Available online at www.sciencedirect.com

ScienceDirect

Journal of Process Control 18 (2008) 266–276

JOURNAL OF PROCESS CONTROL

www.elsevier.com/locate/jprocont

On the interaction between measurement strategy and control performance in semiconductor manufacturing

An-Jhih Su [a], Cheng-Ching Yu [a,*], Babatunde A. Ogunnaike [b]

[a] Department of Chemical Engineering, National Taiwan University, Taipei 106-17, Taiwan
[b] Department of Chemical Engineering, University of Delaware, Newark, DE 19716-3110, USA

Received 4 January 2007; accepted 11 July 2007

Abstract

Manufacturing in the high revenue semiconductor industry involves a highly capital intensive process consisting of more than 300 steps. To ensure stable process operation and ultimately meet the exacting requirements on final product quality, the typical advanced IC fabrication process requires many on-line sensors and off-line metrology tools for acquiring process and product information necessary for effective monitoring and control. However, the high cost associated with these measurement devices has made the economics of metrology a major factor in the industry's quest for world-class manufacturing. In this paper, we first introduce various measurement data types and describe how they feature within an ideal fab-wide control architecture; subsequently, and from a process control point of view, we derive various run-to-run controllers, carry out stability analyses, and analyze control system performance. These results are then applied to the problem of rational metrology strategy selection where the effects of various metrology strategies on control system performance are systematically analyzed. In particular, if control performance takes priority over economics, we present results for determining maximum tolerable sampling intervals, maximum tolerable delay, and measurement priority.

Keywords: Semiconductor manufacturing; Measurement strategy; Controller design; Run-to-run control

1. Introduction

Semiconductor manufacturing is a highly capital intensive venture with high revenue, where maintaining competitiveness demands high equipment efficiency, tight quality control, and a relentless and continuous reduction of material cost. A key factor in achieving these goals is the measurement system, the set of sensors and analyzers needed to acquire critical measurements and information about the process and product. Such information, if processed appropriately can be used to create knowledge, increase equipment efficiency and accelerate yield improvement, ultimately generating increased profit.

1.1. Measurements classification

To construct a device on a wafer, the process flow involves several cycles of lithography, etch, chemical vapor deposition (CVD), chemical mechanical polishing (CMP), etc. as shown schematically in Fig. 1 [19,17]; and many modern microelectronic fabrication plants ("fabs" for short) are equipped with world-class information technology (IT) infrastructure for collecting and storing lots of process and product information. Such data is typically used for fault diagnosis and classification, or to provide information as feedback in run-to-run control to maximize operation efficiency of the equipment. The measurement data collected in the typical fab may be classified as follows:

(1) *Real-time equipment data*: typically from an in situ sensor, and used as a feedback signal for on-line control (for example, on-line temperature measurements can

[*] Corresponding author. Tel.: +886 2 3366 3037; fax: +886 2 2362 3040. *E-mail address:* ccyu@ntu.edu.tw (C.-C. Yu).

0959-1524/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2007.07.005

Notations

CMP	chemical mechanical polishing
D_{eff}	effective delay
d_t	disturbance sequence in production domain
d_t^*	disturbance sequence in sampled-run domain
G_{C}	controller transfer function
G_{P}	process transfer function
j	imaginary root
IMC	internal model control
IMA	integrated-moving-average time series
N_{m}	measurement delay
N_{s}	sampling interval
K_{P}	process gain
$\widehat{K}_{\mathrm{P}}$	estimated process gain
K_{F}	forward loop gain
$K_{\mathrm{F,opt}}$	optimal value of K_{F}
k	ramp slope
PI^2	proportional-plus-double-integral controller
q^{-1}	discrete domain (backward operator)
S	sensitivity function
s	Laplace domain
y_t	quality sequence in production domain
y_t^*	quality sequence in sampled-run domain
a_t	white noise sequence in production domain
a_t^*	white noise sequence in sampled-run domain
τ_{f}	time constant of IMC filter
θ	IMA coefficient
θ^*	IMA coefficient in sampled domain with N_{s} sampling interval
σ^2	sequence variance
ξ	process gain mismatch ($\xi = K_{\mathrm{P}}/\widehat{K}_{\mathrm{P}}$)
ω	frequency
ω_{u}	ultimate frequency

