the proposed CPO method.
Once this happens, it may indicate that the original $\omega$ is too large to be optimal. When
$\omega < 1/d$, in order to make sure that all $\hat{\epsilon}_k$ are stable and reasonable, the range of $\delta$
should be $0 < \delta < 1/\hat{\omega}$. If the metrology delay $d$ varies from run to run, the calculation
of $\hat{\omega}$ can be adjusted as shown in Section 4.2.2.

With this CPO method, three important problems can be solved.

1) The best achievable minimum variance can be calculated only from the
process output data and parameters which are already known. This minimum variance
can be used in Harris’ type performance index;

2) It is possible to detect whether the control performance is optimal or not in
the case with model-plant mismatch and metrology delay. Once the performance is
considered as “optimal” with the CPO method in Section 4.3, the optimal range of $M$ of
CPA II can be decided.

3) It is possible to recommend a controller tuning factor $\omega$ based on previous
process data, so that a dynamic controller tuning factor can be calculated. As the
metrology delay is also considered in the relationship between $\omega$ and $\hat{\omega}$, it can also help
to improve the control performance of the processes with varied metrology delay.
4.4 Summary and Conclusion:

In this chapter, industry data from Texas Instruments are analyzed by CPA II method presented in Chapter 3. Preliminary results showed that CPA II method is available to detect the control performance change in a threaded EWMA control system. However, the data analysis result also revealed a problem that in real cases, the optimal range of $M$ may be not from -0.1 to 0.1.

Simulations are carried out to discuss the influence of metrology delay and plant model match to the optimal range of $M$. Simulation results showed that inaccurate estimation of metrology delay will make CPA II method unreliable. When applying CPA II method to real industrial data, it is important to record the exact value of metrology delay for every run. Simulations also showed that even if the metrology delay varied from run to run, CPA II method is still available to be applied and the change of control performance can still be detected from the figure of $M$.

My research also showed that when the model-plant mismatch cannot be ignored, the optimal range of $M$ will be changed. In order to determine the baseline of “optimal control”, a method named CPO (Control Performance Optimization) was developed in this chapter. Simulations showed that this CPO method is available to investigate whether the current control performance is good or not even if there is model plant mismatch. The CPO method can also calculate the best achievable minimum variance and give suggestions about how to improve the control performance.

Theoretically, CPA II + CPO method presented in Chapters 3 and 4 is a good solution to the control performance assessment and diagnosis problem of threaded EWMA controller in semiconductor manufacturing processes. In order to apply this method to industrial processes, what should be done by practical control engineer is: (1), record the metrology delay for every run; (2), try to determine the baseline of “optimal control performance” by CPO method; (3), calculate the value of M with CPA II method.
With this method, the control performance change caused by stochastic disturbance can be detected and a solution can be suggested.

However, in the future, non-threaded control with high-mix set points will be a popular research topic. In next chapter, a new non-threaded control method which has been developed by Wang[43] will be presented, and the key idea of this CPA II method will be extrapolated to that non-threaded control system.
CHAPTER 5

CPA FOR NON-THREADED HIGH-MIX CONTROL SYSTEM

In this chapter, a non-threaded controller based on state estimation and Kalman filter which has been developed by Wang[43] will be introduced, and then the CPA II method will be applied to this non-threaded control system. Simulations are carried out to investigate the cases with plant model mismatch and improper controller parameter selection but no metrology delay will be considered. Simulation results indicated that CPA II method is also able to detect control performance degradation in a non-threaded control system. We also show that CPA II method is not a good choice to assess the state estimation performance of every special state, but the analysis result of single process state is more sensitive to the change of process control performance. More work about both controller development and assessment should be done in the future.

5.1 NON-THREADED CONTROL OF HIGH-MIX SEMICONDUCTOR MANUFACTURING SYSTEM

In Chapter 1, some non-threaded control methods have been introduced, such as Kalman filter, JADE and so on. Because the process disturbances in semiconductor manufacturing processes are usually considered as IMA time series, Wang[43] developed a new non-threaded control algorithm, which is called the Kalman-IMA method. In this method, the overall process state is defined as the sum of the mean effect and the contribution of every single context. The observed overall state $z_k$ can be described as follows:

$$z_k = \mu + \sum_i s_{i,k} + v_k$$  \hspace{1cm} (5-1)

