AUTHOR QUERIES

Journal id: TPRS_A_303131

Corresponding author: Shi-Shang Jang

Title: Performance assessment of run-to-run control in semiconductor manufacturing based on IMC framework

Dear Author
Please address all the numbered queries on this page which are clearly identified on the proof for your convenience.

Thank you for your cooperation

Query number	Query
1	Please supply names of all authors in reference Edgar et al. 2000 & Firth et al. 2006.
2	Please supply journal title in full for Harris 1999.
3	Please supply names of all authors plus update of publication details for reference Ma et al. 2007.
4	Please supply names of all authors for reference Moyne et al. 2001.
5	Please supply date of symposium plus place of publication and name of publisher for reference Prabhu et al. 2006.
6	Please supply journal title in full for reference Qin 1998.
7	Please supply names of all authors plus journal title in full for reference Qin et al. 2006
8	Please supply journal title in full for reference Stanfelj et al. 1993.
9	Please supply names of all authors for the following reference: Tseng et al. 2003; Wang et al. 2005; Zheng et al. 2006;

International Journal of Production Research
Vol. ??, No. ?, Month?? 2008, 1–27

Taylor & Francis
Taylor & Francis Group

Performance assessment of run-to-run control in semiconductor manufacturing based on IMC framework

Liang Chen[a], Mingda Ma[b], Shi-Shang Jang[c]*,
David Shan-Hill Wang[c] and Shuqing Wang[a]

[a]*State Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, China;* [b]*Center for Control and Guidance Technology, Harbin Institute of Technology, Harbin 150001, China;* [c]*Department of Chemical Engineering, National Tsing-Hua University, Hsin-Chu 30043, Taiwan*

(*Received 6 November 2007; final version received 22 February 2008*)

The objective of this paper is to propose a universal methodology for performance assessment of run-to-run control in semiconductor manufacturing. The slope of the linear semiconductor process model is assumed to be known or subjected to mild plant/model mismatch. Based on an internal model control framework, analytical expressions of minimum variance performance (MVP) and best achievable performance (BAP) for a series of run-to-run control schemes are derived. In the methodology, closed-loop identification is utilised as the first step to estimate the noise dynamics via routine operating data, and numerical optimisation is employed as a second step to calculate the best achievable performance bounds of the run-to-run control loops. The validity of the methodology is justified by examples of performance assessment for EWMA control, double EWMA control and RLS-LT control, even under circumstances where the processes encounter model mismatch, metrology delay and more sophisticated noises. Several essential characteristics of run-to-run control are discovered by performance assessment, and valuable advice is offered to process engineers for improving the run-to-run control performance. Furthermore, a useful application example for online performance monitoring and optimal tuning of run-to-run controller demonstrates the advantage of the methodology.

Keywords: performance assessment; run-to-run control; minimum variance performance; best achievable performance; IMC

1. Introduction

Moore's law is the Bible of semiconductor industry. During the last two decades, sustained technology transitions have been made to keep pace with Moore's law. However, with a shrinking feature sise (0.45 μm or smaller) and an enlarging wafer diameter (300 mm or bigger), the process engineers nowadays are suffering great challenges due to the increasing complexity of semiconductor manufacturing. This makes an emergent appeal to advanced process control (Edgar *et al.* 2000, Qin *et al.* 2006), especially run-to-run (RtR) control (Castillo 1997, Moyne *et al.* 2001).

*Corresponding author. Email: ssjang@mx.nthu.edu.tw

ISSN 0020–7543 print/ISSN 1366–588X online
© 2008 Taylor & Francis
DOI: 10.1080/00207540802029633
http://www.informaworld.com

RtR control, the name given by the semiconductor industry, is a combination of statistical process control (SPC) and engineering process control (EPC) (Sachs *et al.* 1995). Over the last 10 years, a substantial growth of literature exists on various approaches to semiconductor RtR control, including exponentially weighted moving average (EWMA) controller (Ingolfsson and Sachs 1993, Patel and Jenkins 2000, Tseng *et al.* 2003), double EWMA (dEWMA) controller (Castillo 1999, Chen and Guo 2001; or PCC controller, Bulter and Stefani 1994), RLS-LT controller (Wang *et al.* 2005), RtR-IMC controller (Adivikolanu and Zafiriou 2000), RtR-JADE controller (Firth *et al.* 2006), RtR-ANOVA controller (Ma *et al.* 2007) and RtR-MPC (Bode *et al.* 2004). However, most of the above controllers are expensive to implement and maintain; requiring hardware, software, and engineering services to generate sustained benefits. Actually, few RtR controllers in a fab are running at their best performance. This may be due to either incorrect use of controller structure, improper tuning of controller parameters, non-stationary process disturbance, or the mixed product nature of semiconductor manufacturing (Zheng *et al.* 2006). Furthermore, many semiconductor processes exhibit tool ageing problems. This effect of equipment ageing introduces deterministic drift disturbance, and leads to more severe process variations. Preventive maintenance (PM) is needed to keep the product quality on target. However, PM meanwhile results in an additional shift, which further degrades system performance. Six Sigma criteria are basically implemented to evaluate the product quality in semiconductor industry. Some other statistics such as Cpk, that includes the effects of both mean and variance, are also frequently referred to. These criteria, however, do not give any insight into the process nature, and fail to tell whether the system is running at its optimal performance. Therefore, there is a need of monitoring techniques to assess the RtR control performance and identify the underlying causes of the poor performance via closed loop data, which is called performance assessment in the control engineering literature.

Research on performance assessment has received increasing attention since the original work of Harris (1989). Elegant reviews by Qin (1998) and Harris (1999) are available. Harris proposed the use of routine operating data to evaluate the minimum variance performance (MVP) of control loops. MVP represents a lowest bound on the variance of the system when minimum variance control (MVC) is implemented. There are also many related research works, including assessment of single loop feedback and feed-forward control (Stanfelj *et al.* 1993, Desborough and Harris 1993), assessment of cascade control (Ko and Edgar 2000), and assessment of multivariable control (Harris *et al.* 1996, Huang *et al.* 1997, Ko and Edgar 2001a). Moreover, in case MVC cannot be achieved, extensions of the performance assessment to more realistic control are also studied, including assessment of PID-achievable performance (Ko and Edgar 2004), and MPC-achievable performance (Ko and Edgar 2001b).

To our knowledge, there is rather limited research work on performance assessment of semiconductor process control. Prabhu *et al.* (2006) derived the best achieve performance (BAP) of EWMA controller based on their previous contribution of PID-achievable performance assessment (Ko and Edgar 2004). The EWMA controller is first transformed into a discrete integral controller (Sachs *et al.* 1995, Castillo 2001). An iterative solution is then utilised to calculate the performance index via closed loop data. However, the above approach is only suitable for EWMA controller since other RtR controllers are not necessarily of PID-type.

In this paper a more general methodology is proposed to assess the performance of a series of RtR controllers (EWMA, dEWMA, RLS-LT, etc.) based on internal

model control (IMC) structure, since most of the RtR controllers are designed using the IMC framework (Bulter and Stefani 1994, Adivikolanu and Zafiriou 2000, Castillo 2001), or can be represented as IMC controllers (Tseng *et al.* 2003, Wang *et al.* 2005). In this work, it is our purpose to assess the RtR control performance using minimum process information (routine operating data only) and minimum prior knowledge (controller structure only). The following questions are answered:

(1) How to derive the MVP bound of the RtR control system?
(2) How to calculate the BAP bound of a RtR control loop?
(3) Whether the current system performance is good enough?
(4) If not, how to improve the performance? By re-tuning the RtR controller, or designing an advanced RtR controller, or at the worst case modifying the process to reduce the disturbance and the time delay?

In this work, the process model and the IMC representation of RtR control systems are reviewed firstly in Section 2. Based on the IMC framework, the MVP bound can be found using classical theory of performance assessment. The BAP bounds for EWMA and double EWMA controllers are derived analytically in Section 3. Closed loop identification and numerical optimisation are used to solve the rest problem. In Section 4, realistic examples of RtR control performance assessment and detailed simulation verification are presented. Essential characteristics of the RtR controllers are investigated by performance assessment. Controller improvement suggestions are offered to process engineers for improving the RtR control systems. Furthermore, extensions of online performance monitoring and optimal tuning of RtR controllers are given in Section 5. Finally, conclusive remarks are given in Section 6.

2. Background

The problem of performance assessment of RtR control is briefly as represented in Figure 1. The key issue of the above problem is to evaluate the MVP and BAP bounds of the specified RtR control system via closed loop data. In semiconductor manufacturing, the process output is usually described by a linear model.

$$Y_k = \alpha + \beta U_{k-1} + N_k \tag{1}$$

where α models any offset or bias from the target and β is the input-output gain, Y_k is the measured quality characteristic of run k, U_{k-1} is the manipulated variable of previous run, and N_k is the noise disturbance. As commonly used in time series analysis, IMA (1, 1) (first order integrated moving average process) is a useful representation of process disturbance in discrete manufacturing systems (Box and Jenkins 1994). However, in semiconductor manufacturing, many processes suffer from tool wear problems.

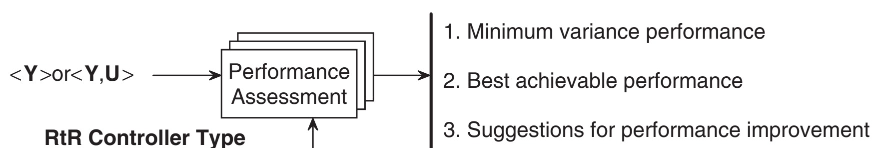

Figure 1. Problem formulation of RtR control performance assessment.

In time series terminology, tool wear represents noise dynamics, i.e. the noise contains an
125 additive drift disturbance. Hence, the following noise model (Castillo 2001, 2002) is adopted in our research.

$$N_k = N_{k-1} + \delta + (1 - \theta z^{-1})\varepsilon_k \quad 0 \le \theta \le 1 \tag{2}$$

where δ is the drift term, and $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$ is a white noise sequence. Note that Equation (2) is especially useful to represent the noise in semiconductor manufacturing as
130 discussed by Castillo (2002).

In semiconductor manufacturing, EWMA-based controllers are the most widely used RtR control schemes. Looking at EWMA-based controllers using IMC framework provides some further insight about how they work. Butler and Stefani (1994) use IMC structure to design their PCC controller, and Adivikolanu and Zafiriou (2000) carried out
135 the performance robustness trade-off research directly based on IMC framework. Most of the RtR control schemes can be understood as an IMC controller (Castillo 2002). The main diversity among these controllers is the filter used in the IMC structure. As shown in Figure 2, EWMA controller uses a simple filter $a_k = \lambda(Y_k - bU_{k-1}) + (1-\lambda)a_{k-1}$ to recursively update the output intercept, and dEWMA controllers use double EWMA
140 filters to compensate process drift, and other more advanced RtR controllers (e.g. RLS-LT controller and RtR-IMC controller) utilise more sophisticated filters, i.e. RLS filter (Wang 2005) and RtR-IMC filter (Adivikolanu 2000), to improve performance and robustness. These filters could be transformed into the EWMA-based formulation under certain assumptions (Wang *et al.* 2005). Hence, without loss of generality, our investigation is
145 carried out using the EWMA-based (EWMA and dEWMA) filters, since it can be easily extended to other filters using IMC framework.