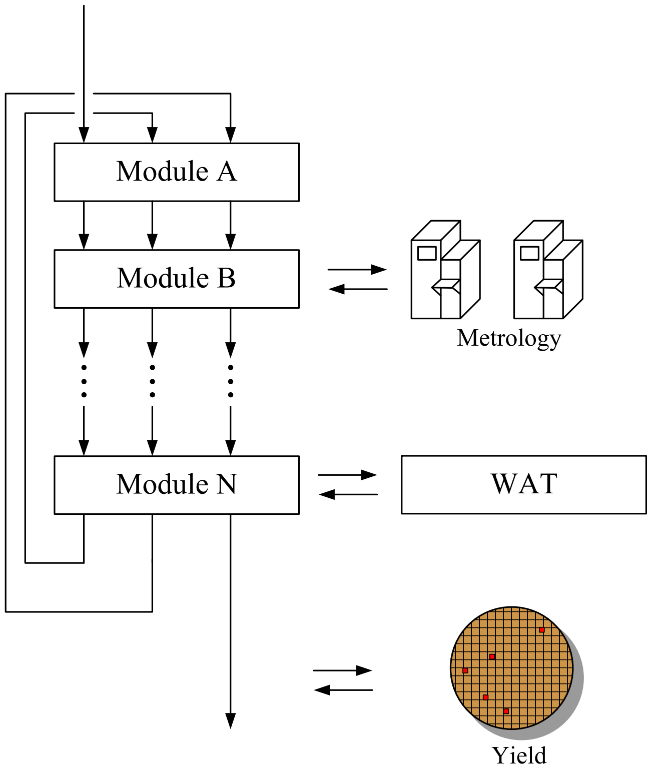

Fig. 1. Multilayer configuration leading to repetitive characteristics in manufacturing. It uses the same module to execute the same function with the metrology tool(s) shared by same module; the WAT test is performed at the completion of each layer; yield is determined toward the end of the entire process.

be used for set point tracking for a given temperature program in a CVD tool). They also can be used to monitor the condition of the equipment, and to build a "health index" to guide decisions on the preventive maintenance. For instance, measurements of chamber pressure can be monitored to check for the existence of a leak and thus determine when maintenance is needed. Typical sampling time ranges from seconds to milliseconds [18].

(2) *Geometric properties data*: measured by metrology tools. A wide variety of semiconductor manufacturing equipment is used to fabricate wafers with desired geometric properties, such as layer thickness and trench structures. The metrology tools used to determine such properties can be divided into two categories: (i) Integrated metrology tools: these are combined with manufacturing equipment and can be configured to measure the quality of every one, or one out of several wafers, (e.g., one out of 4–6 wafers). Such quality information can be used for feedback control on a wafer-to-wafer basis. Typical sampling time for feedback control ranges from 5 to 10 min. (ii) Stand-alone metrology tools: these measure the quality of several wafers in a lot, batchwise, and therefore can only be used for lot-to-lot feedback control. The lots of wafers are typically delivered to the cluster of stand-alone metrology tools to wait in a queue for measurement (Fig. 1). There will obviously be time delays associated with such measurements. Typical sampling time for lot-to-lot control ranges from hours to one day [4,6].

(3) *Wafer acceptance tests (WAT) data:* provide information on many important electric properties that indicate whether or not the device will function correctly after the completion of a metal layer or a structure (Fig. 1). Because many steps are required to complete the construction of a structure, WAT data are available for feedback only after a long time delay. Typical sampling time ranges from days to a week [14,16,20].