where $\mu$ is the mean effect, $v_k$ is the measurement noise which is usually considered as white noise, and $s_{i,k}$ indicates the process disturbance of the $k^{th}$ run comes from context
i. A matrix $H$ is defined to describe which context is included in one special run. The $k^{th}$ column of matrix $H$ corresponds to the $k^{th}$ run, and the $i^{th}$ row of matrix $H$ corresponds to the $i^{th}$ context. If the element $H_{i,k} = 1$, it means the context $i$ is included in the $k^{th}$ run; if $H_{i,k} = 0$, then the context $i$ is not included in the $k^{th}$ run. The extended process state $\psi_k$ is the augmentation of the process state $x_k$ (process disturbance in semiconductor manufacturing processes) with white noise $w_k$.

$$\psi_k = \begin{bmatrix} x_k \\ w_k \end{bmatrix} \quad (5-2)$$

In high-mix semiconductor manufacturing processes, the context of every run in one control loop may consist of different categories. For example, there may be different layers, different products and different tools in one special non-threaded control loop. The context matrix $H$ represents the context information for every run. Then the state space representation of the process should be:

$$\psi_{k+1} = A_k\psi_k + B_ke_k \quad (5-3)$$

$$z_k = C_k\psi_k + v_k \quad (5-4)$$

where $v_k$ is the measure noise, which is considered as white noise with zero mean and variance as $R$. $e_{k,i}$ is also zero mean white noise, which is corresponding to context $i$, and the variance of $e$ is $Q$.

$$A_k = \begin{bmatrix} I & (-\theta)\text{diag}(H_k) \\ 0 & I - \text{diag}(H_k) \end{bmatrix} \quad (5-5)$$

$$B_k = \begin{bmatrix} \text{diag}(H_k) \\ \text{diag}(H_k) \end{bmatrix} \quad (5-6)$$

$$C_k = \begin{bmatrix} \text{diag}(H_k) & 0 \end{bmatrix} \quad (5-7)$$

$$Q_k = \begin{bmatrix} \text{diag}(H_k) \\ \text{diag}(H_k) \end{bmatrix}Q\begin{bmatrix} \text{diag}(H_k) \\ \text{diag}(H_k) \end{bmatrix}^T \quad (5-8)$$

Here $\theta$ is the factor of IMA model, and $\text{diag}(H_k)$ means a diagonal matrix with the $i^{th}$ diagonal element equal to $H_{i,k}$. Then a state estimation method based on Kalman filter[43] can be derived as:

$$\hat{x}_{k+1} = A_k\hat{x}_k + A_kP_kC_k^T(R + C_kP_kC_k^T)^{-1}(\hat{z}_k - C_k\hat{x}_k) \quad (5-9)$$
\[ P_{k+1} = A_k (P_k - P_k C_k^T (R + C_k P_k C_k^T)^{-1} C_k P_k) A_k^T + Q_k \]  

(5-10)

Where \( \hat{z}_k \) is the observed value of \( z_k \).

In a semiconductor manufacturing processes, the process output is

\[ y_k = \beta u_k + z_k \]  

(5-11)

where \( \beta \) is the actual process gain, and \( u_k \) is the manipulated variable. The nominal process gain used in the controller system is \( b \). Then

\[ \hat{z}_k = y_k - b u_k \]  

(5-12)

The process input \( u_k \) can be calculated from:

\[ u_k = \frac{1}{b} (r_k - C_k \hat{x}_k) \]  

(5-13)

As a result

\[ \hat{z}_k - C_k \hat{x}_k = y_k - b u_k - C_k \hat{x}_k = y_k - r_k \]  

(5-14)

As the matrix \( \text{diag}(H_k) \) is not full rank matrix, the system is not observable, thus it is difficult to obtain unbiased estimates of every single state variable. However, it’s still possible to obtain a very good estimate of the sum of all state variables of the contexts included in current run. When applying this method to semiconductor processes, suboptimal control performance may result from the following cases.

(1) the observed variable may be not precise as a result of model plant mismatch;

(2) the elements of matrices \( R \) and \( Q \) may be not precise;

(3) the estimation of parameter \( \theta \) may not correct.

In the following section, simulations are carried out to investigate the influence of those problems, and the CPA II method will be applied to every state and the overall process output.

5.2 General Idea about Non-threaded Control

The key idea of CPA II is to assume there is no plant model mismatch, then optimal estimation of the state variable can lead to optimal control performance. In
EWMA controllers, because the optimal controller tuning factors can be derived, Q filter used in EWMA controller is designed based on the optimal $\omega$.

