Threaded EWMA Controller Tuning and Performance Evaluation in a High-Mixed System

Ming-Da Ma, Chun-Cheng Chang, David Shan-Hill Wong, and Shi-Shang Jang

***Abstract*—The exponentially weighted moving average (EWMA) controller is a very popular run-to-run (RtR) control scheme in the semiconductor industry. However, in any typical step of semiconductor process, many different products are produced on parallel tools. RtR control is usually implemented with a "threaded" control framework, i.e., different controllers are used for different combinations of tools and products. In this paper, the problem of EWMA controller tuning and performance evaluation in a mixed product system are investigated by simulation and time-series analysis. It was found that as the product frequency changed, the tuning guidelines of a threaded EWMA controller were different for different types of tool disturbances. For a stationary ARMA(1,1) noise, the tuning parameter λ should be decreased as product frequency decreases. If the tool exhibits nonstationary tool dynamics, e.g., ARIMA(1,1,1) noise, the tuning parameter should increase as the product frequency decreases.**

***Index Terms*—Exponentially weighted moving average (EWMA), mixed product system, performance evaluation, time series.**

I. Introduction

IN the last two decades, run-to-run (RtR) control that combines statistical process control (SPC) and feedback control has been widely used in the semiconductor manufacturing industry [1], [2], [17]. RtR control adjusts process recipes or inputs from run to run to compensate for various process disturbances to maintain the process output close to a given target. RtR control can efficiently improve the product yield and throughout and reduce scrap, rework, and cycle time.

Exponentially weighed moving average (EWMA) controller is the most popular RtR control scheme in the semiconductor industry [3]. There are many different kinds of EWMA algorithms [4]–[6]. However, all of the above assume that only a single product is fabricated on a tool.

In the semiconductor manufacturing industry, production resembles an automated assembly line in which many similar products with slightly different specifications are manufactured step-by-step on a number of different tools. This constitutes a high-mix production system. RtR control is commonly implemented with a "threaded" framework. In this approach, each specific combination of tool and product is called a thread. Each thread has its own controller. Threaded EWMA control is probably the most popular control architecture, although other frameworks were proposed [8]–[10]. One key advantage of this approach is that the stability of the control system is guaranteed [7]. On the other hand, some commonly encountered questions are whether the controller performance is optimal and how should each thread be tuned differently as the product frequency changes. Controller performance evaluation has seen active development in research and gradual acceptance in industry since the original work of Harris [15]. Recent developments in this subject were reviewed by Qin [16]. Prabhu *et al.* [11] proposed a performance evaluation algorithm for an EWMA controller of a single thread. In this paper, we shall use the performance evaluation technique and time-series analysis to investigate the optimal tuning of a threaded EWMA controller in a high-mixed system.

II. Threaded EWMA Control of a Mixed Product Plant

Consider a static simple linear single-input single-output process performed in a mixed run situation on a single tool

$$y[t] = \alpha_{k[t]} + \beta_{k[t]} u[t] + N[t] \tag{1}$$

where $y[t]$ and $u[t]$ denote values of output and manipulated variable used on the tth run on the tool. $\alpha_{k[t]}$ is offset or bias term and $\beta_{k[t]}$ is the static gain term associated with the product produced on the tth run on the tool. Let us assume that they are relatively independent of time. $N[t]$ is a stochastic noise process associated with the tool. In a threaded approach, a sequence of output and input for each specific product is resampled

$$y[t_k] = \alpha_k + \beta_k u[t_k] + N_k[t_k] \tag{2}$$

where t_k is an index of the number of runs making the kth product that have been carried out. Given a process model $y = b_k u + a_k$ for each product, the offset term can be estimated by EWMA filter

$$a_k[t_k] = \lambda\,(y[t_k] - bu[t_k]) + (1-\lambda) a_k[t_k - 1]. \tag{3}$$

The control action is

$$u[t_k + 1] = \frac{T_k - a_k[t_k]}{b}. \tag{4}$$

Manuscript received November 09, 2007; revised June 18, 2009; June 22, 2009; accepted June 22, 2009.

M.-D. Ma is with the Center for Control and Guidance Technology, Harbin Institute of Technology, Harbin 150001, China.

C.-C. Chang, D. S.-H. Wong, and S.-S. Jang are with the Department of Chemical Engineering, National Tsing-Hua University, Hsin-Chu 30043, Taiwan (e-mail: ssjang@mx.nthu.edu.tw).

Color versions of one or more figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2009.2028219

0894-6507/$26.00 © 2009 IEEE

IEEE Proof Web Version

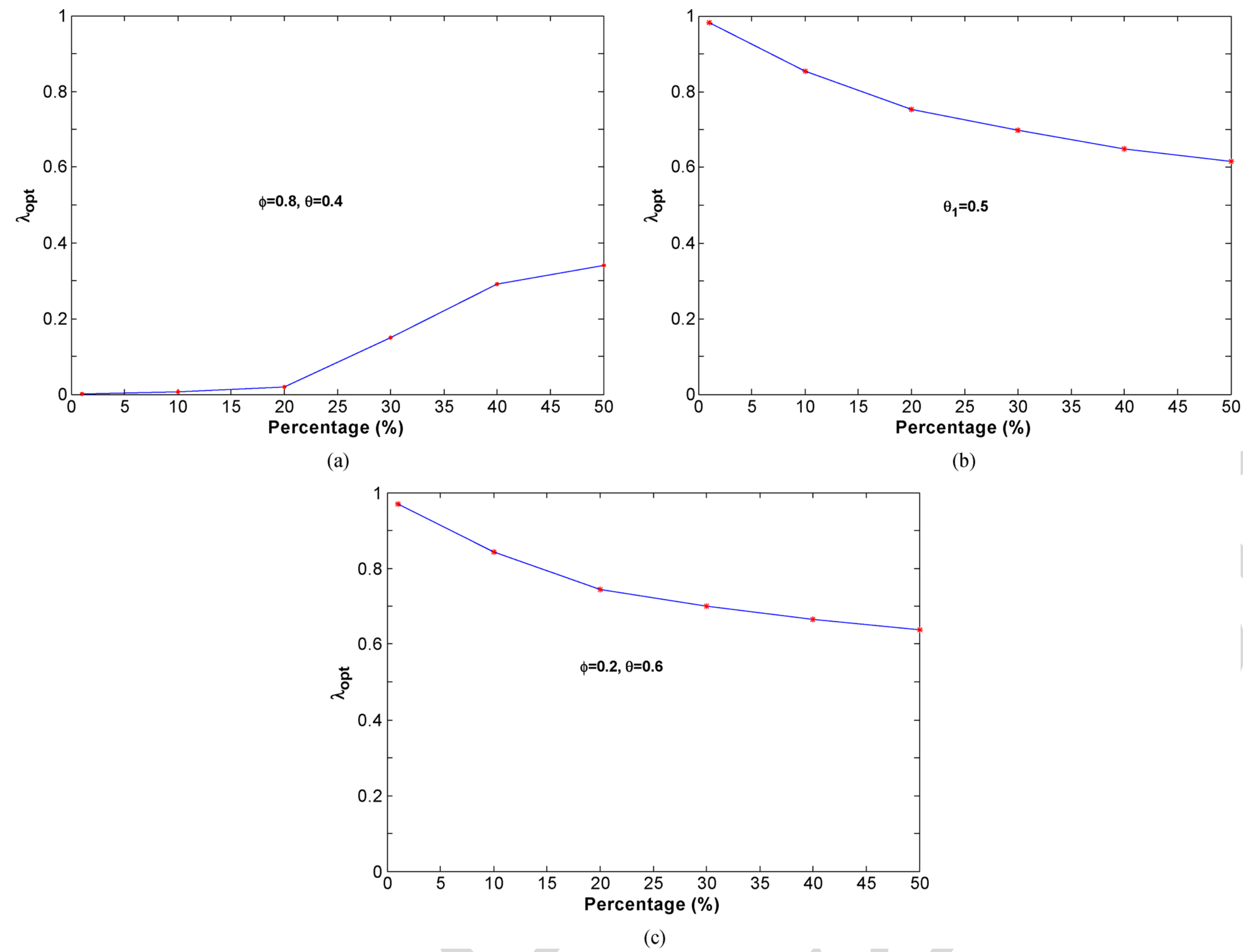

Fig. 1. Effect of production frequency on optimal tuning parameter λ. (a) ARMA(1,1) disturbance. (b) IMA(1,1) disturbance. (c) ARIMA(1,1,1) disturbance.