For an EWMA control system shown in Figure 2, the characteristic equation is

$$1 + \frac{\lambda(\xi - 1)z^{-1}}{1 - (1 - \lambda)z^{-1}} = 0 \tag{3}$$

where $\xi = \beta/b$ represents the model mismatch between the real and the estimated process
150 gain. Solving Equation (3), the stability condition for EWMA controller is $|1 - \lambda\xi| < 1$(Ingolfsson and Sachs 1993). For dEWMA control system, Castillo proved that it is asymptomatically stable if and only if

$$\left|1 - 0.5\xi(\lambda_1 + \lambda_2) + 0.5\sqrt{\xi^2(\lambda_1 + \lambda_2)^2 - 4\lambda_1\lambda_2\xi}\right| < 1$$

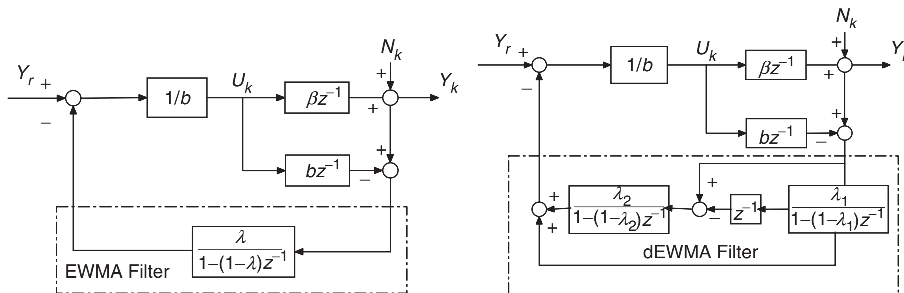

Figure 2. IMC representations of the EWMA and dEWMA controller.

and

$$\left|1 - 0.5\xi(\lambda_1 + \lambda_2) - 0.5\sqrt{\xi^2(\lambda_1 + \lambda_2)^2 - 4\lambda_1\lambda_2\xi}\right| < 1$$

155 (Castillo 1999, 2002). These stability conditions should be satisfied when assessing the RtR control performance.

In the rest of the paper, the topic of performance assessment of RtR control will be carried out step by step using the IMC framework. Since most of the semiconductor 160 manufacturing processes are described by a simple linear model, for simplicity, the input-output gain β is assumed to be accurately estimated by design of experimental (DOE) or regression analysis.

Lemma (IMC performance assessment): *Define* $G_p(z^{-1}) = \beta z^{-d}$ *as the real process model,* $G_u(z^{-1}) = bz^{-d}$ *as the offline estimated process model,* $G_f(z^{-1})$ *as the transfer function of* 165 *filter, and* $G_c(z^{-1})$ *is the inverse of the process model without time delay. For a given IMC system assuming no model mismatch* (*i.e.* $\beta = b$)

$$\begin{cases} Y_k = G_p(z^{-1})U_k + N_k \\ N_k = G_w(z^{-1})\varepsilon_k = \dfrac{B_w(z^{-1})}{A_w(z^{-1})}\varepsilon_k\,. \\ G_{\mathrm{IMC}}(z^{-1}) = G_f(z^{-1})G_c(z^{-1}) \end{cases} \tag{4}$$

The system output can be expressed as

$$Y_k = \underbrace{F(z^{-1})\varepsilon_k}_{\text{Uncontrollable}} + \underbrace{A_w^{-1}(z^{-1})(G(z^{-1}) - B_w(z^{-1})G_f(z^{-1}))\varepsilon_{k-d}}_{\text{Controllable}}. \tag{5}$$

170 **Proof:** See Appendix 1.

Note that this formulation is a direct and simple extension of Qin's work (1998), which acts as the first step for performance assessment of RtR controllers based on IMC framework. When there is model mismatch, the inaccurate estimated process gain will degrade system performance, and affect the results of performance assessment. This 175 proposition will be investigated in detail in Section 4.

Remark 1: Equation (5) gives an insight into the control system. Since the first term on the right-hand side depends on the data up to run k indicating that this term is uncontrollable due to the metrology delay, while the second term is controllable because it only relies on the data d runs ahead. By proper design of the IMC filter for the RtR 180 controller, i.e. $G_f(z^{-1}) = G(z^{-1})/B_w(z^{-1})$ (a minimum variance controller), the theoretical MVP bound can be achieved.

$$\begin{aligned} Y_k &= F(z^{-1})\varepsilon_k \\ \sigma^2_{\mathrm{MVP}} &= \mathrm{Var}(F\varepsilon_k) = E[(F\varepsilon_k)^2] = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2_\varepsilon. \end{aligned} \tag{6}$$

Unfortunately, most of the RtR control loops cannot achieve this theoretical bound due to the non-stationary noise dynamics and the manufacturing complexity. In practice, 185 the real system variance is much higher than the MVP bound.

$$\mathrm{Var}(Y_k) \geq \sigma^2_{\mathrm{MVP}} = \mathrm{Var}(F\varepsilon_k). \tag{7}$$

3. Theoretical developments

In this section, the theories for performance assessment of EWMA-based RtR control systems via closed loop data are developed. The BAP bounds are deduced step by step as explicit functions of noise characteristics and RtR controller settings. Detailed flow sheet of the methodology is depicted in the end of this section.

3.1 *Performance assessment for EWMA control*

EWMA controllers are the bread and butter of semiconductor process control. Hence, there's significant motivation to assess EWMA control loop to evaluate its best achievable performance.

Theorem 1: *Consider an EWMA controlled semiconductor manufacturing process with a nonstationary noise disturbance (IMA (1, 1) with drift).*

$$\begin{cases} Y_k = \alpha + \beta U_{k-1} + N_k \\ a_k = \lambda(Y_k - bU_{k-1}) + (1-\lambda)a_{k-1} \\ U_k = (Y_r - a_k)/b \\ N_k = N_{k-1} + \delta + (1 - \theta z^{-1})\varepsilon_k \end{cases}. \tag{8}$$

EWMA control performance can be evaluated by the long-run mean square error (MSE) of product quality.

$$MSE(Y_k) = \underbrace{(1 + f_1^2 + \cdots + f_{d-1}^2)\sigma_\varepsilon^2}_{\text{Minimum Variance Performance}} + \underbrace{\frac{(1-\lambda-\theta)^2}{1-(1-\lambda)^2}\sigma_\varepsilon^2 + \left(\frac{\delta}{\lambda}\right)^2}_{\text{Objective Performance}}. \tag{9}$$

Proof: See Appendix 2.

The first term on the right-hand side represents the MVP bound, and the second term, namely objective performance, should be minimised to calculate the BAP bound of the EWMA control.

Remark 2:

(a) If the process involves with a deterministic drift noise, the theoretical MVP bound can't be achieved by EWMA controller.
(b) For a white noise ($\delta = 0, \theta = 1$), no further control action is needed (i.e. $\lambda = 0$), a common consensus in literature (Box and Jenkins 1994, Castillo 2002).
(c) For an IMA (1, 1) noise without drift (i.e. $\delta = 0$), optimal EWMA setting should be $\lambda = 1 - \theta$ and this is consistent to the literatures (e.g., Ingolfsson and Sachs 1993, Moyne *et al.* 2001).
(d) Consider the noise $N_k = \delta k + \varepsilon_k (\theta = 1$, deterministic trend), $MSE(Y_k) = \sigma_\varepsilon^2 + \lambda/(2-\lambda)\sigma_\varepsilon^2 + (\delta/\lambda)^2$, a special case of Ingolfsson's work (1993).
(e) Consider the random walk with drift noise ($\theta = 0, \delta \neq 0$), $MSE(Y_k) = 1/\lambda(2-\lambda)\sigma_\varepsilon^2 + (\delta/\lambda)^2$, a special case of Castillo's work (1999).

3.2 *Performance assessment for double EWMA control*

Theorem 1 indicates that the EWMA controller is essentially not sufficient to control a worn out process (process subject to a drift drift), since it trends to be significantly off-target, and its performance degrades rapidly for a server drift noise. The dEWMA controller, however, accounts for a deterministic drift and provides offset-free control (Bulter 1994). This scheme is close to, although not equal to, a minimum variance controller.

Theorem 2: *Consider a semiconductor manufacturing process controlled by the double EWMA controller with a non-stationary noise disturbance (IMA (1, 1) with drift).*

$$\begin{cases} Y_k = \alpha + \beta U_{k-1} + N_k \\ a_k = \lambda_1(Y_k - bU_{k-1}) + (1-\lambda_1)a_{k-1} \\ p_k = \lambda_2(Y_k - bU_{k-1} - a_{k-1}) + (1-\lambda_2)p_{k-1} \\ U_k = (Y_r - a_k - p_k)/b \\ N_k = N_{k-1} + \delta + (1-\theta z^{-1})\varepsilon_k \end{cases} . \tag{10}$$

The output variance for double EWMA control is

$$\begin{aligned} \mathrm{Var}(Y_k) &= \left(1 + f_1^2 + \cdots + f_{d-1}^2\right)\sigma_\varepsilon^2 + \frac{(1-\Phi_2)(\Theta^2+1) - 2\Phi_1\Theta}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}(\Phi_1 - 1 - \theta)^2\sigma_\varepsilon^2 \\ &= \left(1 + f_1^2 + \cdots + f_{d-1}^2\right)\sigma_\varepsilon^2 + \left[\frac{-(2(\Theta^2+1) - 2\Phi_1\Theta/(\Phi_1-2)(\Phi_1+2))}{\Phi_2+1}\right. \\ &\quad \left. + \frac{-((\Theta+1)^2/2(\Phi_1+2))}{\Phi_2-\Phi_1-1} + \frac{((\Theta-1)^2/2(\Phi_1-2))}{\Phi_2+\Phi_1-1}\right](\Phi_1 - 1 - \theta)^2\sigma_\varepsilon^2. \end{aligned} \tag{11}$$

Double EWMA control performance can be expressed as the long-run mean square error of the system

$$MSE(Y_k)\begin{cases} = \underbrace{\sigma_{\mathrm{MVP}}^2}_{\mathrm{MVP}} + \underbrace{\frac{(1-\Phi_2)[(\Phi_2+\theta)^2 + (\Phi_1-\theta-1)^2] + 2\Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}\sigma_\varepsilon^2}_{\text{Objective Performance}} & \lambda_1, \lambda_2 \neq 0 \\ = \underbrace{\sigma_{\mathrm{MVP}}^2}_{\mathrm{MVP}} - \underbrace{\left\{\begin{array}{l}\left[2\dfrac{(\Phi_2+\theta)^2 + (\Phi_1-\theta-1)^2 + \Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(\Phi_2+1)(\Phi_1-2)(\Phi_1+2)}\right. \\ \left. + \dfrac{(\Phi_1-\Phi_2-2\theta-1)^2}{2(\Phi_1+2)(\Phi_2-\Phi_1-1)}\right]\sigma_\varepsilon^2 + \left(\dfrac{\delta}{\lambda_1+\lambda_2}\right)^2\end{array}\right\}}_{\text{Objective Performace}} & \lambda_2 \to 0 \end{cases} . \tag{12}$$

where

$$\begin{cases} \Phi_1 = 2 - \lambda_1 - \lambda_2 \\ \Phi_2 = -(\lambda_1 - 1)(\lambda_2 - 1) \\ \sigma_{\mathrm{MVP}}^2 = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma_\varepsilon^2 \end{cases} . \tag{13}$$

Proof: See Appendix 3.