(4) *Yield data:* measure the percentage of acceptable dies in a wafer, a quality measurement directly related to revenue. Note that a wafer contains multiple useable products, called dies, and the number of dies is determined by the wafer diameter, e.g., 8 and 12 in. Yield data are usually not available until almost one month into production [20].

Fig. 2 shows a schematic depiction of the time-scale characteristics and measurement complexity of the various categories of measurement data. The observed yield depends on WAT results which in turn depend on many geometric properties; each geometric property is itself determined by the recipes used in each manufacturing equipment. Fig. 2 also indicates the strong structural relationship between these variables. Consequently, the desired WAT properties are used to design the target of metrology for equipment run-to-run (R2R) control, and this desired metrology target is in turn used to design tool recipes in what is known in the terminology of process control as a cascade control structure. The R2R control loop is designed to compensate for variability in the equipment while the WAT feedback loop sets the correct metrology target for the lower level controller. Both control loops suffer from the same problem: the measurements have significant time delays, causing significant deterioration in the closed-loop stability characteristics and ultimately the controller performance. To address this problem, it is necessary to augment available, but delayed measurements with "virtual metrology" and "virtual WAT" [14], as shown in Fig. 3. By virtual metrology, we mean the estimation of yet unavailable metrology data from available lower level data, such as temperature, pressure, batch time, etc. Thus the "virtual metrology" unit serves as a soft sensor, providing estimated measurements that can be used for inferential control [23]. Similarly, "virtual WAT" implies the estimation of electrical properties from available metrology data. Including the measurement types and control mechanisms, such an ideal fab-wide control structure is shown in Fig. 4. A similar structure has also been proposed in [4], with the crucial difference that the structure in [4] contains no measurement delay compensation mechanisms comparable to the one we are proposing here.

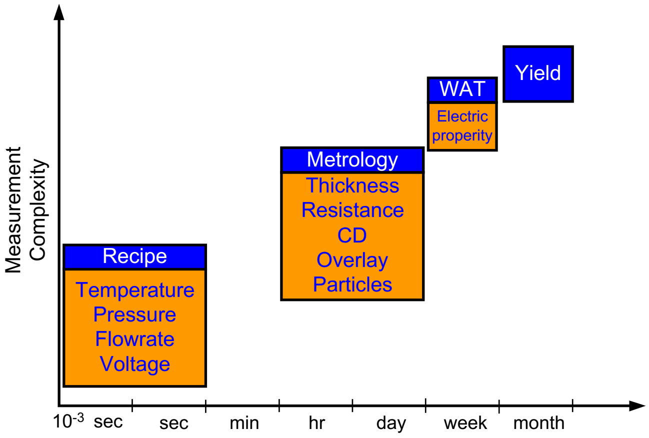

Fig. 2. Measurement complexity and sampling frequencies and data availability times at the recipe, metrology, and WAT levels.

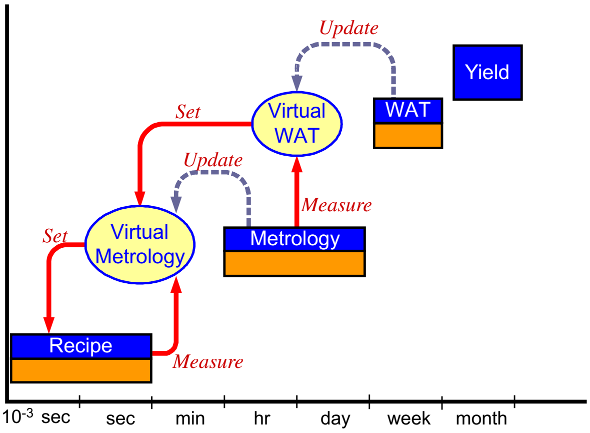

Fig. 3. Coordination between measurements at different time-scales.