Note that the threaded EWMA control is similar to a single product EWMA control, except the disturbance experienced is not the actual change in tool condition $\mathrm{N[t]}$ from run to run but a resampled series $\mathrm{N_k[t_k]}$.

III. Performance Evaluation Using Simulations

Performance assessment is widely implemented in process control now. Usually, minimum variance (MV) is adopted as the benchmark. However, MV may not be easily achieved. Therefore, best achievable performance (BAP) under the current controller is often considered. Prabhu *et al.* [11] proposed a performance assessment method for EWMA control to obtain BAP estimates and optimal λ using closed-loop response data. We shall use this method to evaluate the performance of the mixed product system under threaded EWMA control to investigate how the optimal λ changes as the product frequency change.

A simulation example consisting of one tool and three products 1, 2, 3 is used. Each product is assigned a unique bias. The plant gain is assumed to be known. Three disturbances are considered: ARMA(1,1) (first-order autoregressive moving average), IMA(1,1) (first-order integrated moving average), and ARIMA(1,1,1) (first-order integrated autoregressive-moving average) process, respectively:

Noise I: $(1-\phi\mathrm{B})\mathrm{N[t]} = (1-\theta\mathrm{B})\varepsilon[\mathrm{t}]$ $\phi = 0.8, \theta = 0.4$;
Noise II: $(1-\mathrm{B})\mathrm{N[t]} = (1-\theta\mathrm{B})\varepsilon[\mathrm{t}]$ $\theta = 0.5$;
Noise III: $(1-\phi\mathrm{B})(1-\mathrm{B})\mathrm{N[t]} = (1-\theta\mathrm{B})\varepsilon[\mathrm{t}]$ $\phi = 0.2$, $\theta = 0.6$.

Here "B" stands for backward shift operator. In the simulation, the percentage of product 1 changes from 1% to 50%, product 2 is 30%, and product 3 makes up the rest of the products. For three types of disturbances, 400 sets of disturbance sequences are generated. In each set, the results of at least 100 runs of every product are collected. The production schedule is random. A value of $\lambda = 0.2$ for all threads is used to obtain the closed-loop data. The changes in estimated optimal tuning parameter for thread (product) 1, $\lambda_{1,\mathrm{opt}}$ which is the average value obtained from the above 400 sets of disturbance sequences, are shown in Fig. 1. It is interesting to find out that the tuning rules for the above disturbances are different. For Noise I, which is a stationary ARMA(1,1) noise, the optimal tuning parameter λ decreases as the percentage of the product decreases, as shown in Fig. 1(a). However, for the other two nonstationary noises II and III, the optimal tuning parameter λ increases as the percentage of the product decreases. It is interesting, therefore, to examine if the trends for nonstationary disturbances are general, as shown in Fig. 1(b) and 1(c) using time-series analysis.

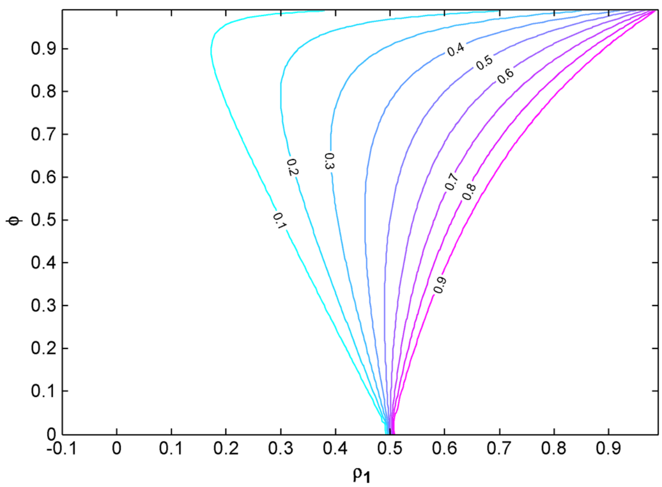

Fig. 2. Contour plots of optimal $\lambda\xi$ for ARMA(1,1) disturbance processes.

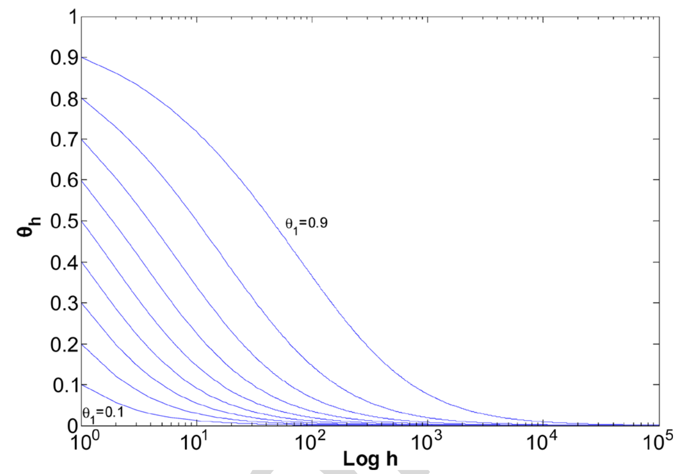

Fig. 3. Changes in parameter θ_{h} when resampling an IMA(1,1) process with parameter θ_1.

IV. Time-Series Analysis

For a simple gain process controlled by an EWMA controller with a single product, the output variance for an ARMA(1,1) disturbance $\mathrm{N[t]}$ is given by (see the Appendix)

$$\sigma_{\mathrm{y}}^2 = \left(\frac{2}{1+\kappa} - 2\rho_1 \frac{1-\kappa}{(1-\phi\kappa)(1+\kappa)}\right)\sigma_{\mathrm{N_t}}^2 \qquad (5)$$

where $\kappa = 1 - \lambda\xi$, $\xi = \beta/\mathrm{b}$, and ρ_1 is the autocorrelation of the disturbance $\mathrm{N[t]}$ at lag 1. The optimal $\lambda\xi$ given ϕ and ρ_1 can be obtained by differentiating the above equations with respect to κ.

Fig. 2 is a contour plot of optimal $\lambda\xi$ at different values of ϕ and ρ_1. In the following analysis, it is assumed that $\xi = 1$. When plant model mismatch is considered, we can simply replace λ with $\lambda\xi$. It should be noted that the optimal value of tuning parameter λ may be different when plant model mismatch exists, but ξ does not affect the changing trend of the optimal EWMA tuning parameter. In other words, the plant model mismatch does not affect the conclusion of this paper.