Remark 3: Special attention should be paid to the variance expression of Equation (11).

(a) Using a partial fraction expansion, when $\lambda_2 \to 0$ $(\Theta \to 1, \Phi_1 \to (2-\lambda_1)$, and $\Phi_2 \to (\lambda_1 - 1))$,

$$\frac{(\Theta-1)^2}{2(\Phi_1-2)(\Phi_2+\Phi_1-1)} \to 0,$$

240 this high order infinitesimal term should be ignored. Castillo (1999) also addressed this feature in his research.

(b) When $\lambda_2 = 0$ (a special case of EWMA controller), the variance term equals to EWMA formulation

$$\mathrm{Var}(Y_k) = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma_\varepsilon^2 + \frac{(1-\lambda_1-\theta)^2}{1-(1-\lambda_1)^2}\sigma_\varepsilon^2.$$

245

Remark 4: Compared with Equation (9), it is intuitive from Equation (12) that the dEWMA controller provides better performance than EWMA controller, since it eliminates output offset term $(\delta/\lambda)^2$, which is the main factor for bad performance. It is suitable for semiconductor processes with tool wearing problems (Bulter and Stefani 250 1994). The optimal settings of dEWMA, however, are not so intuitive to derive compared with the EWMA controller.

3.3 *Performance assessment for RLS-LT control*

As stated in the former section, since most of the RtR control schemes are expressed by IMC framework, their performance can be assessed based on specified IMC filters. Here, 255 the RLS-LT control performance assessment is addressed for illustration. Consider the following RLS-LT RtR control system (Wang *et al.* 2005)

$$\begin{cases} x_k = y_k - bu_{k-1} \\ x_{k+i} = \sum_{j=0}^{n} \omega_j \frac{i^j}{j!} + \varepsilon_{k+i} \\ \tilde{\omega}_k = \tilde{\omega}_{k-1} + Q_k(x_k - \varphi_k^T \tilde{\omega}_{k-1}) \\ Q_k = P_{k-1}\varphi_k(\lambda_{\mathrm{Qin}} + \varphi_k^T P_{k-1}\varphi_k)^{-1} \\ P_k = (I - Q_k\varphi_k^T)P_{k-1}/\lambda_{\mathrm{Qin}} \end{cases} \tag{14}$$

where x_k is the observed variable (when no model mismatch, it represents the process noise); $\omega = [\omega_0\, \omega_1 \ldots \omega_n]^T$ is the model parameters of order n; λ_{Qin} is the forgetting factor 260 that gives more weight to more recent data; P_k, Q_k is the recursive parameters.

It has been proved (Wang *et al.* 2005) that the RLS-LT controller is equivalent to a dEWMA controller with $\lambda_1 = 1 - \lambda_{\mathrm{Qin}}^2$, $\lambda_2 = (1-\lambda_{\mathrm{Qin}})^2$ when sufficient observation is available. Thus we can evaluate its performance via the dEWMA IMC filer. System performance can be expressed by long run mean square error with

$$\begin{cases} \Phi_1 = 2\lambda_{\mathrm{Qin}} \\ \Phi_2 = \lambda_{\mathrm{Qin}}^3(\lambda_{\mathrm{Qin}} - 2) \end{cases}. \tag{15}$$

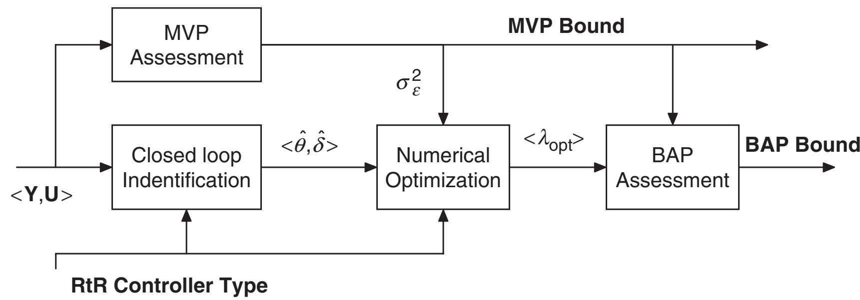

Figure 3. The flowsheet of RtR control performance assessment.

3.4 *General performance assessment*

In the previous section, the performances of EWMA, dEWMA and RLS-LT control are derived analytically as functions of process dynamics (i.e. $\langle\theta,\delta\rangle$) and controller settings (i.e. $\langle\lambda_1,\lambda_2\rangle$). In order to carry out performance assessment via routine operating data, closed loop identification is needed to estimate $\langle\hat{\theta},\hat{\delta}\rangle$ from the input-output data, numerical optimisation is also needed to minimise the objective performance in Equations (9) and (12). The integration of identification, optimisation and performance assessment is described in the flow sheet shown in Figure 3.

Below, we state the detailed methods used for closed loop identification and numerical optimisation in our approach.

Closed-loop identification. For an EWMA control system, the closed loop identification can be easily carried out using Equation (A14) and Equation (A15), $(1-\varphi z^{-1})(Y_k-\mu_Y)=(1-\theta z^{-1})\varepsilon_k$. For a set of given output data $\langle\mathbf{Y}\rangle$, $\mu_Y=\delta/\lambda$ can be estimated as output mean value, i.e. $\hat{\mu}_Y=\bar{Y}$. Then an ARMA (1, 1) model can be fit to the $(Y_k-\hat{\mu}_Y)$ time series to identify $\hat{\varphi}$ and $\hat{\theta}$. Note here $\varphi=1-\lambda$, hence $\hat{\delta}=\hat{\mu}_Y(1-\hat{\varphi})$.

For a double EWMA controller, the above method doesn't work because the output offset is completely eliminated by drift compensation of the second EWMA filter. Therefore, the output data alone don't provide any information about the noise dynamics. Additional input data are needed. Consider the input-output data set $\langle\mathbf{Y},\mathbf{U}\rangle$, it's easy to get the time series of $\langle\nabla\mathbf{Y},\nabla\mathbf{U}\rangle$, which provide useful information for closed loop identification. According to $\nabla Y_k-\delta=\beta\nabla U_{k-1}+(1-\theta z^{-1})\varepsilon_k$, δ can be estimated as $\hat{\delta}=\overline{\nabla Y}$, then an ARMAX (0, 1, 1) time series can be fit to model the data set of $(\nabla Y_k-\hat{\delta})$ to identify the parameters of $\hat{\theta}$ and $b=\hat{\beta}$.

Numerical optimisation. Any standard optimisation package can be used to solve the following problems.

For EWMA controller

$$\begin{aligned} \min f(\lambda) &= \frac{(1-\lambda-\theta)^2}{1-(1-\lambda)^2}\sigma_\varepsilon^2+\frac{\delta^2}{\lambda^2} \\ &s.t.\quad 0<\lambda<1 \\ &|1-\lambda\xi|<1 \text{ (Stability Condition)}, \end{aligned} \tag{16}$$

$$\mathrm{BAP}_{\mathrm{EWMA}}=\sigma_{\mathrm{MVP}}^2+f(\lambda_{\mathrm{opt}})=\sigma_\varepsilon^2+\frac{(1-\lambda_{\mathrm{opt}}-\theta)^2}{1-(1-\lambda_{\mathrm{opt}})^2}\sigma_\varepsilon^2+\frac{\delta^2}{\lambda_{\mathrm{opt}}^2}. \tag{17}$$

295 For double EWMA controller

$$\min f(\Phi_1,\Phi_2)\begin{cases} = \dfrac{(1-\Phi_2)[(\Phi_2+\theta)^2+(\Phi_1-\theta-1)^2]+2\Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}\sigma_\varepsilon^2, \lambda_1,\lambda_2\neq 0 \\ = -\Bigg[2\dfrac{(\Phi_2+\theta)^2+(\Phi_1-\theta-1)^2+\Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(\Phi_2+1)(\Phi_1-2)(\Phi_1+2)} \\ \quad +\dfrac{(\Phi_1-\Phi_2-2\theta-1)^2}{2(\Phi_1+2)(\Phi_2-\Phi_1-1)}\Bigg]\sigma_\varepsilon^2+\left(\dfrac{\delta}{\lambda_1+\lambda_2}\right)^2, \lambda_2\to 0 \end{cases}, \tag{18}$$

$$\text{s.t. } \Phi_1 = 2-\lambda_1-\lambda_2$$
$$\Phi_2 = -(\lambda_1-1)(\lambda_2-1)$$
$$0<\lambda_1<1,\ 0<\lambda_2<1$$
$$\begin{cases} \left|1-0.5\xi(\lambda_1+\lambda_2)+0.5\sqrt{\xi^2(\lambda_1+\lambda_2)^2-4\lambda_1\lambda_2\xi}\right|<1 \\ \left|1-0.5\xi(\lambda_1+\lambda_2)-0.5\sqrt{\xi^2(\lambda_1+\lambda_2)^2-4\lambda_1\lambda_2\xi}\right|<1 \end{cases} \text{(Stability Condition)}, \tag{19}$$

$$\text{BAP}_{\text{dEWMA}} = \sigma^2_{\text{MVP}} + f(\Phi_1^{\text{opt}}, \Phi_2^{\text{opt}}). \tag{20}$$

300

4. Examples

4.1 *An EWMA case study*

A simulated chemical mechanical polishing (CMP) process developed by
305 SEMATECH (Moyne 2001) is studied. A typical diagram of CMP process is shown in Figure 4. The control of CMP process is known to be difficult because of

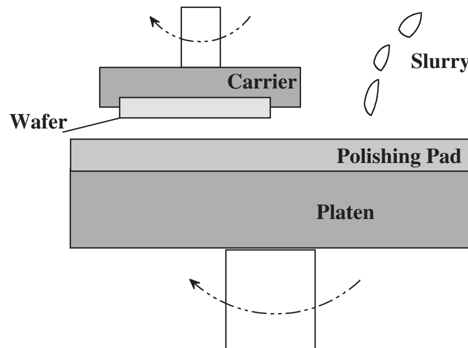

Figure 4. The diagram of CMP process.

poor understanding of the process, worn out of polishing pads, and the lack of in-situ sensor.

$$\begin{cases} y_{1,k} = 1563.5 + 159.3u_{1,k-1} + \dfrac{\delta_1}{1-z^{-1}} + \dfrac{1-\theta_1 z^{-1}}{1-z^{-1}}\varepsilon_{1,k} \\ y_{2,k} = 254 + 32.6u_{2,k-1} + \dfrac{\delta_2}{1-z^{-1}} + \dfrac{1-\theta_2 z^{-1}}{1-z^{-1}}\varepsilon_{2,k} \end{cases}. \tag{21}$$

310 The noise parameters are: $\delta_1 = -5.7$, $\delta_2 = -0.6$, $\theta_1 = 0.7$, $\theta_2 = 0.65$, and $\varepsilon_{1,k} \sim N(0, 10^2)$, $\varepsilon_{2,k} \sim N(0, 8^2)$. For the CMP process, the manipulated variables are platen speed (u_1), and polishing down-force (u_2), while the controlled variables are removal rate (y_1), and within-wafer non-uniformity (y_2). It is observed that the removal rate has the tendency to decrease as the polishing pad wears out rapidly (Chen and Guo 2001), which indicates a large drift
315 term involved in the noise model.