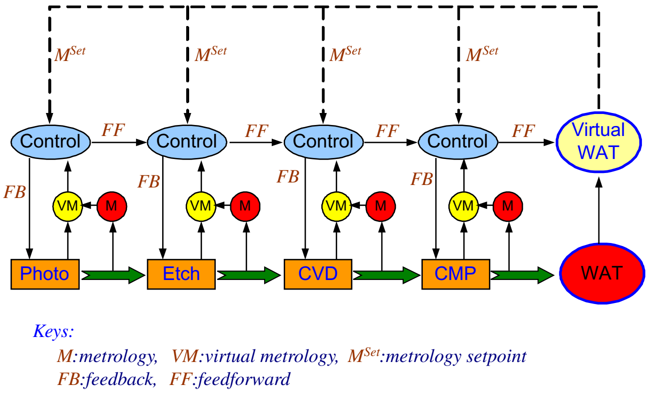

Fig. 4. An ideal fab-wide control architecture.

1.2. Control

The R2R control strategy remains popular in industrial practice and has, for years, attracted a lot of research attention. Del Castillo [2] summarizes results on the design and performance of the most popular R2R controllers: EWMA and double-EWMA. Chen et al. [15] discuss the effects of sequencing on incoming wafers for run-to-run control; and Qin and Good [8] analyze the stability of double-EMWA controllers in the presence of metrology delay, and extended the results to MIMO controllers [9]. Qin [26] also discusses the fab-wide control structure for electrical parameters.

Our current work is motivated by the following questions: (i) what is the effective metrology delay in discrete systems; and (ii) how does the measurement strategy (e.g., sampling interval) influence stability and control performance? Some of these issues have been partially confronted in previous studies. For example, Tseng and Hsu [10] discuss the statistically appropriate number of samples required to obtain an accurate model, and hence guarantee the stability of R2R control; Jula et al. [6] compare the economic impact of in situ, in-line, and off-line metrology systems, while Nurani et al. [7] provide an integrated framework for designing an optimal defect sampling strategy for wafer inspection. Lensing [24] mentions dynamic sampling in metrology to measure just enough to characterize systematic sources of variance and correct them using wafer level control. Moyne [27] discusses financial return-on-investment (ROI) analysis of run-to-run control. However, none of these papers explicitly relate control performance issues to sampling strategies in a comprehensive manner.

1.3. Measurement strategies and controller performance

In order to reduce cost, a common measurement option is to employ one stand-alone metrology tool that is shared by several pieces of equipment. However, this results in metrology delay for R2R control. Under these conditions, determining the maximum tolerable delay that will guarantee control stability and performance is clearly an important issue. Similarly, for equipment with an integrated metrology tool, the critical question is: how frequently should a wafer be sampled for feedback measurement in order to ensure that control performance and high throughput objectives are met?

In the fab, the measurement delay and sampling interval are typically not constant but change depending on many factors. Our goal in this paper is to identify the relationship between measurement delay or sampling interval and the control performance, and illustrate how these results can be used to determine appropriate measurement strategies.

In this study, we consider there are no product-relative terms, since high-mix product problem is very common nowadays [25]. Without these terms, it is more clearly to realize the influence of sampling. The rest of the paper is organized as follows: in Section 2, we define the variables used in this paper, and derive an expression for the effective time delay in a R2R control system; we also discuss how to derive appropriate R2R controllers. In Section 3, we analyze control performance, explicitly dealing with the effects of sampling intervals and effective delay. The application of these results to the rational design of measurement strategies are presented in Section 4 where various concepts including maximum tolerable sampling intervals, tolerable delay, and measurement priority are introduced and discussed. Conclusions follow in Section 5.

2. Effective time delay in R2R control

2.1. Definition

As mentioned earlier, there are two types of basic strategies employed in equipment R2R control: lot-to-lot and wafer-to-wafer. The following analyses of stability and performance can be applied to either type. For simplicity, we use the term "runs" to refer to wafers in the case of wafer-to-wafer control, and lots in the case of lot-to-lot control.