If the product is manufactured at regular intervals of every h runs, with h being a positive integer, then its threaded EWMA controller will face a disturbance that is a resampled sequence of the original sequence at every hth time point. If an ARMA(1,1) process with parameters (ϕ, θ) is resampled, the resulting process $\mathrm{M_h[t]}$ is also an ARMA(1,1) process [13]

$$(1-\phi_{\mathrm{h}}\mathrm{B})\mathrm{M_h[t]} = (1-\theta_{\mathrm{h}}\mathrm{B})\varepsilon_{\mathrm{h}}[\mathrm{t}] \qquad (6)$$

with

$$\phi_{\mathrm{h}} = \phi^{\mathrm{h}} \quad \rho_1(\mathrm{h}) = \phi^{\mathrm{h}-1}\rho_1 \qquad (7)$$

where $\varepsilon_{\mathrm{h}}[\mathrm{t}]$ is a new white noise process with σ_{h} and $\rho_1(\mathrm{h})$ is the autocorrelation of $\mathrm{M_h[t]}$. Since both ϕ_{h} and $\rho_1(\mathrm{h})$ decrease with h, from Fig. 2, we can infer that optimal λ must decrease with production frequency when the tool disturbance is an ARMA(1,1) noise.

If a process is stationary and the samples are taken less frequently in time, the autocorrelation of the sampled data will decrease. When the sampling interval is sufficiently large, the data will appear to be uncorrelated. It is well known that when the disturbance is white noise, no control action should be added, i.e., $\lambda = 0$. Therefore, we can draw the conclusion that, for stationary disturbances, the optimal EWMA controller gain decreases as percentage of the products decreases.

If an IMA(1,1) process $\mathrm{N[t]}$ is sampled at every h time point, the resulting process is denoted by $\mathrm{M_h[t]}$ and is also an IMA(1,1) process

$$(1-\mathrm{B})\mathrm{M_h[t]} = (1-\theta_{\mathrm{h}}\mathrm{B})\eta_{\mathrm{h}}[\mathrm{t}] \qquad (8)$$

with

$$\frac{\mathrm{h}(1-\theta_1)^2}{\theta_1} = \frac{(1-\theta_{\mathrm{h}})^2}{\theta_{\mathrm{h}}} \qquad (9)$$

and $\eta_{\mathrm{h}}[\mathrm{t}]$ is a new white noise process [12, pp. 526–528]. Fig. 3 shows the changes of θ_{h} with the sampling interval h. It can be seen that θ_{h} decreases as h increases. Box *et al.* [12, pp. 496–497] have already showed that EWMA statistic provides the minimum mean square error forecast for an IMA(1,1) process and the optimal EWMA controller gain is $\lambda = 1 - \theta$. Therefore, the optimal EWMA controller tuning parameter increases as the production frequency decreases for IMA(1,1) process.

Given any arbitrary covariance or correlation sequence with only a finite number of nonzero elements, there is a finite moving average process corresponding to the sequence [14, pp. 224–225]. Hence an ARIMA(p,1,q) series can be approximated by an IMA$(1, \mathrm{q}')$ series

$$(1-\mathrm{B})\mathrm{N[t]} = \left(1-\theta_1\mathrm{B}-\theta_2\mathrm{B}^2-\cdots-\theta_{\mathrm{q}'}\mathrm{B}^{\mathrm{q}'}\right)\varepsilon[\mathrm{t}]. \qquad (10)$$

IEEE Proof Web Version

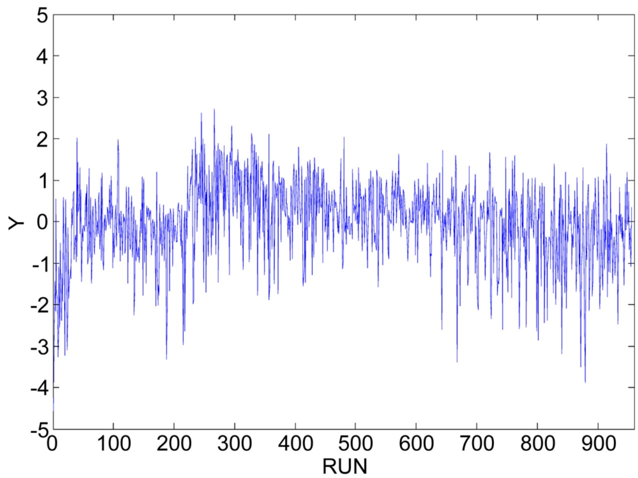

Fig. 4. Plant response under threaded EWMA controller.

If this series is sampled at intervals of h units and $\mathrm{h} > q'-1$, then the resulting process can be represented by an IMA(1,1) series [13]

$$(1-\mathrm{B})\mathrm{M_h}[\mathrm{t}] = (1-\theta_\mathrm{h}\mathrm{B})\varepsilon_\mathrm{h}[\mathrm{t}]. \tag{11}$$

Hence, our conclusion on IMA(1,1) processes can be extended to other nonstationary disturbances.

The above analysis was performed for production at regular intervals. However, in actual production, occurrence of each thread is random unless special attention is given to scheduling. Furthermore, the results are obtained using asymptotic mean square analysis. In actual production, when production frequency is low, the entire quota of a minority product may not contain enough samples for statistical significant analysis.

V. Industrial Example

In this section, an industrial data set from wafer etching process is used to test the effectiveness of the proposed algorithm. The example includes 956 runs. There are 32 products fabricated on two tools and 59 threads all together. The data were standardized to have overall mean zero and overall variance of one. The wafer etching production data were originally under the control of thread EWMA algorithm. The controlled output is plotted in Fig. 4.

The original tuning parameter of EWMA controllers for all threads is 0.2. The disturbance on tool 1 can be roughly approximated as an IMA(1,1) process $(1-B)N_t = (1-0.83B)a_t$ if we neglect the variability caused by different products. A thread of tool 1 with the largest counts number is used as a reference, and performance assessment procedures proposed by Prabhu *et al.* [11] are implemented. The optimal tuning parameter calculated for this reference product is $\lambda_{\mathrm{opt,ref}} = 0.31$. Hence, the parameter is estimated to be $\theta_{\mathrm{h,ref}} = 0.79$. The tuning parameters of other threads are adjusted according to their frequency ratio with respect to the reference component

$$\frac{\mathrm{h_i}}{\mathrm{h_{ref}}} = \frac{(1-\theta_{\mathrm{h,i}})^2\theta_{\mathrm{h,ref}}}{(1-\theta_{\mathrm{h,ref}})^2\theta_{\mathrm{h,i}}} \tag{12}$$

$$\lambda_\mathrm{i} = 1-\theta_{\mathrm{h,i}}. \tag{13}$$

Similar tuning procedures are also applied to tool 2. However, it should be noted that the tuning parameters for the threads whose fabricating counts are less then a certain number, e.g., 15, remain unchanged. This is due to the fact that the statistical properties of the stochastic process may not hold when there are few samples. The stimulated retuned process output is plotted in Fig. 5 and the standard deviation of the output is 0.89, about 10% improvement from the current plant response. The best achievable performance of the industrial data is 0.85 (the optimal EWMA controller gains are obtained by an exhaustive search method).