Two EWMA schemes ($\lambda = 0.15$) are adopted to control both SISO loops. Performance assessment is carried out for both EWMA controllers via output data $\langle \mathbf{Y_1}, \mathbf{Y_2} \rangle$. The identified noise models are $N_{1,k} = N_{1,k-1} - 5.815 + (1 - 0.693z^{-1})\varepsilon_{1,k}$ and $N_{2,k} = N_{2,k-1} - 0.605 + (1 - 0.679z^{-1})\varepsilon_{2,k}$, which correspond well with the real noise. The
320 calculated MVP bounds for loop 1 and loop 2 are 98.6485 and 62.8406 respectively, which also show good agreement with the theoretical value 10^2 and 8^2. Detailed results are summarised in Table 1.

Remark 5:

(a) For a deterministic drift noise, EWMA controller will inevitably result in output
325 offset, $\mu_Y = \delta/\lambda$ As shown in Table 1, the removal rate control has large offset since it involves with severe drift.

(b) The optimal discount factor should be close to 1 when the drift effect is dominating. Consider loop 1 for example, the optimal settings under a series of δ/σ are depicted in Figure 5. Obviously, when the drift is more dominating
330 (i.e. $\delta/\sigma \to 1$), the optimal setting value increased steadily to 1, and when $\delta/\sigma \to 0$, the optimal setting is close to 0.3 (the theoretical optimal value, $\lambda = 1 - \theta$).

(c) For a drift process, EWMA controller can't achieve the MVP bound. The best achievable performance for loop 1 is 174.2162, a much higher bound than the minimum variance performance.

335 (d) The BAP bound for loop 2 is as close as the MVP bound which indicates that for a small drift noise process ($\delta/\sigma \to 0$), a single EWMA controller is sufficient and should be preferred because its tuning procedure is more straightforward compared with other RtR controller. This explains why the EWMA controller is

Table 1. Performance assessment of EWMA control.

Performance	Loop 1	Loop 2	Loop 2 ($\delta = 0$)
Output offset	−37.2700	−3.9110	0.5595
Optimal λ	0.8604	0.4140	0.3387
MVP	98.6485	62.8406	63.5299
EWMA-BAP	174.2162	65.1203	63.5675
System performance	1495.1000	86.0305	74.5140

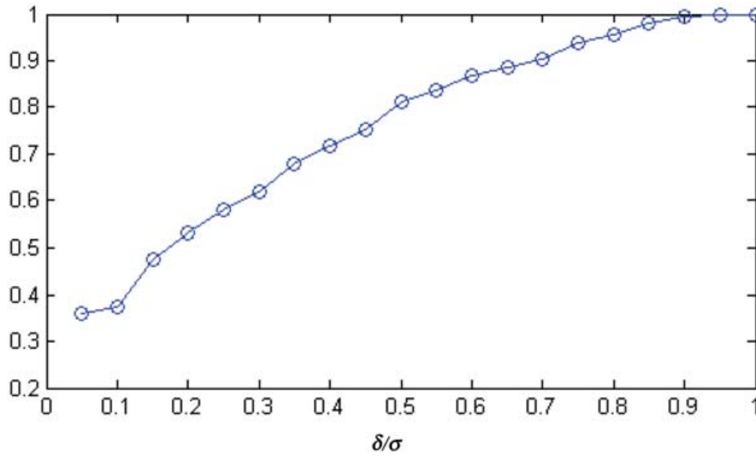

Figure 5. The optimal EWMA setting under different drift effect.

so popular in semiconductor process control since most of the processes involve with relatively small drift. A special case is shown in Table 1. If no drift exists in the process noise, the output offset of loop 2 is 0.5595 (almost zero) and the MVP bound can be achieved by EWMA control for an optimal setting 0.3387.

For the convenience of performance analysis, two performance indices are defined in this article.

$$P_1 = \frac{\text{MVP}}{\text{BAP}}, \tag{22}$$

$$P_2 = \frac{\text{BAP}}{\text{SYS}}, \tag{23}$$

where MVP, BAP and SYS represent the minimum variance performance, best achievable performance and system performance ($\text{SYS} = (\sum_{i=1}^{N} Y_i^2)/(N-1)$) respectively. Both normalised performance indices are restricted to $(0, 1]$. Large index value is preferred for better performance. For example, $P_1 = 1$ indicates the RtR controller is superior enough as a minimum variance controller, and $P_2 = 1$ means the RtR control system is running at its best achievable performance.

For an EWMA controller with a fixed discount factor (used most frequently in real semiconductor manufacturing), its performance indices are shown in Figure 6. As the drift term becomes more dominating, the P_2 index descends steeply indicating a rapid degradation in system performance, while the P_1 index descends steadily and the BAP bound is still acceptable even in the worst case. This indicates the benefit of using an optimal auto-tuning EWMA controller (see detail in Section 5).

4.2 *A double EWMA case*

As demonstrated in Figure 6, an EWMA controller is intrinsically not suitable for a process with severe drift. When δ/σ increases the EWMA P_1 index decreases

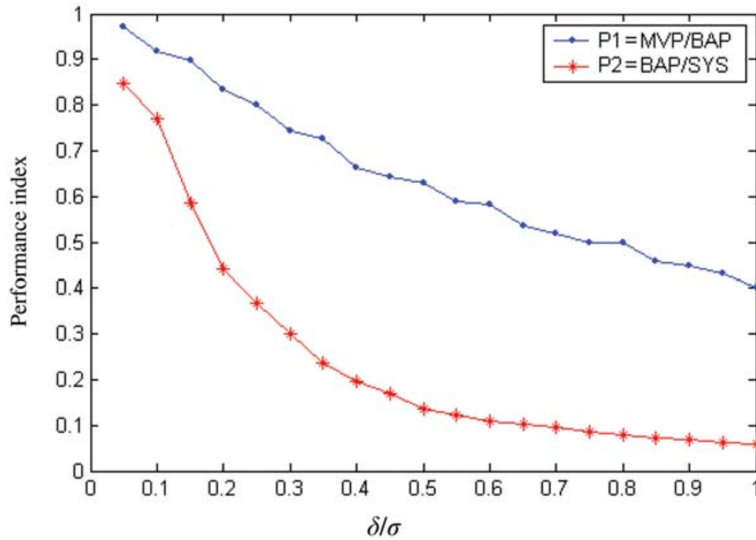

Figure 6. Performance indexes of EWMA control with a fixed discount factor.

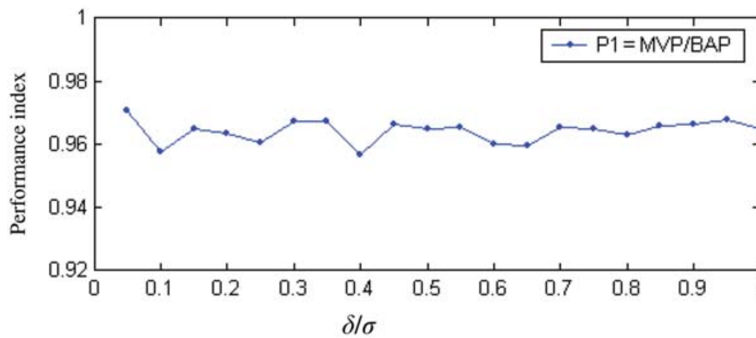

Figure 7. Performance index of double EWMA control.

inevitably due to large output deviation. Contrary to EWMA control, no such deterministic trend is observed in Figure 7 for the double EWMA controller. For the
365 same process and same noise disturbance, the dEWMA controller shows a much higher P_1 index of around 0.96. Evidently, a dEWMA controller is more suitable for relatively large process drift.

For the CMP process considered in the EWMA case, we design a dEWMA controller to handle the severe drift effect in removal rate control. The results of performance
370 assessment are summarised in Table 2. Output offset (−0.1824, almost zero) and MVP bound (103.520, 3 close to 10^2) show good agreement with the theoretical studies.

Remark 6:

(a) According to different optimal settings, three dEWMA-BAP bounds are derived in Table 2. Special attention should be paid when choosing the optimal setting. It is a
375 trade-off between smaller BAP bound and better transient performance.

Table 2. Performance assessment of double EWMA controller.

Performance	Loop 1 (Removal rate control) ($\lambda_1 = 0.15$, $\lambda_2 = 0.2$)			Loop 1 ($\lambda_2 = 0$) $\delta \neq 0$	Loop 1 ($\lambda_2 = 0$) $\delta = 0$
Output offset		−0.1824		−36.8646	0.1109
MVP		103.5203		106.9960	100.2484
System performance		111.2743		1475.0000	109.0489
dEWMA-BAP	103.7696	106.9989	108.1324	180.7300	101.2395
Optimal (λ_1, λ_2)	(0.34, 0.01)	(0.16, 0.16)	(0.177, 0.1)	0.8490	0.2966

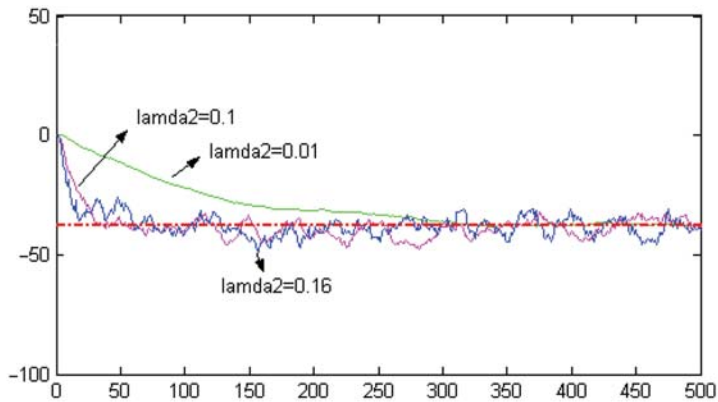

Figure 8. Transient processes of P_k in dEWMA control.