In the non-threaded control system derived by Wang[43], the optimal control condition is precise estimation of parameters, $R$ and $Q$.

According to the theory of Kalman filter, the post estimation of state variable is:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + P_k C_k^T (R + C_k P_k C_k^T)^{-1} (\hat{z}_k - C_k \hat{x}_k)$$  \hspace{1cm} (5-15)

Then the estimated disturbance should be:

$$\hat{e}_k = \hat{x}_{k|k} - \hat{x}_{k|k-1} = P_k C_k^T (R + C_k P_k C_k^T)^{-1} (\hat{z}_k - C_k \hat{x}_k)$$  \hspace{1cm} (5-16)

When $R$, $Q$ and $\theta$ are precisely estimated, then the estimation will be optimal. Assume there is no plant model mismatch, then when the state estimation is optimal, the control performance is also optimal.

When metrology delay and set point change are not considered, only the stochastic error will be included in the process output error, and the optimal Kalman filter will have the minimum mean square error with the estimated error $\hat{e}_k = \hat{x}_{k|k} - \hat{x}_{k|k-1}$ as white noise.

Because $\hat{z}_k - C_k \hat{x}_k = y_k - r_k$ is the process output error, then the filter of CPA II method in this case should be $P_k C_k^T (R + C_k P_k C_k^T)^{-1}$. Theoretically, if the estimated disturbance is white noise, then the corresponding process state is well estimated. However, the control performance is not related to a single process state but the combination of all states. As a result, the combined estimated disturbance should also be investigated. In the case without metrology delay or set point change, this combined estimated disturbance is just the same as the process output error. When the control performance is optimal, the process output error should be a sum of some independent white noises, which is also white noise.
5.3 Simulation of CPA II for Non-threaded Control

In this section, simulations with non-threaded control method were made and then CPA II method was applied to both single state variables and the overall process output error[43]. The process context consists of three different categories, \{tool, layer, product\}, and \(n_{tool} = 2, n_{layer} = 2, n_{product} = 3\). Here, we use T1, T2 to denote tools, L1, L2 to denote the two layers, and P1, P2, P3 to denote three products. The context realization was selected randomly based on given probability density, and the frequency of different context items are shown in Table 5-1. The total number of runs in the simulation is \(N=1000\). The variances and IMA model parameters of the measurement noise is 0.04, and the variance of every context item are also shown in Table 5-1. In simulations 5-1 to 5-4, before \(k = 500\), the value of \(R\) used in non-threaded control system is 0.04, and the value of \(\theta\) and \(Q\) shown in Eqs. (5-9) and (5-10) are exactly the same as the simulation set up. As shown in Wang’s work[43], the control performance is optimal in this range. After \(k = 500\), disturbances were introduced to different parameters in different simulations, and as a result, the control performance should be suboptimal. Details of the parameters change are shown in Table 5-2.

Table 5-1 Parameters set up in Simulation 5-1, 2, 3, 4

<table>
<thead>
<tr>
<th>Context</th>
<th>Frequency</th>
<th>Variance</th>
<th>(\theta)</th>
<th>Initial value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.5</td>
<td>1.6e-3</td>
<td>0.75</td>
<td>0.2</td>
</tr>
<tr>
<td>T2</td>
<td>0.5</td>
<td>1.6e-3</td>
<td>0.85</td>
<td>0.5</td>
</tr>
<tr>
<td>L1</td>
<td>0.5</td>
<td>2.5e-3</td>
<td>0.60</td>
<td>1.1</td>
</tr>
<tr>
<td>L2</td>
<td>0.5</td>
<td>2.5e-3</td>
<td>0.50</td>
<td>0.7</td>
</tr>
<tr>
<td>P1</td>
<td>0.33</td>
<td>0.04</td>
<td>0.70</td>
<td>2.3</td>
</tr>
<tr>
<td>P2</td>
<td>0.33</td>
<td>0.04</td>
<td>0.80</td>
<td>1.2</td>
</tr>
<tr>
<td>P3</td>
<td>0.33</td>
<td>0.04</td>
<td>0.90</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 5-2 Parameter change after \(k=500\) in simulations 5-1 to 5-4