VI. Conclusion

In this paper, the effect of production frequency on optimal tuning of threaded EWMA controller in a high-mixed production system was studied. Simulation shows that, for stationary disturbances, the optimal EWMA controller gain decreases as the production frequency decreases. For nonstationary disturbances, the EWMA controller gain increases as the production frequency decreases. The conclusions are supported by resampled time-series analysis at fixed intervals. Different semiconductor manufacturing processes have different tool dynamics. Processes with rapid tool wear, such as chemical mechanical polishing, are likely to have a tool disturbance nonstationary process. Processes such as photolithography and overlay may be more stationary. According to our study reported in this paper, the tuning guidelines for products with varied lot count will be different for different processes.

Appendix

For the control system described in (2)–(4), the EWMA controller is equivalent to a discrete integral controller with gain $\mathrm{k_i} = \lambda/\mathrm{b}$ from the viewpoint of control engineering. If the process target is set equal to zero, the output y can be expressed as

$$\mathrm{y}[\mathrm{t}] = \frac{1-\mathrm{B}}{1-\kappa\mathrm{B}}\mathrm{N}[\mathrm{t}] \tag{A1}$$

where $\kappa = 1-\lambda\beta/\mathrm{b}$. Taking long division, we get

$$\mathrm{y}[\mathrm{t}] = \sum_{\mathrm{j}=0}^{\infty}\psi_\mathrm{j}\mathrm{N_{t-j}} \tag{A2}$$

with

$$\psi_0 = 1, \quad \psi_\mathrm{j} = \kappa^{\mathrm{j}-1}(\kappa-1), \quad \mathrm{j} \geq 1.$$

Hence the variance of the output is

$$\sigma_\mathrm{y}^2 = \left(\sum_{\mathrm{j}=0}^{\infty}\sum_{\mathrm{k}=0}^{\infty}\psi_\mathrm{j}\psi_\mathrm{k}\rho_{|\mathrm{j}-\mathrm{k}|}\right)\sigma_{\mathrm{N_t}}^2$$
$$= \left(\sum_{\mathrm{j}=0}^{\infty}\psi_\mathrm{j}^2 + 2\psi_0\sum_{\mathrm{k}=1}^{\infty}\psi_\mathrm{k}\rho_\mathrm{k}\right.$$

IEEE Proof Web Version

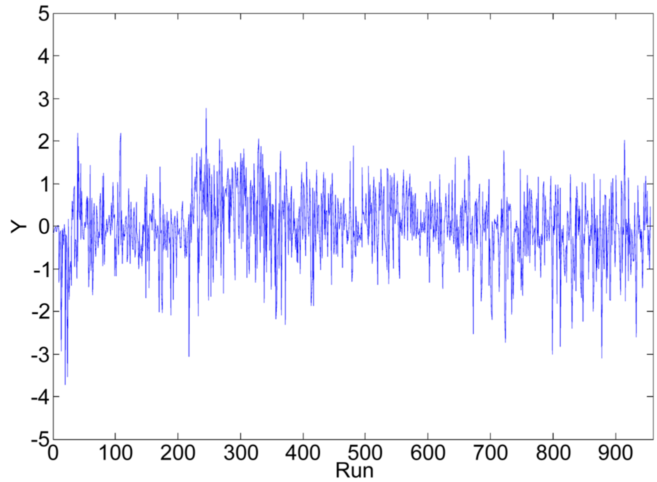

Fig. 5. Simulated response under the control of retuned threaded EWMA controller.

$$+2\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\psi_j\psi_{j+k}\rho_k\Bigg)\,\sigma_{N_t}^2. \tag{A3}$$

For ARMA(1,1) process $(1-\phi B)N[t]=(1-\theta B)\varepsilon[t]$, we have

$$\rho_j=\phi^{j-1}\rho_1,\quad j\geq 1. \tag{A4}$$

By substituting ψ_j and (A4) into (A3), we have

$$\sigma_y^2=\left(\frac{2}{1+\kappa}-2\rho_1\frac{1-\kappa}{(1-\phi\kappa)(1+\kappa)}\right)\sigma_{N_t}^2. \tag{A5}$$

References

[1] J. Moyne, *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2001.
[2] E. D. Castillo, *Statistical Process Adjustment for Quality Control*. New York: Wiley, 2002.
[3] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Trans. Semicond. Manuf.*, vol. 8, pp. 26–43, 1995.
[4] N. S. Patel and S. T. Jenkins, "Adaptive optimization of run-to-run controllers: The EWMA example," *IEEE Trans. Semicond. Manuf.*, vol. 13, pp. 97–107, 2000.
[5] S. T. Tseng, A. B. Yeh, F. Tsung, and Y. Y. Chan, "A study of variable EWMA controller," *IEEE Trans. Semicond. Manuf.*, vol. 16, pp. 633–643, 2003.
[6] T. H. Smith and D. S. Boning, "Artificial neural network exponentially weighted moving average controller for semiconductor processes," *J. Vac. Sci. Technol.*, vol. 15, pp. 236–239, 1997.
[7] Y. Zheng, Q. H. Lin, D. S. H. Wong, S. S. Jang, and K. Hui, "Stability and performance analysis of mixed product run-to-run control," *J. Process Contr.*, vol. 16, pp. 431–443, 2006.
[8] S. K. Firth, W. J. Campbell, A. Toprac, and T. F. Edgar, "Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes," *IEEE Trans. Semicond. Manuf.*, vol. 19, pp. 298–315, 2006.
[9] A. J. Pasadyn and T. F. Edgar, "Observability and state estimation for multiple product control in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 18, pp. 592–604, 2005.
[10] M. D. Ma, C. C. Chang, D. S. H. Wong, and S. S. Jang, "Identification of tool and product effects in a mixed product and parallel tool environment," *J. Process Contr.*, 2008, accepted for publication.
[11] A. V. Prabhu, T. F. Edgar, and R. Chong, "Performance assessment of run-to-run EWMA controllers," in *Proc. Int. Symp. Adv. Contr. Chem. Process.*, Brazil, 2006, pp. 1127–1132.
[12] G. E. P. Box, G. M. Jenkins, and G. Reinsel, *Time Series Analysis Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.
[13] J. F. MacGregor, "Optimal choice of the sampling interval for discrete process control," *Technometrics*, vol. 18, pp. 151–160, 1976.
[14] T. W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.
[15] T. J. Harris, "Assessment of control loop performance," *Can. J. Chem. Eng.*, vol. 67, pp. 856–861, 1989.
[16] S. J. Qin, "Control performance monitoring—A review and assessment," *Comput. Chem. Eng.*, vol. 23, pp. 173–186, 1998.
[17] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*. Englewood Cliffs, NJ: Wiley, 2006.

Ming-Da Ma photograph and biography not available at the time of publication.

Chun-Cheng Chang, photograph and biography not available at the time of publication.

David Shan-Hill Wong, photograph and biography not available at the time of publication.

Shi-Shang Jang, photograph and biography not available at the time of publication.