For example, the optimal setting (0.34, 0.01) should be avoided. Although it offers the smallest BAP bound, it results in bad transient performance. Using Equations (A24) and (A26), $\mu_Y = (-\lambda_1 E[p_{k-2}] + \delta)/(\lambda_1 + \lambda_2)$ and $\lim_{k\to\infty} E[p_k] = \delta/\lambda_1$, the process transients of P_k for all the optimal settings are compared in Figure 8. As expected, all the processes asymptomatically converge to the value of δ/λ_1 (about 38). However, for $\lambda_2 = 0.01$ (too small), the transient performance is extremely poor, requiring nearly 300 runs to eliminate the drift effect. While for larger discounter factors $\lambda_2 = 0.16$ and $\lambda_2 = 0.1$, the output deviation can be eliminated just after a few runs. Therefore, the optimal setting $(0.16, 0.16)$ or $(0.177, 0.1)$ should be preferred since they guarantee both long run and transient performance.

(b) The P_1 index for dEWMA control is close to 1 (see Figure 7). When it is necessary to reduce the output variation (e.g. achieving a system performance about 50), it's useless to re-tune the dEWMA controller or devise a novel controller. The only way is to modify the process to reduce the noise variance. This conclusion isn't intuitive, but is underlined by performance assessment. It does provide insights and better understanding of process nature.

(c) In Table 2, the special case of $\lambda_2 = 0$ (an EWMA controller) is displayed. It is remarkable that the performance assessment results are almost the same as that of EWMA control shown in Table 1.

Table 3. Performance assessment of RLS-LT control.

	Loop 1 (Removal rate control)	
Performance	dEWMA	RLS-LT
MVP bound	104.2002	104.2002
BAP bound	107.9782	105.3120
Optimal setting	(0.2004, 0.1)	0.8538

Table 4. Performance assessment for EWMA control with model mismatch.

	$\xi = \beta/\mathrm{b}$ (EWMA)					
Performance	0.5	1.0	1.5	2.0	2.5	3.0
MVP bound	119	100	107	107	107	101
BAP bound	236	183	167	154	149	137
Optimal setting	0.95	0.85	0.68	0.58	0.51	0.45

(d) The RLS-LT control is a special case of double EWMA control under certain circumstances. A RLS-LT control case study was also carried out in this section. Based on the same input output data, using the same identification method, we compare the performance assessment results for both controllers in Table 3. We repeat the experiment several times, and find the BAP bound of the RLS-LT controller is a little better than that of the dEWMA controller.

4.3 *Robustness of the methodology*

In this part, we will evaluate the robustness of the performance assessment methodology, i.e. consider the accuracy of the results in case of model mismatch, more sophisticated noise disturbance, and metrology delay.

4.3.1 *Model mismatch*

Let us re-consider the removal rate control of the CMP process. To implement the EWMA or dEWMA controller, the process gain should be estimated in advance by off-line identification. Normally, good estimation can be realised ($b \approx \beta$) since the semiconductor process is described by a linear model. However, inaccurate estimation or model mismatch does exist in practical application due to inadequate data set or time variant process dynamics. In such cases, inaccurate performance assessment results occur. The performance assessment results of EWMA and dEWMA control under different model mismatch is summarised in Tables 4 and 5, respectively.

Remark 7:

(a) Model mismatch has little influence on MVP bound calculation, but has a strong effect on BAP bound assessment, especially for the EWMA controller.

Table 5. Performance assessment for dEWMA control with model mismatch.

	$\xi=\beta/b$ (dEWMA)					
Performance	0.5	1.0	1.5	2.0	2.5	3.0
MVP bound	119	100	107	107	107	101
BAP bound	124	106	111	110	112	106
Optimal setting	(0.1, 0.31)	(0.11, 0.11)	(0.1, 0.1)	(0.1, 0.1)	(0.1, 0.1)	(0.1, 0.1)

Table 6. Performance assessment for EWMA and dEWMA control under other noises.

Value		IMA (1, 2)	ARIMA (1, 1, 1)
$(\hat{\delta}, \hat{\theta})$		(−5.68, 0.96)	(−7.48, 0.29)
BAP	EWMA	221.75	169.15
	dEWMA	118.06	110.10

An over-estimated process gain ($\xi<1$) will generate a large BAP bound, and for under-estimated gains ($\xi<1$), BAP bound and optimal setting trends to decline. However, for a moderate model mismatch $\xi\in[0.8, 1.5]$, the accuracy of BAP bound calculation will be above 90% (when $\xi=0.8$, BAP bound ≈ 198), which is acceptable in practice.

(b) The robustness of performance assessment of dEWMA control under model mismatch is shown in Table 5. The results are consistent with those in Table 2.

(c) As suggested by former research, an over-estimated gain b will guarantee long run stability (Castillo 1999). It is true but meanwhile results in worse performance assessment results.

(d) Pay attention to the stability condition when $\xi>2$ for EWMA and $\xi>1$ for dEWMA, because for $0<\lambda<1$ the stability is no longer guaranteed in these situations (Castillo 1999, 2002). This is the reason why a smaller discounter factor is always preferred (Edgar *et al.* 2000, Moyne *et al.* 2001).

4.3.2 *Sophisticated noises*

On occasion, semiconductor processes may involve more sophisticated noise disturbance. For example, IMA (1, 2) noise with drift $(1-z^{-1})N_k=\delta+(1-\theta_1-\theta_2)\varepsilon_k$ or ARIMA (1, 1, 1) noise with drift $(1-z^{-1})(1-\phi)N_k=\delta+(1-\theta)\varepsilon_k$. Successful performance assessment of such processes lies greatly with closed-loop identification. If the identification method is robust enough to approximate the real noise disturbance with an estimated noise model $(1-z^{-1})\hat{N}_k=\hat{\delta}+(1-\hat{\theta})\varepsilon_k$, then the results are worthy of confidence.

Again, we consider the removal rate control of CMP process. Two process disturbances, namely IMA (1, 2) with drift ($\delta=-5.7$, $\theta_1=0.7$, $\theta_2=0.3$) and ARIMA (1, 1, 1) with drift ($\delta=-5.7$, $\phi=0.25$, $\theta=0.5$) are engaged in the simulation studies. The results for EWMA and dEWMA controller are listed in Table 6. Furthermore, we compare the identified noise with the real process noise. As shown in Figures 9 and 10, the identified noise models can accurately capture the dynamics of real disturbances. Hence, the corresponding performance assessment results are believable. It should be

XML Template (2008) [18.4.2008–5:39pm] [1–27]
{TANDF_FPP}TPRS/TPRS_A_303131.3d (TPRS) [First Proof]

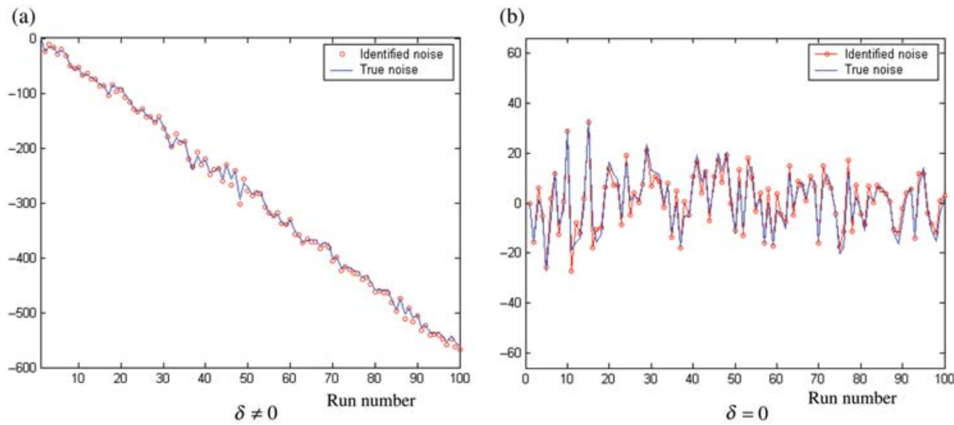

Figure 9. Compare the identified noise with IMA (1, 2) with drift.

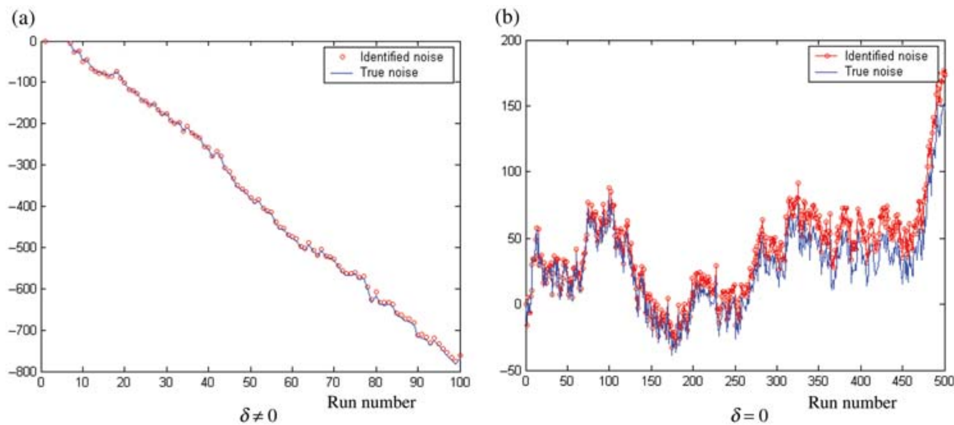

Figure 10. Compare the identified noise with ARIMA (1, 1, 1) with drift.

noted that although the BAP bounds for EWMA control changes greatly, those for dEWMA control are more stable for different noises.

4.3.3 *Metrology delay*

The last case considered is metrology delay introduced by *ex situ* measure equipment in semiconductor manufacturing. Performance assessment of RtR control system with metrology delay is directly considered in our research, see Equations (9), (12) and (13). P_1 index for both EWMA and dEWMA control in terms of a series of metrology delay is plotted in Figure 11. It is observed that the minimum variance performance is more easily achieved with a larger metrology delay. This phenomenon is not intuitive, but does account for the severe effect of large metrology delay. As the delay increases, it becomes the most dominant factor of performance degradation (i.e. in Equations (9) and (12), $MVP >> Objective\ performance$), and P_1 index goes asymptomatically to 1.