<table>
<thead>
<tr>
<th>Simulation</th>
<th>(1 \leq k \leq 1000)</th>
<th>(\beta)</th>
<th>(Q_k)</th>
<th>(\theta)</th>
<th>(\theta)</th>
<th>(R)</th>
<th>(\text{var}(y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 5-1</td>
<td>501 \leq k \leq 1000</td>
<td>0.05</td>
<td>precise</td>
<td>precise</td>
<td>0.04</td>
<td></td>
<td>0.1216</td>
</tr>
<tr>
<td>Simulation 5-2</td>
<td>501 \leq k \leq 1000</td>
<td>0.05</td>
<td>0.7</td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.1590</td>
</tr>
<tr>
<td>Simulation 5-3</td>
<td>501 \leq k \leq 1000</td>
<td>0.05</td>
<td>(I\times0.1)</td>
<td>0.04</td>
<td>0.1</td>
<td>0.0965</td>
<td></td>
</tr>
<tr>
<td>Simulation 5-4</td>
<td>501 \leq k \leq 1000</td>
<td>0.05</td>
<td>0.1</td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.0925</td>
</tr>
</tbody>
</table>

92
5.3.1 Plant Model Mismatch

As shown in Table 5-2, in this simulation, a step change was introduced to the process gain at $k=501$, and as a result plant model mismatch occurred. The process output error and results of CPA II analysis of the overall process are shown in Fig.5-1. Table 5-2 shows that the variance of the process output is somewhat larger than before. It’s difficult to tell whether the control performance is optimal or not from the CPA II method, but an apparent change can be detected in Fig. 5-1(b), which indicates a process excursion at $k = 500$. Because the process output error is a combination of different noises rather than one single white noise, it is not valid to assume -0.1 to 0.1 as optimal control range in the CPA II analysis result. Variance information shown in Table 5-2 indicates that when plant model mismatch happens, obvious degradation of control performance may happen and can be detected by both CPA II analysis and variance change.

![Figure 5-1 Process output and CPA II analysis of overall process (Simulation 5-1)](image)

The actual and estimated values of every state variable are shown in Fig. 5-2, where the discrete points denote the real values and the solid line is the estimated value. The corresponding CPA II analysis results are also shown in this figure. From Fig. 5-2, it can be found after the change at $k=501$, the estimation results of some state variables ($x_{l1}$, and $x_{l2}$) are not as precise as before. However, the process output error results shown in Fig. 5-1 indicate that the overall estimation and control in this range was still satisfying. In Fig. 5-2, the CPA II analysis of the state $x_{p3}$ shows a very obvious change,
thus the CPA II result applied to single state variables is not a valid standard to assess whether the control performance is good or not. However, it is possible to provide some process monitoring information about whether there some change has occurred.

Figure 5-2 Actual value, estimated value and CPA II analysis result of every single state variable (Simulation 5-1)
5.3.2 Improper Selection of $\theta$

In Eq.(5-5), matrix $A_k$ includes the IMA model parameter $\theta$, which is in fact unknown and difficult to estimate precisely in industrial practice. Theoretically, optimal control performance requires that the value of $\theta$ must be precisely estimated. However, research of Wang[43] has shown that when improper value of $\theta$ was selected in $A_k$, the state estimation result was not been obviously influenced. In Simulation 5-2, before $k=500$, precise values of $\theta$ were used to achieve optimal control, and after $k=501$, all values of $\theta_i$ for every context are selected as 0.7, which is not precise enough. The process output and CPA II analysis of overall output are shown in Fig. 5-3, and it can be found that as reported Wang[43], no obvious control performance degradation can be detected. The variance shown in Table 5-2 also proved that the control performance is still satisfying and in fact even better in this case. The result of variance analysis yields the same conclusions as the CPA II analysis result.

![Figure 5-3 Process output and CPA II analysis of overall process (Simulation 5-2)]
The actual and estimated values of every single state variable are also shown in Fig. 5-4 together with the CPA II analysis result of every state variable. It is interesting that the variables with estimated $\theta$ far away from the real $\theta$ did not show obvious degradation of state estimation performance. For example, for the context $p_3$, the real $\theta$ is
0.9 and for $p_1$, the real $\theta$ is 0.7. When the value of $\theta$ in $A_k$ is selected as 0.7 for all state variables, no obvious difference of state estimation performance can be detected between the two variables. This is because in the Kalman-IMA method, the estimation of different variables is highly corrected and can influence every other variable in the subsequent estimation steps with Equations. (5-9) and (5-10). In the CPA II analysis result of $p_3$, some change can be detected but this change was not observed in other state variables. In conclusion, CPA II analysis of every single state variable is not a valid method for the control or state estimation performance assessment, but sometimes, a specific variable may be sensitive to the change of the process which cannot be detected by the CPA II analysis of the overall output. The CPA II analysis of the overall output is still beneficial to assess the control performance in this Kalman-IMA method. However, as this method is so efficient and robust in the case without metrology delay, no obvious performance degradation happens in the case with plant model mismatch or improper selection of $\theta$.