IEEE Proof
Web Version

Threaded EWMA Controller Tuning and Performance Evaluation in a High-Mixed System

Ming-Da Ma, Chun-Cheng Chang, David Shan-Hill Wong, and Shi-Shang Jang

***Abstract*—The exponentially weighted moving average (EWMA) controller is a very popular run-to-run (RtR) control scheme in the semiconductor industry. However, in any typical step of semiconductor process, many different products are produced on parallel tools. RtR control is usually implemented with a "threaded" control framework, i.e., different controllers are used for different combinations of tools and products. In this paper, the problem of EWMA controller tuning and performance evaluation in a mixed product system are investigated by simulation and time-series analysis. It was found that as the product frequency changed, the tuning guidelines of a threaded EWMA controller were different for different types of tool disturbances. For a stationary ARMA(1,1) noise, the tuning parameter λ should be decreased as product frequency decreases. If the tool exhibits nonstationary tool dynamics, e.g., ARIMA(1,1,1) noise, the tuning parameter should increase as the product frequency decreases.**

***Index Terms*—Exponentially weighted moving average (EWMA), mixed product system, performance evaluation, time series.**

I. Introduction

In the last two decades, run-to-run (RtR) control that combines statistical process control (SPC) and feedback control has been widely used in the semiconductor manufacturing industry [1], [2], [17]. RtR control adjusts process recipes or inputs from run to run to compensate for various process disturbances to maintain the process output close to a given target. RtR control can efficiently improve the product yield and throughout and reduce scrap, rework, and cycle time.

Exponentially weighed moving average (EWMA) controller is the most popular RtR control scheme in the semiconductor industry [3]. There are many different kinds of EWMA algorithms [4]–[6]. However, all of the above assume that only a single product is fabricated on a tool.

In the semiconductor manufacturing industry, production resembles an automated assembly line in which many similar products with slightly different specifications are manufactured step-by-step on a number of different tools. This constitutes a high-mix production system. RtR control is commonly implemented with a "threaded" framework. In this approach, each specific combination of tool and product is called a thread. Each thread has its own controller. Threaded EWMA control is probably the most popular control architecture, although other frameworks were proposed [8]–[10]. One key advantage of this approach is that the stability of the control system is guaranteed [7]. On the other hand, some commonly encountered questions are whether the controller performance is optimal and how should each thread be tuned differently as the product frequency changes. Controller performance evaluation has seen active development in research and gradual acceptance in industry since the original work of Harris [15]. Recent developments in this subject were reviewed by Qin [16]. Prabhu *et al.* [11] proposed a performance evaluation algorithm for an EWMA controller of a single thread. In this paper, we shall use the performance evaluation technique and time-series analysis to investigate the optimal tuning of a threaded EWMA controller in a high-mixed system.

Manuscript received November 09, 2007; revised June 18, 2009; June 22, 2009; accepted June 22, 2009.

M.-D. Ma is with the Center for Control and Guidance Technology, Harbin Institute of Technology, Harbin 150001, China.

C.-C. Chang, D. S.-H. Wong, and S.-S. Jang are with the Department of Chemical Engineering, National Tsing-Hua University, Hsin-Chu 30043, Taiwan (e-mail: ssjang@mx.nthu.edu.tw).

Color versions of one or more figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSM.2009.2028219

II. Threaded EWMA Control of a Mixed Product Plant

Consider a static simple linear single-input single-output process performed in a mixed run situation on a single tool

$$\mathrm{y}[\mathrm{t}] = \alpha_{\mathrm{k}[\mathrm{t}]} + \beta_{\mathrm{k}[\mathrm{t}]}\mathrm{u}[\mathrm{t}] + \mathrm{N}[\mathrm{t}] \tag{1}$$

where $\mathrm{y}[\mathrm{t}]$ and $\mathrm{u}[\mathrm{t}]$ denote values of output and manipulated variable used on the tth run on the tool. $\alpha_{\mathrm{k}[\mathrm{t}]}$ is offset or bias term and $\beta_{\mathrm{k}[\mathrm{t}]}$ is the static gain term associated with the product produced on the tth run on the tool. Let us assume that they are relatively independent of time. $\mathrm{N}[\mathrm{t}]$ is a stochastic noise process associated with the tool. In a threaded approach, a sequence of output and input for each specific product is resampled

$$\mathrm{y}[\mathrm{t_k}] = \alpha_{\mathrm{k}} + \beta_{\mathrm{k}}\mathrm{u}[\mathrm{t_k}] + \mathrm{N_k}[\mathrm{t_k}] \tag{2}$$

where $\mathrm{t_k}$ is an index of the number of runs making the kth product that have been carried out. Given a process model $\mathrm{y} = \mathrm{b_k u} + \mathrm{a_k}$ for each product, the offset term can be estimated by EWMA filter

$$\mathrm{a_k}[\mathrm{t_k}] = \lambda\left(\mathrm{y}[\mathrm{t_k}] - \mathrm{bu}[\mathrm{t_k}]\right) + (1-\lambda)\mathrm{a_k}[\mathrm{t_k}-1]. \tag{3}$$

The control action is

$$\mathrm{u}[\mathrm{t_k}+1] = \frac{\mathrm{T_k} - \mathrm{a_k}[\mathrm{t_k}]}{\mathrm{b}}. \tag{4}$$