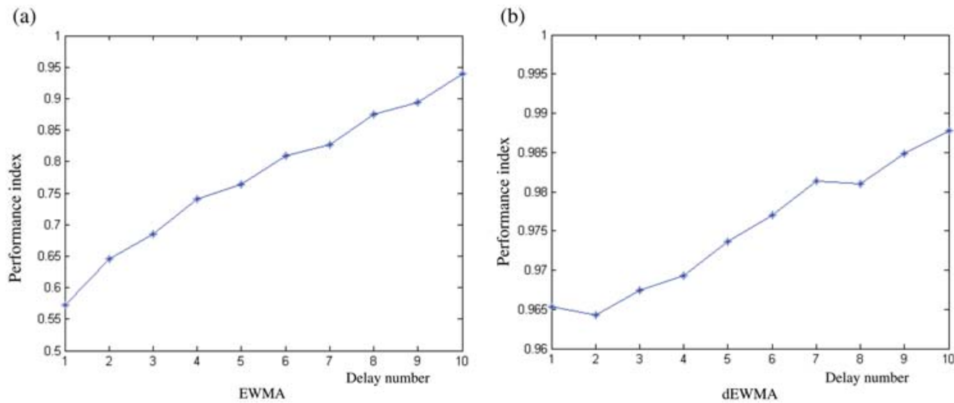

Figure 11. Trend of P_1 index with metrology delay.

Similar conclusions can be found in Qin (1998) for PI performance assessment and Prabhu *et al*. (2006) for EWMA performance assessment. This tendency is essential and 460 profound. It indicates that for a RtR control system involved with large metrology delay, the most important suggestion for performance improvement is to modify process dynamics to reduce time delay (e.g. using a *in situ* metrology equipment) rather than to devise a novel controller.

5. Extensions

465
5.1 *Online performance monitoring*

In this section, a simple but useful extension of the above method is provided for online performance monitoring. The method is easy to implement and efficient in computation, and is possible for real-time application. The moving window technique is used here to calculate the online performance indices. The window length is selected as 50 runs. Shorter 470 window length is also possible, but will results in poor results of closed loop identification.

Again, take the removal rate control as an illustrative example. The lifetime of the polishing pad is quite limited due to the wear out process. PMs are scheduled after certain runs to maintain the product on target. Figure 12 shows a typical PM scheduling in semiconductor manufacturing. The worn out effect introduces a deterministic drift, and 475 produces a gradual descent in process noise. PM is performed every 300 runs to update a new polishing pad. The drift rate changes slightly between different PMs, resulting in time-variant performance indices shown in Figure 13. It is noted that a large process drift (between runs 300–600) results in a bad system performance ($P_2 \approx 0.15$), and a small drift term (between runs 600–900) leads to good system performance ($P_2 \approx 0.3$). However, 480 inaccurate index should also be noted in Figure 13. These upsets are caused by the inaccuracy of the input-output time series. For example, the indices between runs 600–650 are calculated based on the mixed data set of previous and current manufacturing.

As shown in Figure 13, the system performance is really poor suggested by a small P_2 index, and significant performance improvement can be achieved since a large P_1 index 485 indicates the BAP bound is quite agreeable (125% of MVP bound). Note the BAP bound can be easily achieved by optimal tuning of the controller.

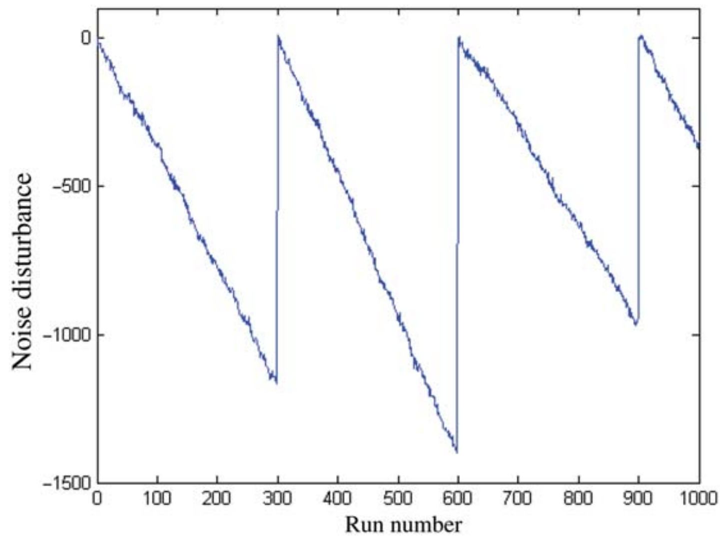

Figure 12. PM scheduling in semiconductor manufacturing.

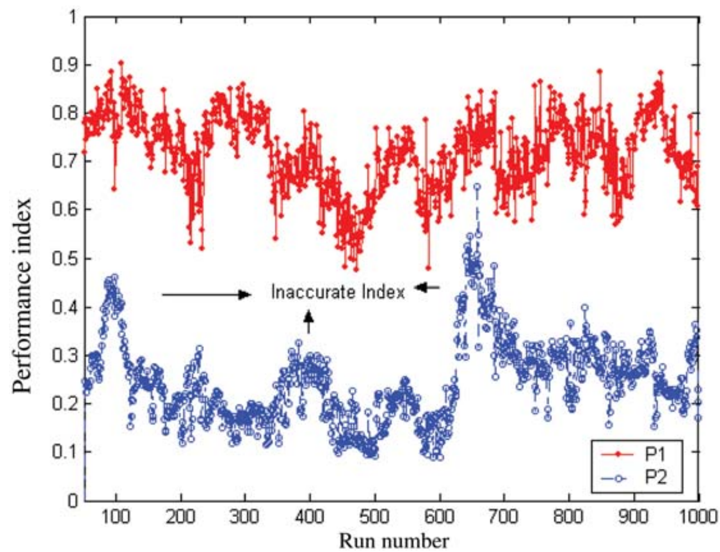

Figure 13. Online performance monitoring of CMP process.

5.2 *Online optimal tuning RtR controller*

In this part, a further application of online tuning RtR controller is shown. The basic idea for online optimal tuning is to update the controller settings with the optimal value computed by online BAP assessment. This strategy is simple yet proved to be effective. As shown in Figure 14, significant improvement is made by online tuning. The first 50-run data are used as the training set.

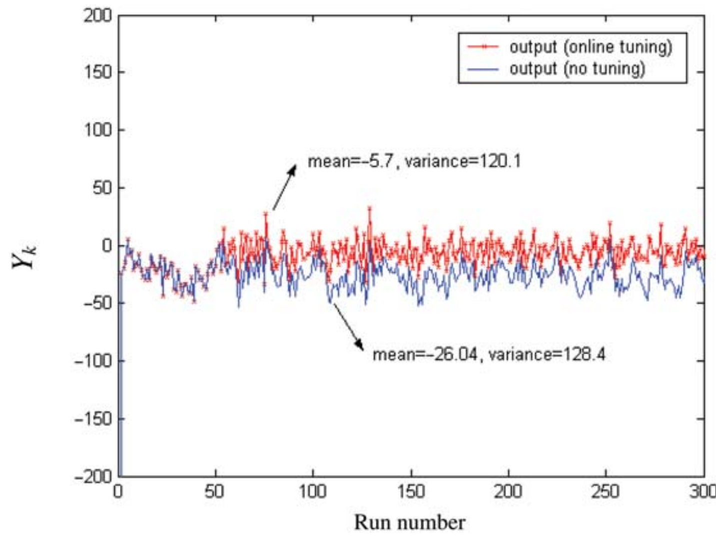

Figure 14. Compare the process outputs with or without online tuning.

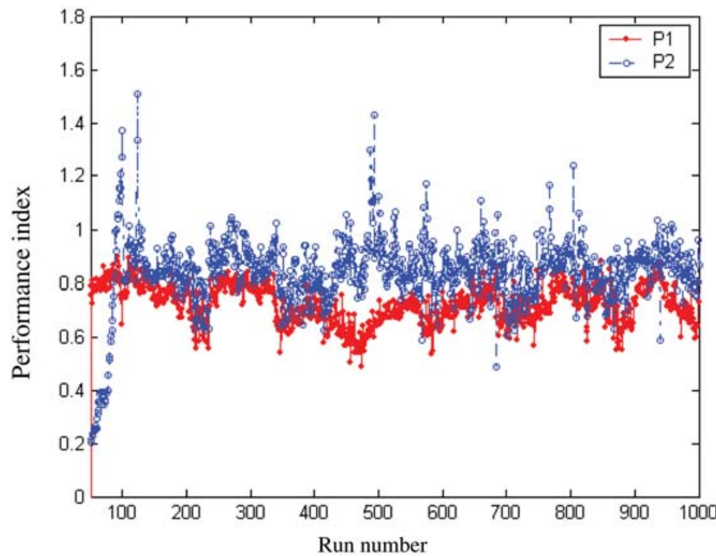

Figure 15. Performance index of RtR control with online tuning.

The online monitored performance index with optimal tuning is further investigated. As shown in Figure 15, the performance index $P_2 \approx 0.95$ indicates great improvement of system performance, and it also means the EWMA controller is running almost at its best achievable performance. Because both MVP and BAP bounds are invariant term, the P_1 index keeps unchanged as expected. Again, some inaccurate indices arose due to the inaccuracy of the output data.

The optimal time-variant discounter factors are plotted in Figure 16. For the mean value of the optimal setting during the three periods,

XML Template (2008) [18.4.2008–5:40pm] [1–27]
{TANDF_FPP}TPRS/TPRS_A_303131.3d (TPRS) [First Proof]

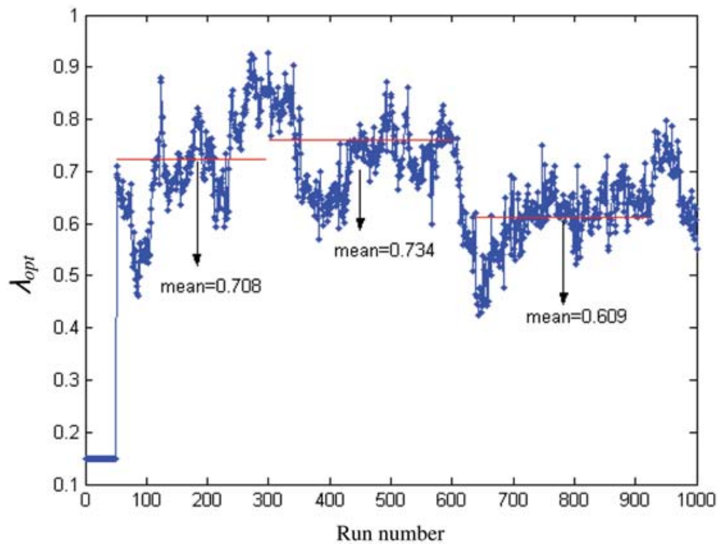

Figure 16. Time-variant optimal settings.

mean $(\lambda_2^{opt}) >$ mean $(\lambda_1^{opt}) >$ mean (λ_3^{opt})(note $\delta_2 > \delta_1 > \delta_3$), this is consistent with the conclusion stated in Remark 5(b).