### 5.3.3 Improper Selection of R or Q

In Equations (5-9) and (5-10), in order to achieve optimal state estimation, the value of $R$ and $Q_k$ should be selected precisely. However, it is difficult to obtain precise values and usually only an estimate will be employed. In Simulation 5-3, after $k = 501$, the value of $R$ is changed from the precise value 0.04 to 0.1, which is similar to the process output variance. In Simulation 5-4, after $k = 501$, the value of $Q_k$ was selected as $I \times 0.1$, which was also related to process output variance but not the precise value. The process output and the CPA II analysis of both simulations are shown in Fig. 5-5 and 5-6 respectively.

Results shown in Figs. 5-5 and 5-6 show the robustness of the Kalman-IMA method, and both process variance and CPA II analysis showed that even if the value of $R$ and $Q_k$ are not estimated precisely, the control performance is still good and could be
considered as optimal. No obvious change of overall process control performance can be detected from CPA II analysis result.

![Figure 5-5](image_url)

**Figure 5-5.** Process output and CPA II analysis of overall process (Simulation 5-3)

![Figure 5-6](image_url)

**Figure 5-6.** Process output and CPA II analysis of overall process (Simulation 5-4)

### 5.4 SUMMARY

In this chapter, a new non-threaded control algorithm developed by Wang[43] was introduced and CPA II method was applied to this control system. Simulations show that when metrology delay and set point change are not considered, the Kalman-IMA method did performed well in state estimation, and the process control performance is not very sensitive to the selection of controller factors, such as $\theta$, $R$ and $Q_k$. Simulations also show that CPA II analysis of single state variables can detect the process and control performance change even if the process output did not have obvious degradation.
Simulations in this chapter did not show obvious control performance degradation because Wang’s work only considered the case without metrology delay. In EWMA controller system, when metrology delay is considered, improper selection of controller parameter can cause unstable performance. As analyzed in previous chapters, set point change will also contribute significantly to the process output error as part of the deterministic error. When metrology delay and set point change are considered in future research for non-threaded control of high-mix processes, more suboptimal behavior may occur and CPA II method may be able to provide additional contributions to the process monitoring and control performance assessment/diagnosis.
CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 SUMMARY OF RESEARCH

The error source of the Run-to-run EWMA control system is analyzed, and a Control Performance Assessment method (CPA I) is generated based on the error source analysis. In CPA I, ARMAX regression is directly applied to the process output error, and the performance index is defined based on the variance of the regression results. CPA I method has been demonstrated that it can provide information about control performance degradation in both threaded and non-threaded EWMA controllers, and CPA I method is also possible to diagnose the plant model mismatch in the cases with multiple set points.

The study of CPA method (I) indicated that it is possible to investigate the control performance from ARMAX regression of process output error. It is not only possible to determine whether the control performance is good or not, but also possible to offer diagnosis information of suboptimal behavior.

6.1.2 CPA (II) based on analysis of estimated disturbance

In the CPA (II) method, the error source of run-to-run EWMA controller system is reconsidered as a combination of observer error and plant model mismatch error. In the case when there is no set point change, a time series called “estimated disturbance” is generated from the Q filter, and this disturbance signal is considered as the observer error. When the estimated disturbance is white noise, the control performance is optimal.

The CPA II method is developed on the ARMAX regression result of the estimated disturbance. A recursive regression method is applied for this estimation and simulations shown that this new CPA II method is sensitive to the performance degradation caused by improper controller tuning or model plant mismatch in the case
without set point change. Simulations also showed that the CPA II method works very well in the case with metrology delay and high-mix process parameters.