IEEE Proof Print Version

0894-6507/$26.00 © 2009 IEEE

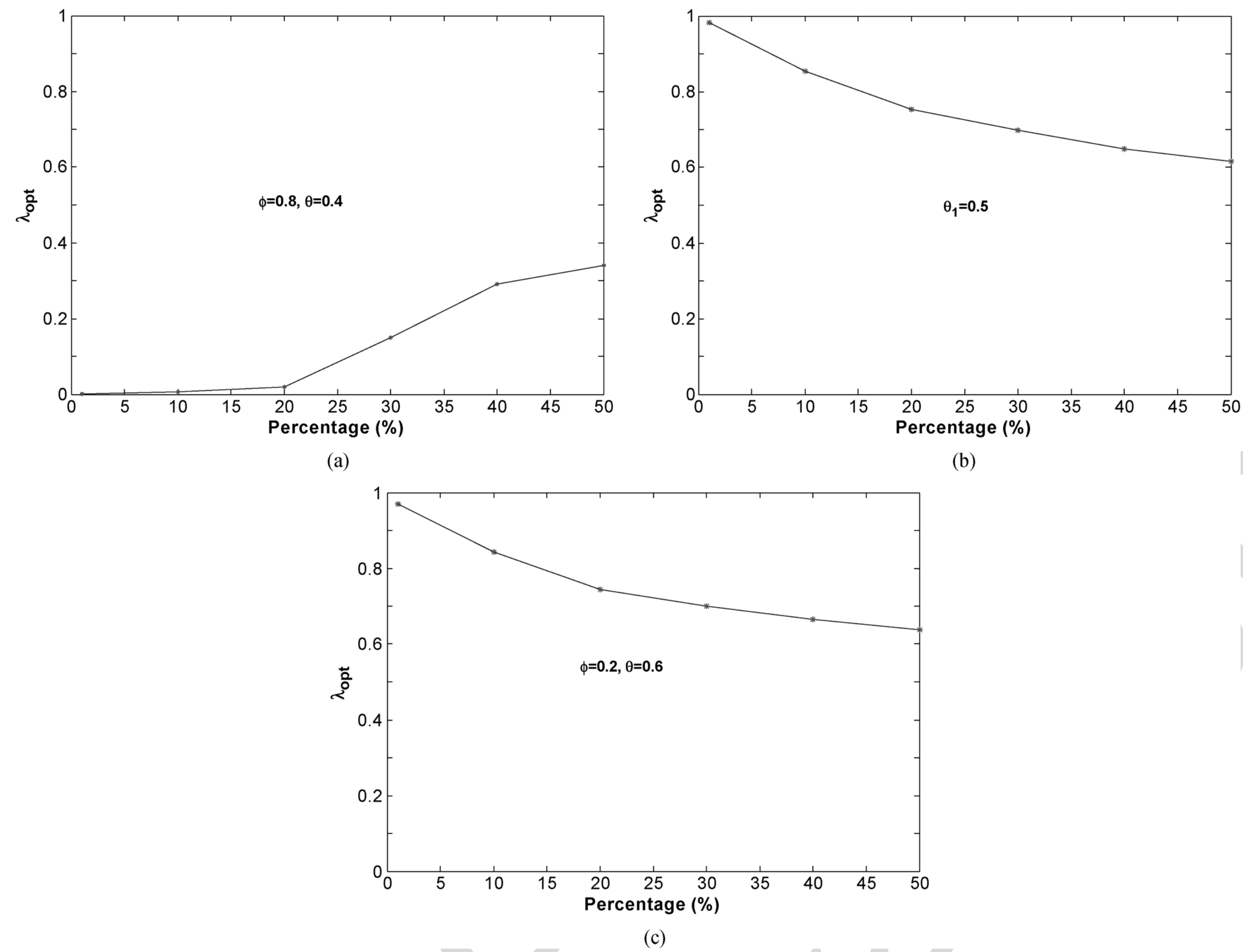

Fig. 1. Effect of production frequency on optimal tuning parameter λ. (a) ARMA(1,1) disturbance. (b) IMA(1,1) disturbance. (c) ARIMA(1,1,1) disturbance.

Note that the threaded EWMA control is similar to a single product EWMA control, except the disturbance experienced is not the actual change in tool condition $\mathrm{N}[\mathrm{t}]$ from run to run but a resampled series $\mathrm{N_k}[\mathrm{t_k}]$.

III. Performance Evaluation Using Simulations

Performance assessment is widely implemented in process control now. Usually, minimum variance (MV) is adopted as the benchmark. However, MV may not be easily achieved. Therefore, best achievable performance (BAP) under the current controller is often considered. Prabhu *et al.* [11] proposed a performance assessment method for EWMA control to obtain BAP estimates and optimal λ using closed-loop response data. We shall use this method to evaluate the performance of the mixed product system under threaded EWMA control to investigate how the optimal λ changes as the product frequency change.

A simulation example consisting of one tool and three products 1, 2, 3 is used. Each product is assigned a unique bias. The plant gain is assumed to be known. Three disturbances are considered: ARMA(1,1) (first-order autoregressive moving average), IMA(1,1) (first-order integrated moving average), and ARIMA(1,1,1) (first-order integrated autoregressive-moving average) process, respectively:

Noise I: $(1-\phi \mathrm{B})\mathrm{N}[\mathrm{t}] = (1-\theta \mathrm{B})\varepsilon[\mathrm{t}]$ $\phi = 0.8, \theta = 0.4$;
Noise II: $(1-\mathrm{B})\mathrm{N}[\mathrm{t}] = (1-\theta \mathrm{B})\varepsilon[\mathrm{t}]$ $\theta = 0.5$;
Noise III: $(1-\phi \mathrm{B})(1-\mathrm{B})\mathrm{N}[\mathrm{t}] = (1-\theta \mathrm{B})\varepsilon[\mathrm{t}]$ $\phi = 0.2$, $\theta = 0.6$.

Here "B" stands for backward shift operator. In the simulation, the percentage of product 1 changes from 1% to 50%, product 2 is 30%, and product 3 makes up the rest of the products. For three types of disturbances, 400 sets of disturbance sequences are generated. In each set, the results of at least 100 runs of every product are collected. The production schedule is random. A value of $\lambda = 0.2$ for all threads is used to obtain the closed-loop data. The changes in estimated optimal tuning parameter for thread (product) 1, $\lambda_{1,\mathrm{opt}}$ which is the average value obtained from the above 400 sets of disturbance sequences, are shown in Fig. 1. It is interesting to find out that the tuning rules for the above disturbances are different. For Noise I, which is a stationary ARMA(1,1) noise, the optimal tuning parameter λ decreases as the percentage of the product decreases, as shown in Fig. 1(a). However, for the other two nonstationary noises II and III, the optimal tuning parameter λ increases as the percentage of the product decreases. It is interesting, therefore, to examine if the trends for nonstationary disturbances are general, as shown in Fig. 1(b) and 1(c) using time-series analysis.

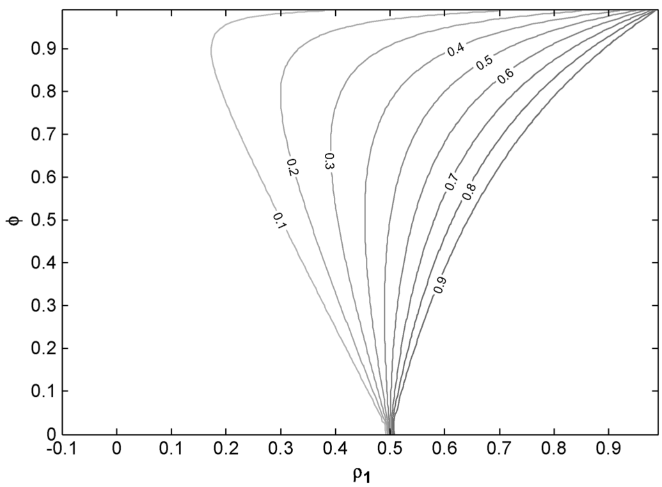

Fig. 2. Contour plots of optimal $\lambda\xi$ for ARMA(1,1) disturbance processes.

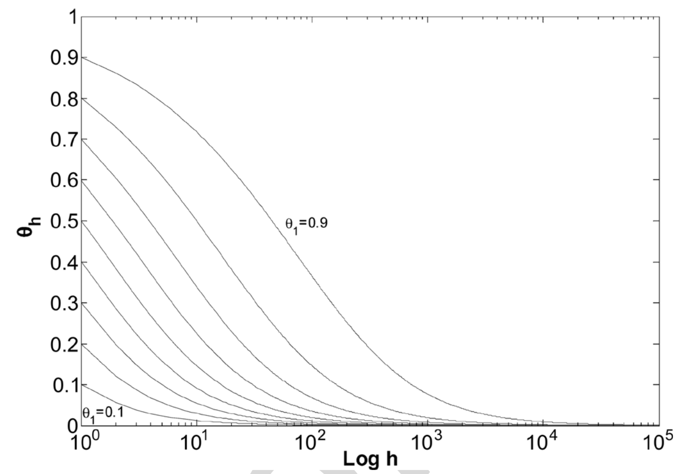

Fig. 3. Changes in parameter θ_{h} when resampling an IMA(1,1) process with parameter θ_1.

IV. Time-Series Analysis

For a simple gain process controlled by an EWMA controller with a single product, the output variance for an ARMA(1,1) disturbance $\mathrm{N[t]}$ is given by (see the Appendix)

$$\sigma_{\mathrm{y}}^2 = \left(\frac{2}{1+\kappa} - 2\rho_1 \frac{1-\kappa}{(1-\phi\kappa)(1+\kappa)}\right)\sigma_{\mathrm{N_t}}^2 \tag{5}$$

where $\kappa = 1 - \lambda\xi$, $\xi = \beta/\mathrm{b}$, and ρ_1 is the autocorrelation of the disturbance $\mathrm{N[t]}$ at lag 1. The optimal $\lambda\xi$ given ϕ and ρ_1 can be obtained by differentiating the above equations with respect to κ.