6. Conclusion

There's significant incentive to assess run-to-run control performance via closed loop data to ensure the system is running at its optimal point. Compared with former work of Prabhu (2006), the methodology proposed in this paper is more straightforward, and can be used to evaluate a series of RtR control systems. Based on IMC framework, more rigorous expressions for performance assessment are derived from control engineering point of view. Our methodology is carried out and validated by examples of EWMA control, double EWMA control and RLS-LT control. In practice, for any RtR controller, performance assessment can be deduced following a very similar way as long as it can be represented by IMC structure.

Several essential characteristics of the RtR controllers are explored by performance assessment. Though some of these results have already existed in the literature, we derive them in a more systematic way, and give more theoretical explanations. Below, we summarise some fundamental results as guidelines for process engineers of run-to-run control.

(1) Most of the semiconductor processes involve relatively small drift, hence the EWMA controller is to be considered when designing a RtR control system, since it provides agreeable performance and easy of tuning.
(2) If the process suffers from severe drift, it is a wise choice to use a controller with drift compensation, e.g. double EWMA controller, RLS-LT controller, etc. In such a situation, re-tuning the EWMA controller will not provide any benefit.

525 (3) If the process input-output gain cannot be well estimated (with large departure), the double EWMA controller is preferred since its performance and stability trade-off can be easily realised using a pair of small discounter factors (too small value should be avoid for poor transient performance).

(4) For an extremely poor performance, thorough assessment is needed. Pay 530 more attention to process dynamics rather than the controller. The underlying causes may be a large metrology delay or a non-stationary noise. In such situation, improved process dynamics is the only way to achieve better performance.

(5) Online performance assessment is an effective way for system performance 535 monitoring to avoid product failure and reduce rework rate.

More trivial guidelines can be found in the paper.

References

Adivikolanu, S. and Zafiriou, E., 2000. Extension and performance/robustness tradeoffs of the EWMA run-to-run controller by internal model control structure. *IEEE Transactions of* 540 *Electronic Packaging and Manufacture*, 23, 56–68.

Bode, C.A., Ko, B.-S., and Edgar, T.F., 2004. Run-to-run control and performance monitoring of overlay in semiconductor manufacturing. *Control Engineering Practice*, 12, 893–900.

Bulter, S.W. and Stefani, J.A., 1994. Supervisory run-to-run control of polysilicon gate etch using in situ ellipsometry. *IEEE Transactions of Semiconductor Manufacturing*, 7 (2), 193–201.

545 Box, G.E.P. and Jenkins, G.M., 1994. *Time series analysis: forecasting and control*. Englewood Cliffs, NJ: Prentice Hall.

Castillo, E. Del., 1997. Run-to-run process control: literature review and extensions. *Journal of Quality Technology*, 92 (2), 184–196.

Castillo, E. Del., 1999. Long run and transient analysis of a double EWMA feedback controller. *IIE* 550 *Transactions*, 31, 1157–1169.

Castillo, E. Del., 2001. Some properties of EWMA feedback quality adjustment schemes for drifting disturbance. *Journal of Quality Technology*, 33 (2), 153–166.

Castillo, E. Del., 2002. *Statistical process adjustment for quality control*. New York: John Wiley and Sons.

555 Chen, A. and Guo, R.S., 2001. Aged-based double EWMA controller and its application to CMP processes. *IEEE Transactions of Semiconductor Manufacturing*, 14 (1), 11–19.

Desborough, L.D. and Harris, J., 1993. Performance assessment for univariate feedforward/feedback control. *Canadian Journal of Chemical Engineering*, 71, 605–616.

1 Edgar, T.F., *et al.*, 2000. Automatic control in microelectronics manufacturing: practices, challenges 560 and possibilities. *Automatica*, 36, 1567–1603.

1 Firth, S.K., *et al.*, 2006. Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes. *IEEE Transactions of Semiconductor Manufacturing*, 19 (3), 298–315.

Harris, T.J., 1989. Assessment of control loop performance. *Canadian Journal of Chemical* 565 *Engineering*, 67, 856–861.

Harris, T.J., Seppala, C.T., and Desborough, L.D., 1996. Performance assessment of multivariable feedback controllers. *Automatica*, 32 (11), 1505–1518.

2 Harris, T.J., 1999. A review of performance monitoring and assessment techniques for univariate and multivariate control systems. *Journal of Proc. Control*, 9, 1–17.

570 Huang, B., Shah, S.L., and Kwok, E.K., 1997. Good, bad or optimal? Performance assessment of multivariable processes. *Automatica*, 33 (6), 1175–1183.

XML Template (2008) [18.4.2008–5:40pm] [1–27]
{TANDF_FPP}TPRS/TPRS_A_303131.3d (TPRS) [First Proof]

Ingolfsson, A. and Sachs, E., 1993. Stability and sensitivity of an EWMA controller. *Journal of Quality Technology*, 25 (4), 271–287.

Ko, B.-S. and Edgar, T.F., 2000. Performance assessment of cascade control. *AIChE Journal*, 46 (2), 281–291.

Ko, B.-S. and Edgar, T.F., 2001a. Performance assessment of multivariable feedback control system. *Automatica*, 37 (6), 899–905.

Ko, B.-S. and Edgar, T.F, 2001b. Performance assessment of constrained model predictive control systems. *AIChE Journal*, 47 (6), 1363–1371.

Ko, B.-S. and Edgar, T.F., 2004. PID control performance assessment: the single-loop case. *AIChE Journal*, 50 (6), 1211–1218.

Ma, M.D., Jang, S.S. *et al.*, 2007. State estimation of a mixed run plant. Submitted to *Automatica*. ▮▮.

Moyne, J., *et al.*, 2001. *Run-to-run control in semiconductor manufacturing*. Florida: CRC Press.

Patel, N.S. and Jenkins, S.T., 2000. Adaptive optimisation of run-to-run controllers: the EWMA example. *IEEE Transactions of Semiconductor Engineering*, 13 (1), 97–107.

Prabhu, A.V., Edgar T.F. and Chong R., 2006. Performance assessment of run–to–run EWMA controllers. *International symposium on advanced control of chemical processes*, ▮▮, Gramado, Brazil. ▮▮: ▮▮, 1127–1132.

Qin, S.J., 1998. Control performance monitoring – a review and assessment. *Comp. and Chemical Engineering*, 23, 173–186.

Qin, S.J., *et al.*, 2006. Semiconductor manufacturing process control and monitoring: a fab–wide framework. *Journal of Proceedings Control*, 16, 179–191.

Sachs, E., Hu, A., and Ingolfsson, A., 1995. Run by run process control: combining SPC and feedback control. *IEEE Transactions of Semiconductor Manufacturing*, 8 (1), 26–43.

Stanfelj, N., Marlin, T.E., and MacGregor, J.F., 1993. Monitoring and diagnosing process control performance: the single loop case. *Ind. Engineering and Chemical Research*, 67 (10), 856–861.

Tseng, S.-T., *et al.*, 2003. A study of variable EWMA controller. *IEEE Transactions of Semiconductor Manufacturing*, 16 (4), 633–643.

Wang, J., Qin, S.J., *et al.*, 2005. Recursive least squares estimation for run–to–run control with metrology delay and its application to STI etch process. *IEEE Transactions of Semiconductor Manufacturing*, 18 (2), 309–319.

Zheng, Y., *et al.*, 2006. Stability and performance analysis of mixed product run-to-run control. *Journal of Proc. Control*, 16, 431–443.

Appendix (Proofs)

A1. Proof of Lemma

For the IMC control system, the process output can be described by the following z-transfer function

$$Y_k = \frac{G_w(z^{-1}) - G_w(z^{-1})G_{\mathrm{IMC}}(z^{-1})G_u(z^{-1})}{1 + G_{\mathrm{IMC}}(z^{-1})(G_p(z^{-1}) - G_u(z^{-1}))}\varepsilon_k. \tag{A1}$$

If $G_p(z^{-1}) = G_u(z^{-1})$.

$$Y_k = (G_w(z^{-1}) - G_w(z^{-1})G_{\mathrm{IMC}}(z^{-1})G_u(z^{-1}))\varepsilon_k. \tag{A2}$$

Based on the Diophantion equation, the noise model can be expressed as

$$G_w(z^{-1}) = \frac{B_w(z^{-1})}{A_w(z^{-1})} = F(z^{-1}) + \frac{z^{-d}G(z^{-1})}{A_w(z^{-1})} \text{ or } B_w(z^{-1}) = A_w(z^{-1})F(z^{-1}) + z^{-d}G(z^{-1}). \tag{A3}$$

Substitute it into Equation (A2)

$$\begin{aligned} Y_k &= (F(z^{-1}) + \frac{z^{-d}G(z^{-1})}{A_w(z^{-1})})\varepsilon_k - \beta\frac{B_w(z^{-1})}{A_w(z^{-1})}G_{\mathrm{IMC}}(z^{-1})\varepsilon_{k-d} \\ &= F(z^{-1})\varepsilon_k + A_w^{-1}(z^{-1})(G(z^{-1}) - \beta B_w(z^{-1})G_{\mathrm{IMC}}(z^{-1}))\varepsilon_{k-d} \end{aligned} \quad \text{(A4)}$$

615 Here, $G_{IMC}(z^{-1}) = G_f(z^{-1})G_c(z^{-1}) = \beta^{-1}G_f(z^{-1})$ is the IMC representation of the RtR controllers. Hence, the process output can be expressed by the following equation.

$$Y_k = \underbrace{F(z^{-1})\varepsilon_k}_{\text{Uncontrollable}} + \underbrace{A_w^{-1}(z^{-1})(G(z^{-1}) - B_w(z^{-1})G_f(z^{-1}))\varepsilon_{k-d}}_{\text{Controllable}}. \quad \text{(A5)}$$

620

A2. Proof of Theorem 1

According to the time series analysis, the mean square error of system output is

$$\mathrm{MSE}(Y_k) = \mathrm{Var}(Y_k) + \mu_Y^2 \quad \text{(A6)}$$

where μ_Y is defined as mean deviation from target, $\mu_Y = E[Y_k] - Y_r$ (in the following paper, without loss of generality, we assume a zero set point, i.e. $\mu_Y = E[Y_k]$).

625 It is well known for a white noise added to the output channel, $E[Y_k] = 0$. However, for a non-stationary disturbance like IMA (1, 1) with deterministic drift, Box *et al.* proved it can be expressed as

$$N_k = \frac{1 - \theta z^{-1}}{1 - z^{-1}} e_k \quad \text{(A7)}$$

where the coloured noise $e_k \sim (\delta/(1-\theta), \sigma_\varepsilon^2)$ indicates the drift term could result in a nonzero

630 output ($E[Y_k] \neq 0$), but has no effect on the output variance.