Industry data from industrial high-mix semiconductor manufacturing processes (Texas Instruments) are analyzed by CPA II method. Preliminary results showed that CPA II method is available to detect the control performance change in a threaded EWMA control system. However, the data analysis result also revealed a problem that in real cases, the optimal range of $M$ may be not from -0.1 to 0.1.

The influence of metrology delay and plant model match to the optimal range of $M$ has been discussed. Simulation results showed that inaccurate estimation of metrology delay will make CPA II method unreliable. When applying CPA II method to real industrial data, it is important to record the exact value of metrology delay for every run. Simulations also showed that even if the metrology delay varied from run to run, CPA II method is still available to be applied and the change of control performance can still be detected from the figure of $M$.

Research in this dissertation also showed that when the model-plant mismatch cannot be ignored, the optimal range of $M$ will be changed. In order to determine the baseline of “optimal control”, a method named CPO (Control Performance Optimization) was developed in this chapter. Simulations showed that this CPO method is available to investigate whether the current control performance is good or not even if there is model plant mismatch. The CPO method can also calculate the best achievable minimum variance and give suggestions about how to improve the control performance.

Theoretically, CPA II + CPO method is a good solution to the control performance assessment and diagnosis problem of threaded EWMA controller in semiconductor manufacturing processes. In order to apply this method to industrial processes, what should be done by practical control engineer is: (1), record the metrology delay for every run; (2), try to determine the baseline of “optimal control performance” by CPO method; (3), calculate the value of $M$ with CPA II method. With this method, the
control performance change caused by stochastic disturbance can be detected and a solution can be suggested.

A new non-threaded control algorithm developed by Wang[43] was introduced and CPA II method was applied to this control system. Simulations show that when metrology delay and set point change are not considered, the Kalman-IMA method did performed well in state estimation, and the process control performance is not very sensitive to the selection of controller factors, such as $\theta$, $R$ and $Q_k$. Simulations also show that CPA II analysis of single state variables can detect the process and control performance change even if the process output did not have obvious degradation.

Simulations did not show obvious control performance degradation because Wang’s work only considered the case without metrology delay. In EWMA controller system, when metrology delay is considered, improper selection of controller parameter can cause unstable performance. When metrology delay and set point change are considered in future research for non-threaded control of high-mix processes, more suboptimal behavior may occur and CPA II method may be able to provide additional contributions to the process monitoring and control performance assessment/diagnosis.

6.2 **RECOMMENDATIONS TO FUTURE WORK**

6.2.1 **Estimated Disturbance Computation**

In this dissertation, the key points of CPA II method are: 1) compute the estimated disturbance; 2) check the difference between estimated disturbance and white noise.

ARMAX regression did a very good job in checking whether the estimated disturbance is white noise or not, because it can also provide some diagnosing information besides the assessment results.
The estimated disturbance in this dissertation is computed by a filter, which is designed based on information of process model and metrology delay. However, the real industry will be more complicated and as a result, it may be difficult to achieve the theoretical model as shown in Chapter 3.

As the key idea of the computation of estimated disturbance is try to figure out the difference between the real and estimated values of some specific state variables or parameters, it is also possible to apply other statistical methods to carry out this work. In recent work of Sun and Qin[115], which is about the control performance of model predictive control, the process disturbance is estimated from the residual of an ARMAX regression. As the process output error of EWMA run-to-run control can also be expressed by ARMAX regression, it is also possible to calculate the process disturbance and the estimated disturbance with similar methods.

In the future, non-threaded controller based on state estimation may be more and more popular, and then it will be also an interesting topic to investigate the estimation performance which is closely related to the control performance.

6.2.2 Online Adaptive Tuning

The method of tuning, or determination of the suggested value controller factor, has been presented in this work as an off-line process (CPO in Chapter 4). However, via the CPO method in Chapter 4, it is possible to achieve an online tuning suggestion for the controller factor, and if an online optimization of the controller tuning factor is implemented, it is quite possible to reduce the process excursion.

In current semiconductor industry, such “dynamic” tuning factor is also applied in some processes, but the practical experience showed that sometimes the adjustment of the controller maybe not such helpful because real industry is more complicated than simulation, and the change in “dynamic” controller factor may provide no obvious benefits but a lot of trouble in both process control and monitoring.
As a result, the online application of the CPO method still need more research, and the trade in between higher controller performance and simplifier manipulation should be considered.