Fig. 2 is a contour plot of optimal $\lambda\xi$ at different values of ϕ and ρ_1. In the following analysis, it is assumed that $\xi = 1$. When plant model mismatch is considered, we can simply replace λ with $\lambda\xi$. It should be noted that the optimal value of tuning parameter λ may be different when plant model mismatch exists, but ξ does not affect the changing trend of the optimal EWMA tuning parameter. In other words, the plant model mismatch does not affect the conclusion of this paper.

If the product is manufactured at regular intervals of every h runs, with h being a positive integer, then its threaded EWMA controller will face a disturbance that is a resampled sequence of the original sequence at every hth time point. If an ARMA(1,1) process with parameters (ϕ, θ) is resampled, the resulting process $\mathrm{M_h[t]}$ is also an ARMA(1,1) process [13]

$$(1 - \phi_{\mathrm{h}}\mathrm{B})\mathrm{M_h[t]} = (1 - \theta_{\mathrm{h}}\mathrm{B})\varepsilon_{\mathrm{h}}[\mathrm{t}] \tag{6}$$

with

$$\phi_{\mathrm{h}} = \phi^{\mathrm{h}} \quad \rho_1(\mathrm{h}) = \phi^{\mathrm{h}-1}\rho_1 \tag{7}$$

where $\varepsilon_{\mathrm{h}}[\mathrm{t}]$ is a new white noise process with σ_{h} and $\mathrm{h}\rho_1(\mathrm{h})$ is the autocorrelation of $\mathrm{M_h[t]}$. Since both ϕ_{h} and $\rho_1(\mathrm{h})$ decrease with h, from Fig. 2, we can infer that optimal λ must decrease with production frequency when the tool disturbance is an ARMA(1,1) noise.

If a process is stationary and the samples are taken less frequently in time, the autocorrelation of the sampled data will decrease. When the sampling interval is sufficiently large, the data will appear to be uncorrelated. It is well known that when the disturbance is white noise, no control action should be added, i.e., $\lambda = 0$. Therefore, we can draw the conclusion that, for stationary disturbances, the optimal EWMA controller gain decreases as percentage of the products decreases.

If an IMA(1,1) process $\mathrm{N[t]}$ is sampled at every h time point, the resulting process is denoted by $\mathrm{M_h[t]}$ and is also an IMA(1,1) process

$$(1 - \mathrm{B})\mathrm{M_h[t]} = (1 - \theta_{\mathrm{h}}\mathrm{B})\eta_{\mathrm{h}}[\mathrm{t}] \tag{8}$$

with

$$\frac{\mathrm{h}(1-\theta_1)^2}{\theta_1} = \frac{(1-\theta_{\mathrm{h}})^2}{\theta_{\mathrm{h}}} \tag{9}$$

and $\eta_{\mathrm{h}}[\mathrm{t}]$ is a new white noise process [12, pp. 526–528]. Fig. 3 shows the changes of θ_{h} with the sampling interval h. It can be seen that θ_{h} decreases as h increases. Box *et al.* [12, pp. 496–497] have already showed that EWMA statistic provides the minimum mean square error forecast for an IMA(1,1) process and the optimal EWMA controller gain is $\lambda = 1 - \theta$. Therefore, the optimal EWMA controller tuning parameter increases as the production frequency decreases for IMA(1,1) process.

Given any arbitrary covariance or correlation sequence with only a finite number of nonzero elements, there is a finite moving average process corresponding to the sequence [14, pp. 224–225]. Hence an ARIMA(p,1,q) series can be approximated by an $\mathrm{IMA}(1, \mathrm{q}')$ series

$$(1 - \mathrm{B})\mathrm{N[t]} = \left(1 - \theta_1\mathrm{B} - \theta_2\mathrm{B}^2 - \cdots - \theta_{\mathrm{q}'}\mathrm{B}^{\mathrm{q}'}\right)\varepsilon[\mathrm{t}]. \tag{10}$$

IEEE Proof Print Version

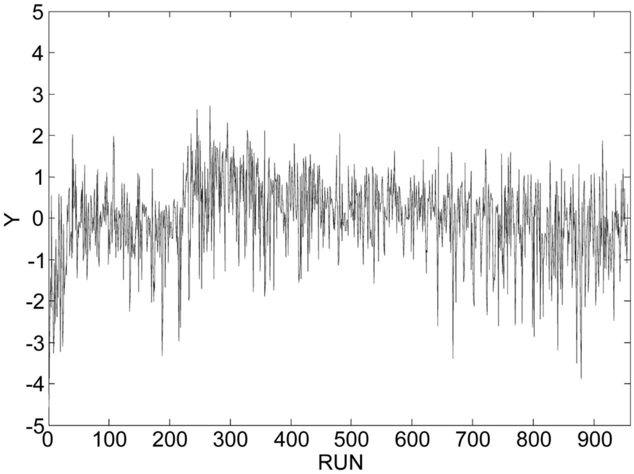

Fig. 4. Plant response under threaded EWMA controller.

If this series is sampled at intervals of h units and $\mathrm{h} > q'-1$, then the resulting process can be represented by an IMA(1,1) series [13]

$$(1-\mathrm{B})\mathrm{M_h}[\mathrm{t}] = (1-\theta_{\mathrm{h}}\mathrm{B})\varepsilon_{\mathrm{h}}[\mathrm{t}]. \tag{11}$$

Hence, our conclusion on IMA(1,1) processes can be extended to other nonstationary disturbances.

The above analysis was performed for production at regular intervals. However, in actual production, occurrence of each thread is random unless special attention is given to scheduling. Furthermore, the results are obtained using asymptotic mean square analysis. In actual production, when production frequency is low, the entire quota of a minority product may not contain enough samples for statistical significant analysis.

V. Industrial Example

In this section, an industrial data set from wafer etching process is used to test the effectiveness of the proposed algorithm. The example includes 956 runs. There are 32 products fabricated on two tools and 59 threads all together. The data were standardized to have overall mean zero and overall variance of one. The wafer etching production data were originally under the control of thread EWMA algorithm. The controlled output is plotted in Fig. 4.

The original tuning parameter of EWMA controllers for all threads is 0.2. The disturbance on tool 1 can be roughly approximated as an IMA(1,1) process $(1-B)N_t = (1-0.83B)a_t$ if we neglect the variability caused by different products. A thread of tool 1 with the largest counts number is used as a reference, and performance assessment procedures proposed by Prabhu *et al.* [11] are implemented. The optimal tuning parameter calculated for this reference product is $\lambda_{\mathrm{opt,ref}} = 0.31$. Hence, the parameter is estimated to be $\theta_{\mathrm{h,ref}} = 0.79$. The tuning parameters of other threads are adjusted according to their frequency ratio with respect to the reference component

$$\frac{\mathrm{h_i}}{\mathrm{h_{ref}}} = \frac{(1-\theta_{\mathrm{h,i}})^2\theta_{\mathrm{h,ref}}}{(1-\theta_{\mathrm{h,ref}})^2\theta_{\mathrm{h,i}}} \tag{12}$$

$$\lambda_{\mathrm{i}} = 1-\theta_{\mathrm{h,i}}. \tag{13}$$

Similar tuning procedures are also applied to tool 2. However, it should be noted that the tuning parameters for the threads whose fabricating counts are less then a certain number, e.g., 15, remain unchanged. This is due to the fact that the statistical properties of the stochastic process may not hold when there are few samples. The stimulated retuned process output is plotted in Fig. 5 and the standard deviation of the output is 0.89, about 10% improvement from the current plant response. The best achievable performance of the industrial data is 0.85 (the optimal EWMA controller gains are obtained by an exhaustive search method).