Sampling interval, N_s: the number of runs between samples. For example, the specific value N_s means that after a measured run, the next run to be selected for measurement is the N_s^{th} run; alternatively, that one out of every N_s runs will be measured. One may increase the sampling interval (i.e., making fewer measurements per lot) to achieve higher throughput rate (no processing interruption from measurement delay); however, because the sampling interval has a significant effect on stability and control performance, it should be chosen judiciously. Therefore, it is important to determine the maximum tolerable sampling interval required to guarantee quality. N_s is an integer with a minimum value of 1.

Metrology delay, N_m: the number of runs between when the measurement is made and when the metrology data is available for feedback. The specific value N_m means that data from a measured run will be available at the next N_m^{th} run. Because of this definition, the metrology delay should not only include the measurement time, but also the queue time and material transportation time. It should be noted that even if the measurement can be completed immediately, the result can only be applied to the next run for feedback control. Thus, N_m, also an integer, has a minimum value of 1. For some modules with several sequential processing steps (e.g., washing, baking, main process), it is possible to have several wafers held simultaneously in the module; however, the manipulated variable (tool settings) can only be applied to the wafers at the entrance of the module (i.e., fresh wafers entering the module). In this case, the metrology delay is the difference in the number of runs between the measured wafer and the wafer at the entrance.

2.2. Determination of effective time delay

It is well-known that the presence and magnitude of a time delay are important determinants of the characteristics of a control loop. In general, if each run is sampled (so that $N_m = 1$), the delay in the control loop is equivalent to the metrology delay. However, when the sampling interval is greater than 1, the effective time delay in the control loop, D_{eff}, is a function of sampling interval and metrology delay as follows

$$D_{eff} = \text{round}_{up}\left(\frac{N_m}{N_s}\right), \tag{1}$$

where because this number must be an integer, the indicated function, $\text{round}_{up}(.)$, means that the computed number must be rounded up to the next higher integer. From the perspective of the sampled-run domain, D_{eff} may be interpreted as meaning that the data available for use at the current sampling instant is from the previous D_{eff}^{th} "measured-run". Fig. 5 shows how the system representation in the production domain is converted into the sampled-run domain for analysis; here, G_C is the controller,

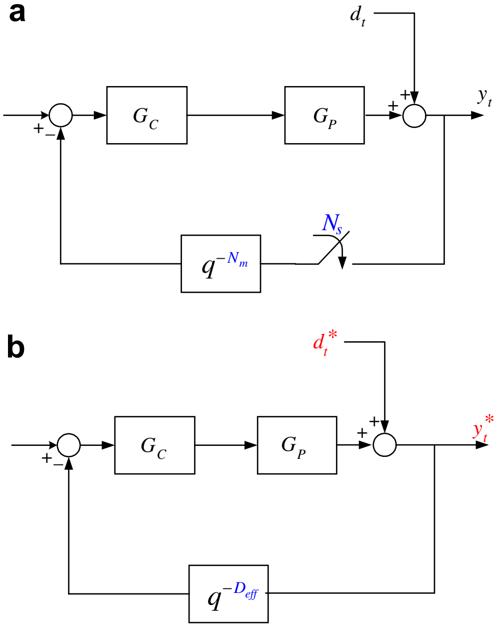

Fig. 5. The control system in: (a) product domain and (b) sampled-run domain.

G_P the process, d the disturbance, y the metrology or quality data, and q is the standard discrete system backshift operator. The "star" variables d^* and y^* refer to the sampled-run domain versions of the original variables d and y in the product domain (See Section 3 for more detail.).

It is more convenient to carry out closed-loop analyses in the sampled-run domain rather than in the original discrete-time domain; specifically, closed-loop stability and controller performance analyses are more easily carried out in the sampled-run domain. We use following two examples to illustrate the meaning of effective delay with constant or variable N_m and N_s.