6.2.3 Control Performance Diagnosis

An obvious drawback of the work in this dissertation is: although the change of degradation of control performance can be detected with CPA II and a new controller tuning factor can be suggested with CPO, it is still not so clear what the exact cause of the control performance change. All the work in this dissertation is focused on the mathematical work related to the controller, but the process may provide more useful information. Although the process output cannot be achieved until the end of a run, there are also some other variables that can be measured online, such as temperature, pressure, and so on. Those online measurement values are not directly included in the controller algorithm; however, they can affect some parameters, such as the process gain. Therefore, control performance diagnosis should include some information from the process variables which may not be included in the controller scheme.
Appendices
APPENDIX A

Optimal EWMA Tuning Factor with Metrology Delay without Set Point Change

When there is no model plant mismatch, no set point change, and the effective metrology delay is non-zero, the output error is

\[ e_k = \frac{1 - \theta q^{-1}}{1 - q^{-1} + \omega q^{-1-d}} q^{-d} \varepsilon_k \]  \hspace{1cm} (2-14)

Similar to the work of Wang [116], we can also calculate the variance of \( e_k \), and then find the optimal value of \( \omega \) that minimizes the variance of \( e_k \). However, due to the term \( \omega q^{-1-d} \), it is not so easy to calculate the variance of \( e_k \), so an alternative method is explained below.

As indicated before, when there is no model-plant mismatch, the best state estimation of \( a_k \) leads to the best control performance. When metrology delay is not zero

\[ a_k = \hat{a}_{k|k-1-d} \]  \hspace{1cm} (A-1)

The minimum mean square error (MMSE) forecast of IMA(1,1) series with lead time \( l=d+1 \) can be calculated by the method provided in Castillo[117]. Then the MMSE forecast is

\[ \hat{a}_{k|k-1-d} = (1 - \theta) a_{k-1-d} + \theta \hat{a}_{k-1|k-2-d} \]  \hspace{1cm} (A-2)

When there is no model-plant mismatch,

\[ a_{k-1-d} = y_{k-1} - bu_{k-1-d} \]  \hspace{1cm} (A-3)

With this optimal forecast, the best achievable performance is

\[ \min(\text{var}(\varepsilon)) = \min(\text{var}(a - \alpha)) = (1 + d \cdot (1 - \theta)^2) \text{var}(\varepsilon) \]  \hspace{1cm} (A-4)
In the current EWMA controller,

\[ \hat{a}_{k|k-1-d} = \omega (y_{k-1} - b u_{k-1}) + (1 - \omega) \hat{a}_{k-1|k-2-d} \]  

(A-5)

Assume all the estimation and process data before the \( k^{th} \) run are optimal. In order to make the \( k^{th} \) run also optimal, we need to make

\[ \omega (y_{k-1} - b u_{k-1}) + (1 - \omega) \hat{a}_{k-1|k-2-d} \]

\[ = (1 - \theta)(y_{k-1} - b u_{k-1-d}) + \theta \hat{a}_{k-1|k-2-d} \]  

(A-6)

Before the \( k^{th} \) run, the control performance is optimal. If the set point to kept at 0, then we have

\[ \hat{a}_{k-1|k-2-d} = a_{k-1} = \frac{1 - \theta}{1 - q^{-1}} \epsilon_{k-2-d} \]  

(A-7)

\[ b u_{k-1} = -a_{k-1} = \frac{1 - \theta}{1 - q^{-1}} \epsilon_{k-2-d} \]  

(A-8)

\[ b u_{k-1-d} = -a_{k-1} - d = \frac{1 - \theta}{1 - q^{-1}} \epsilon_{k-2-2d} \]  

(A-9)

\[ y_{k-1} = \frac{1 - \theta q^{-1} - q^{-1-d} + \theta q^{-1-d}}{1 - q^{-1}} \epsilon_{d-1} \]  

(A-10)

Substitute Equations (A-7 to A-10) to Eq. (A-6), then

\[ [\omega(1 - \theta q^{-1} - q^{-1-d} + \theta q^{-1-d})] \epsilon_{k-d-1} = [(1 - \theta)(1 - q^{-1})] \epsilon_{k-d-1} \]  

(A-11)

Eq. (A-11) can be rewritten as

\[ \left[ \omega\left(1 - q^{-1} + (1 - \theta)q^{-1}(1 - q^{-d})\right) \right] = [(1 - \theta)(1 - q^{-1})] \]  

(A-12)