Based on the lemma for IMC performance assessment, an EWMA filter is used in EWMA control system

$$G_f^{\mathrm{EWMA}} = \frac{\lambda}{1 - (1-\lambda)z^{-1}}. \quad \text{(A8)}$$

Substituting Equations (A7) and (A8) into Equation (A5), we get

$$Y_k = Fe_k + \frac{1}{1 - z^{-1}}\left(G - \frac{\lambda(1 - \theta z^{-1})}{1 - (1-\lambda)z^{-1}}\right)e_{k-d}. \quad \text{(A9)}$$

635 Apply Equation (A7) to Diophantion equation, $G = 1 - \theta$, and we finally get

$$Y_k = Fe_k + \frac{1 - \lambda - \theta}{1 - (1-\lambda)z^{-1}}e_{k-d}. \quad \text{(A10)}$$

According to the classical theory of time series analysis, the output variance for system (A10) is

$$\mathrm{Var}(Y_k) = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma_\varepsilon^2 + \frac{(1 - \lambda - \theta)^2}{1 - (1-\lambda)^2}\sigma_\varepsilon^2. \quad \text{(A11)}$$

640 To calculate the nonzero mean deviation μ_Y, an EWMA system in Equation (8) is transformed into Equation (A12). Here, the notation ∇ is an abbreviation of $1-z^{-1}$.

$$\begin{cases} \nabla Y_k = \beta \nabla U_{k-1} + \nabla N_k \\ \nabla U_k = -\dfrac{\nabla a_k}{b} = -\dfrac{\lambda Y_k}{b} \\ \nabla N_k = \delta + (1-\theta z^{-1})\varepsilon_k \end{cases}. \tag{A12}$$

Convert Equation (A12) into a concise expression (A13) and (A14)

645

$$(1-(1-\lambda)z^{-1})Y_k = \delta + (1-\theta z^{-1})\varepsilon_k \tag{A13}$$

$$Y_k - \frac{\delta}{1-(1-\lambda)z^{-1}} = \frac{1-\theta z^{-1}}{1-(1-\lambda)z^{-1}}\varepsilon_k. \tag{A14}$$

Hence, the mean deviation μ_Y can be calculated as

$$\mu_Y = E(Y_k) = E\left[\frac{\delta}{1-(1-\lambda)z^{-1}}\right] = \frac{\delta}{\lambda}. \tag{A15}$$

650 Substitute Equations (A11) and (A15) into Equation (A6), finally, the mean square error of the EWMA feedback control loop is derived.

$$\mathrm{MSE}(Y_k) = \underbrace{(1+f_1^2+\cdots+f_{d-1}^2)\sigma_\varepsilon^2}_{\text{Minimum Variance Performance}} + \underbrace{\frac{(1-\lambda-\theta)^2}{1-(1-\lambda)^2}\sigma_\varepsilon^2 + \left(\frac{\delta}{\lambda}\right)^2}_{\text{Objective Performance}}. \tag{A16}$$

A3. Proof of Theorem 2

655 The IMC representation for double EWMA filter is

$$G_f^{\mathrm{dEWMA}} = \frac{(\lambda_1+\lambda_2)(1-z^{-1}) + \lambda_1\lambda_2 z^{-1}}{(1-(1-\lambda_1)z^{-1})(1-(1-\lambda_2)z^{-1})}. \tag{A17}$$

Using the lemma for IMC performance assessment, we get the formulation of dEWMA control output

$$Y_k = Fe_k + \frac{(1-\lambda_1-\lambda_2-\theta) - ((1-\lambda_1)(1-\lambda_2)-\theta)z^{-1}}{(1-(1-\lambda_1)z^{-1})(1-(1-\lambda_2)z^{-1})}e_{k-d}. \tag{A18}$$

660 When $\lambda_2 = 0$, the formulation reduces to

$$Y_k = Fe_k + \frac{1-\lambda_1-\theta}{1-(1-\lambda_1)z^{-1}}e_{k-d},$$

i.e. Equation (A10), a special case of EWMA control.

Note the second term on the right-hand side is an ARMA (2, 1) process. For a general ARMA (2, 1)

$$G_\omega(z^{-1}) = \frac{1-\Theta z^{-1}}{1-\Phi_1 z^{-1} - \Phi_2 z^{-2}},$$

665 its variance can be computed as (Box 1994)

$$\mathrm{Var}(G_\omega) = \frac{(1-\Phi_2)(\Theta^2+1) - 2\Phi_1\Theta}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}. \quad \text{(A19)}$$

For the ARMA (2,1) process in Equation (A18)

$$\begin{cases} \Phi_1 = 2-\lambda_1-\lambda_2 \\ \Theta = \dfrac{(1-\lambda_1)(1-\lambda_2)-\theta}{1-\lambda_1-\lambda_2-\theta} \\ \Phi_2 = -(\lambda_1-1)(\lambda_2-1) \end{cases}. \quad \text{(A20)}$$

670 Therefore, the output variance for dEWMA control loop is

$$\begin{aligned} \mathrm{Var}(Y_k) &= (1+f_1^2+\cdots+f_{d-1}^2)\sigma_\varepsilon^2 + \frac{(1-\Phi_2)(\Theta^2+1)-2\Phi_1\Theta}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}(\Phi_1-1-\theta)^2\sigma_\varepsilon^2 \\ &= (1+f_1^2+\cdots+f_{d-1}^2)\sigma_\varepsilon^2 + \left[\frac{-(2(\Theta^2+1)-2\Phi_1\Theta/(\Phi_1-2)(\Phi_1+2))}{\Phi_2+1}\right. \\ &\quad \left. + \frac{-((\Theta+1)^2/2(\Phi_1+2))}{\Phi_2-\Phi_1-1} + \frac{((\Theta-1)^2/2(\Phi_1-2))}{\Phi_2+\Phi_1-1}\right](\Phi_1-1-\theta)^2\sigma_\varepsilon^2. \end{aligned} \quad \text{(A21)}$$

Below we calculate mean deviation μ_Y for dEWMA control. Similarly, we rearrange dEWMA system by multiplying $\nabla = 1 - z^{-1}$.

675

$$\nabla Y_k = \beta\nabla U_{k-1} + \nabla N_k = -(\nabla a_{k-1} + \nabla p_{k-1}) + \delta + (1-\theta z^{-1})\varepsilon_k \quad \text{(A22)}$$

$$(1-(1-\lambda_1-\lambda_2)Z^{-1})Y_k = -\lambda_1 p_{k-2} + \delta + (1-\theta Z^{-1})\varepsilon_k. \quad \text{(A23)}$$

The mean deviation is expressed as

$$\mu_Y = \frac{-\lambda_1 E[p_{k-2}] + \delta}{\lambda_1+\lambda_2}. \quad \text{(A24)}$$

Argon Chen (2001) proved for a PCC (i.e. dEWMA) controller, the expectation of the second
680 EWMA filter $E[P_{k-2}]$ is

$$\begin{aligned} p_k &= \frac{\alpha\lambda_2}{\lambda_2-\lambda_1}\left((1-\lambda_2)^k - (1-\lambda_1)^k\right) + \sum_{j=1}^{k}\left[\frac{\lambda_2}{\lambda_2-\lambda_1}\left((1-\lambda_2)^{k-j} - (1-\lambda_1)^{k-j}\right)\cdot\lambda_1(\alpha+N_k)\right. \\ &\quad \left. + (1-\lambda_2)^{k-j}\lambda_2(\alpha+N_k)\right]. \end{aligned} \quad \text{(A25)}$$

In this paper, $N_k = N_{k-1} + \delta + (1-\theta z^{-1})\varepsilon_k$, $E[N_k] = \delta k$. When k approaches infinity

$$\lim_{k\to\infty} E[p_k] = \frac{\delta}{\lambda_1}. \quad \text{(A26)}$$

Note that $E[a_k] \to \alpha + \delta(n+1) - \delta/\lambda_1$, so $E[a_k + p_k] \to \alpha + \delta(n+1)$ serves as an asymptomatically
685 unbiased one step ahead forecasting of the process state. This explains why dEWMA control scheme can provide offset–free control. For details, see (Chen and Guo 2001, Castillo 1999).

For the special case of EWMA control (i.e. $\lambda_2 = 0$), according to Equation (A25), it's easy to find out $E[p_k] = 0$. Hence, the long–run mean deviation of the output is expressed as

$$\begin{cases} \mu_Y = 0 & \text{dEWMA} \quad (\lambda_1 \neq 0, \lambda_2 \neq 0) \\ \mu_Y = \delta/\lambda_1 & \text{EWMA} \quad (\lambda_2 = 0) \end{cases} \quad \text{(A27)}$$

690 Hence, for a dEWMA control system, its performance is only affected by the output variance, i.e. $\mathrm{MSE}(Y_k) = \mathrm{Var}(Y_k)$. However, if the second discount factor is too small (i.e. $\lambda_2 \to 0$), it may result in severe transient effects, and the output deviation can't be eliminated after a few runs. Therefore, we finally describe the dEWMA control performance as below.

$$MSE(Y_k)\begin{cases} = \underbrace{\sigma^2_{\mathrm{MVP}}}_{\mathrm{MVP}} + \underbrace{\dfrac{(1-\Phi_2)[(\Phi_2+\theta)^2 + (\Phi_1-\theta-1)^2] + 2\Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(1+\Phi_2)(\Phi_2+\Phi_1-1)(\Phi_2-\Phi_1-1)}\sigma^2_\varepsilon}_{\text{Objective Performance}} & \lambda_1, \lambda_2 \neq 0 \\ = \underbrace{\sigma^2_{\mathrm{MVP}}}_{\mathrm{MVP}} - \underbrace{\left\{ \begin{aligned} &\left[2\frac{(\Phi_2+\theta)^2 + (\Phi_1-\theta-1)^2 + \Phi_1(\Phi_2+\theta)(\Phi_1-\theta-1)}{(\Phi_2+1)(\Phi_1-2)(\Phi_1+2)} \right. \\ &\left. + \frac{(\Phi_1-\Phi_2-2\theta-1)^2}{2(\Phi_1+2)(\Phi_2-\Phi_1-1)} \right]\sigma^2_\varepsilon + \left(\frac{\delta}{\lambda_1+\lambda_2}\right)^2 \end{aligned} \right\}}_{\text{Objective Performace}} & \lambda_2 \to 0 \end{cases} \tag{A28}$$

695 where

$$\begin{cases} \Phi_1 = 2 - \lambda_1 - \lambda_2 \\ \Phi_2 = -(\lambda_1 - 1)(\lambda_2 - 1) \\ \sigma^2_{\mathrm{MVP}} = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2_\varepsilon \end{cases}. \tag{A29}$$