Example 2.1. Consider the wafer-to-wafer control configuration shown in Fig. 6a, with $N_s = 3$ (i.e., we sample every three wafers), and with a metrology delay that varies from 1 to 3 ($1 \leqslant N_m \leqslant 3$). We obtain from Eq. (1) that $D_{eff} = 1$ for all three values of N_m. In this case there is only a unit delay in the sampled-run domain control system.

Example 2.2. Consider another wafer-to-wafer control strategy, this time in a CMP equipment with an integrated metrology tool that has a delay of three wafers. To maintain high throughput, a decision has been made that only six wafers in a lot of 25 wafers will be measured. A process engineer, believing that the process is less stable early in the lot, decides to take more samples during this period; specifically, the 2nd , 3rd , 5th , 10th , 20th and 25th wafers are chosen to be sampled. It is easier to analyze the effective delay via Fig. 6b, which shows clearly that in the sampled-run domain, the delay is 1 or 2. Because the delay is variable, one must tune the controller conservatively on the basis of the worst case. However, if we rearrange the sampling to be uniform, so that $N_s = 4$, by applying Eq. (1) the result will be an effective delay of only 1. From a

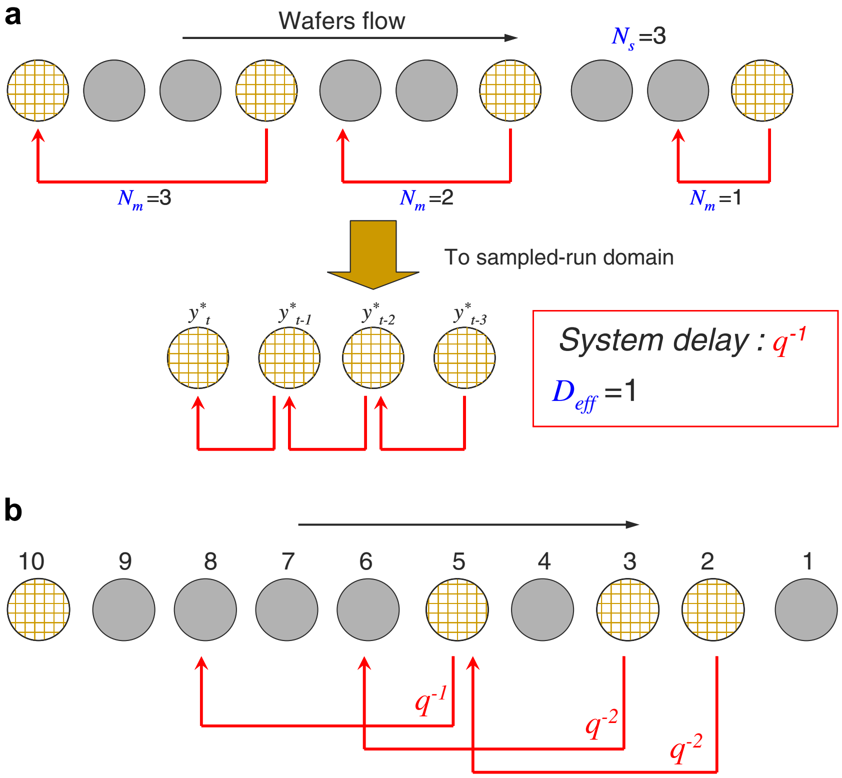

Fig. 6. Effective delay of: (a) Example 2.1 and (b) Example 2.2.

process control point of view, this latter system that has only a unit delay will have better control performance and enjoy a wider range of stable controller parameters than the previous control system.

2.3. Controller derivation and stability

In R2R control of semiconductor manufacturing, the most popular strategies are the EWMA controller, a pure integral controller [12] designed to deal with shift or step-like disturbances, and the double EWMA controller, a proportional-double-integral controller [11] designed to deal with drift or ramp disturbances. These control strategies have been discussed extensively elsewhere as off-the-shelf controllers available for use in semiconductor manufacturing; here we will discuss instead how to derive appropriate controllers with no preconceptions, and subsequently how to ensure robust stability.