VI. Conclusion

In this paper, the effect of production frequency on optimal tuning of threaded EWMA controller in a high-mixed production system was studied. Simulation shows that, for stationary disturbances, the optimal EWMA controller gain decreases as the production frequency decreases. For nonstationary disturbances, the EWMA controller gain increases as the production frequency decreases. The conclusions are supported by resampled time-series analysis at fixed intervals. Different semiconductor manufacturing processes have different tool dynamics. Processes with rapid tool wear, such as chemical mechanical polishing, are likely to have a tool disturbance nonstationary process. Processes such as photolithography and overlay may be more stationary. According to our study reported in this paper, the tuning guidelines for products with varied lot count will be different for different processes.

Appendix

For the control system described in (2)–(4), the EWMA controller is equivalent to a discrete integral controller with gain $\mathrm{k_i} = \lambda/\mathrm{b}$ from the viewpoint of control engineering. If the process target is set equal to zero, the output y can be expressed as

$$\mathrm{y}[\mathrm{t}] = \frac{1-\mathrm{B}}{1-\kappa\mathrm{B}}\mathrm{N}[\mathrm{t}] \tag{A1}$$

where $\kappa = 1-\lambda\beta/\mathrm{b}$. Taking long division, we get

$$\mathrm{y}[\mathrm{t}] = \sum_{\mathrm{j}=0}^{\infty}\psi_{\mathrm{j}}\mathrm{N}_{\mathrm{t-j}} \tag{A2}$$

with

$$\psi_0 = 1, \quad \psi_{\mathrm{j}} = \kappa^{\mathrm{j}-1}(\kappa-1), \quad \mathrm{j} \geq 1.$$

Hence the variance of the output is

$$\begin{aligned}\sigma_{\mathrm{y}}^2 &= \left(\sum_{\mathrm{j}=0}^{\infty}\sum_{\mathrm{k}=0}^{\infty}\psi_{\mathrm{j}}\psi_{\mathrm{k}}\rho_{|\mathrm{j-k}|}\right)\sigma_{\mathrm{N_t}}^2 \\ &= \left(\sum_{\mathrm{j}=0}^{\infty}\psi_{\mathrm{j}}^2 + 2\psi_0\sum_{\mathrm{k}=1}^{\infty}\psi_{\mathrm{k}}\rho_{\mathrm{k}}\right.\end{aligned}$$

IEEE Proof Print Version

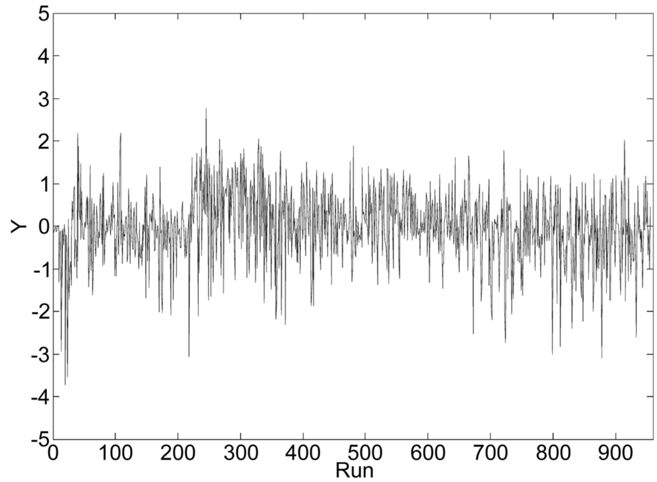

Fig. 5. Simulated response under the control of retuned threaded EWMA controller.

$$+2\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\psi_j\psi_{j+k}\rho_k\Bigg)\sigma^2_{N_t}. \qquad (A3)$$

For ARMA(1,1) process $(1-\phi B)N[t]=(1-\theta B)\varepsilon[t]$, we have

$$\rho_j=\phi^{j-1}\rho_1, \quad j\geq 1. \qquad (A4)$$

By substituting ψ_j and (A4) into (A3), we have

$$\sigma^2_y=\left(\frac{2}{1+\kappa}-2\rho_1\frac{1-\kappa}{(1-\phi\kappa)(1+\kappa)}\right)\sigma^2_{N_t}. \qquad (A5)$$

REFERENCES

[1] J. Moyne, *Run-to-Run Control in Semiconductor Manufacturing*. Boca Raton, FL: CRC Press, 2001.
[2] E. D. Castillo, *Statistical Process Adjustment for Quality Control*. New York: Wiley, 2002.
[3] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Trans. Semicond. Manuf.*, vol. 8, pp. 26–43, 1995.
[4] N. S. Patel and S. T. Jenkins, "Adaptive optimization of run-to-run controllers: The EWMA example," *IEEE Trans. Semicond. Manuf.*, vol. 13, pp. 97–107, 2000.
[5] S. T. Tseng, A. B. Yeh, F. Tsung, and Y. Y. Chan, "A study of variable EWMA controller," *IEEE Trans. Semicond. Manuf.*, vol. 16, pp. 633–643, 2003.
[6] T. H. Smith and D. S. Boning, "Artificial neural network exponentially weighted moving average controller for semiconductor processes," *J. Vac. Sci. Technol.*, vol. 15, pp. 236–239, 1997.
[7] Y. Zheng, Q. H. Lin, D. S. H. Wong, S. S. Jang, and K. Hui, "Stability and performance analysis of mixed product run-to-run control," *J. Process Contr.*, vol. 16, pp. 431–443, 2006.
[8] S. K. Firth, W. J. Campbell, A. Toprac, and T. F. Edgar, "Just-in-time adaptive disturbance estimation for run-to-run control of semiconductor processes," *IEEE Trans. Semicond. Manuf.*, vol. 19, pp. 298–315, 2006.
[9] A. J. Pasadyn and T. F. Edgar, "Observability and state estimation for multiple product control in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 18, pp. 592–604, 2005.
[10] M. D. Ma, C. C. Chang, D. S. H. Wong, and S. S. Jang, "Identification of tool and product effects in a mixed product and parallel tool environment," *J. Process Contr.*, 2008, accepted for publication.
[11] A. V. Prabhu, T. F. Edgar, and R. Chong, "Performance assessment of run-to-run EWMA controllers," in *Proc. Int. Symp. Adv. Contr. Chem. Process.*, Brazil, 2006, pp. 1127–1132.
[12] G. E. P. Box, G. M. Jenkins, and G. Reinsel, *Time Series Analysis Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1994.
[13] J. F. MacGregor, "Optimal choice of the sampling interval for discrete process control," *Technometrics*, vol. 18, pp. 151–160, 1976.
[14] T. W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.
[15] T. J. Harris, "Assessment of control loop performance," *Can. J. Chem. Eng.*, vol. 67, pp. 856–861, 1989.
[16] S. J. Qin, "Control performance monitoring—A review and assessment," *Comput. Chem. Eng.*, vol. 23, pp. 173–186, 1998.
[17] G. S. May and C. J. Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control*. Englewood Cliffs, NJ: Wiley, 2006.

Ming-Da Ma photograph and biography not available at the time of publication.

Chun-Cheng Chang, photograph and biography not available at the time of publication.

David Shan-Hill Wong, photograph and biography not available at the time of publication.

Shi-Shang Jang, photograph and biography not available at the time of publication.

IEEE Proof Print Version