That means

\[ \{\omega(1 - q^{-1})[1 + (1 - \theta)(1 + q^{-1} + q^{-2} + \cdots + q^{-d})]\} = [(1 - \theta)(1 - q^{-1})] \]  

(A-13)
as a result, the optimal $\omega$ can be calculated as

$$\omega_{opt} = \frac{1 - \theta}{1 + (1 - \theta)d}$$  \hspace{1cm} (A-14)

If there is model-plant mismatch, the optimal $\omega$ should be

$$\omega_{opt} = \frac{1 - \theta}{1 + (1 - \theta)d} \frac{b}{\beta}$$  \hspace{1cm} (2-17)
APPENDIX B

ARMAX Model Used in MATLAB

The ARMAX model structure is

\[ y(t) + a_1 y(t-1) + \ldots + a_{n_a} y(t-n_a) = b_1 u(t-n_k) + \ldots + b_{n_b} y(t-n_k-n_b+1) + c \varepsilon(t) + \ldots + c_{n_c} \varepsilon(t-n_c) \]  

(A-1)

A more compact way to write the difference equation is

\[ A(q) y(t) = B(q) u(t-n_k) + C(q) \varepsilon(t) \]  

(B-1)

where

- \( y(t) \) — Output at time \( t \).
- \( n_a \) — Number of poles.
- \( n_b \) — Number of zeroes plus 1.
- \( n_c \) — Number of \( C \) coefficients.
- \( n_k \) — Number of input samples that occur before the input affects the output, also called the dead time in the system. For discrete systems with no dead time, there is a minimum 1-sample delay because the output depends on the previous input and \( n_k = 1 \).

- \( y(t-1) \ldots y(t-n_a) \) — Previous outputs on which the current output depends.
- \( u(t-n_k) \ldots u(t-n_k-n_b+1) \) — Previous and delayed inputs on which the current output depends.
- \( \varepsilon(t-1) \ldots \varepsilon(t-n_c) \) — White-noise disturbance value.

The parameters \( n_a, n_b, \) and \( n_c \) are the orders of the ARMAX model, and \( n_k \) is the delay. \( q \) is the delay operator. Specifically,

\[ A(q) = 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a} \]  

(B-3)

\[ B(q) = b_1 + b_2 q^{-1} + \ldots + b_{n_b} q^{-n_b+1} \]  

(B-4)

\[ C(q) = 1 + c_1 q^{-1} + \ldots + c_{n_c} q^{-n_c} \]  

(B-5)

If the data is a time series, which has no input channels and one output channel, then the MATLAB function “armax” calculates an ARMAX model for the time series

\[ A(q) y(t) = C(q) \varepsilon(t) \]  

(B-6)
BIBLIOGRAPHY


8. Adivikolanu, S. and Zafiriou, E., *Extensions and performance/robustness tradeoffs of the EWMA run-to-run controller by using the Internal Model Control


17. Butler, S.W., *Texas Instruments' overview of control needs for manufacturing in the 90's*, in *AEC/APC Workshop V, SEMATECH1993*


25. Campbell, W.J.; Model predictive run-to-run control of chemical mechanical planarization; PhD; The University of Texas at Austin; 1999.


66. Ettaleb, L.; Control loop performance assessment and oscillation detection; PhD; University of British Columbia; 1999.


102. Prabhu, A.V.; Performance monitoring of run-to-run control systems used in semiconductor manufacturing; Doctoral; The University of Texas at Austin; 2008.


114. Stuber, J.D. *Multiloop photolithography control using hierarchical context information for APC models*. 2003. SPIE.


VITA

Xiaojing Jiang was born in Jinan, Shandong Province, China in 1986. After graduation from Zhangqiu No. 1 High School, Shandong, China in 2002, Xiaojing entered Zhejiang University in Hangzhou City, Zhejiang, China. She received the degree of Bachelor of Science in Chemical Engineering from the Zhejiang University in June of 2006. She also received a second degree of the Chu Kochen Honors Program in Zhejiang University, 2006. Xiaojing graduated in June of 2008 with a Master of Science Degree in Chemical Engineering in Zhejiang University. She entered graduate school at the University of Texas at Austin in August, 2008 and worked as a graduate research assistant. During her study at the University of Texas at Austin, she also served as a Teaching Assistant in 2010 and the spring of 2012.

Permanent address: lvxueyimei@gmail.com

This dissertation was typed by the author.