On the premise that rational control system design should be based on sufficient knowledge about the process and the input signal types (setpoints to track and/or disturbances to reject), one can use such information along with the internal model control (IMC) principle [13] (or equivalently direct synthesis concepts [21]) to derive appropriate controllers required to achieve specified performance objectives. In particular, it is customary in semiconductor manufacturing to model the batch-to-batch operation as a pure gain process, i.e., $G_\mathrm{P} = K_\mathrm{P}$, with $\widehat{G}_\mathrm{P} = \widehat{K}_\mathrm{P}$ as the best estimate of the true process gain. It is easy to show that according to the IMC principle (or via direct synthesis), the standard feedback controller for a pure gain process subject to shift disturbances (type-1), is

$$G_\mathrm{C}(s) = \frac{1}{\widehat{K}_\mathrm{P}} \cdot \frac{1}{\tau_\mathrm{f} s}, \tag{2}$$

where τ_f is a tuning parameter that determines the speed of the desired closed-loop response. If τ_f is small, the closed-loop system is aggressive but less robust; conversely, the closed-loop system is more robust but also more sluggish for larger values of τ_f. In discrete form, Eq. (2) is

$$G_\mathrm{C}(q^{-1}) = \frac{1}{\widehat{K}_\mathrm{P}} \cdot \frac{1}{\tau_\mathrm{f}(1 - q^{-1})}. \tag{3}$$

This is a pure integral controller, identical in form to the EWMA controller, confirming why EWMA controllers deal effectively with step-like disturbances.

Many pieces of process equipment in semiconductor manufacturing are subject to drift disturbances, such as the decay of polish pad efficiency in a CMP tool, or undesired deposition on side walls in a CVD tool. For simplicity, it is customary to idealize this degradation of efficiency as a ramp disturbance to a time-invariant process model instead of using a more complicated time-varying process model. The IMC design for dealing with such a ramp (type-2) disturbance results in a feedback controller of the form

$$G_\mathrm{C}(s) = \frac{1}{\widehat{K}_\mathrm{P}} \frac{2\tau_\mathrm{f} + 1}{\tau_\mathrm{f}^2 s^2} \tag{4}$$

or in discrete form

$$G_\mathrm{C}(q^{-1}) = \frac{1}{\widehat{K}_\mathrm{P}} \frac{\frac{2}{\tau_\mathrm{f}} - \left(\frac{2}{\tau_\mathrm{f}} - \frac{1}{\tau_\mathrm{f}^2}\right) q^{-1}}{1 - 2q^{-1} + q^{-2}}. \tag{5}$$

This is a proportional-double-integral controller (PI^2) with the same form as a double EWMA controller but with only one tuning parameter, τ_f. For more complicated process models and disturbances, this same procedure can be used to derive R2R controllers appropriate to the known information, no matter how complicated. Note how the nature of the process model and the disturbance structure determine the appropriate controller, emphasizing the importance of understanding the process and disturbance structure before implementing feedback controllers. One should not just choose an existing controller arbitrarily without incorporating such knowledge.

We now consider the stability characteristics of Eqs. (3) and (5). From Fig. 5a, the closed-loop characteristic equation is

$$1 + G_\mathrm{C}(q^{-1}) \cdot G_\mathrm{P}(q^{-1}) \cdot q^{-D_\mathrm{eff}} = 0, \tag{6}$$

whose roots must all be located inside the unit circle in the complex plane for stability. In the following derivation, plant/model mismatch is allowed (because of the simplicity of the pure gain model) with K_P as the true but unknown process gain, and $\widehat{K}_\mathrm{P}$ as the model estimate. In combination with τ_f, the controller tuning constant, the static part of the forward transfer function is defined as

$$K_\mathrm{F} = \frac{K_\mathrm{P}}{\widehat{K}_\mathrm{P} \tau_\mathrm{f}}$$

The characteristic equation can be solved analytically for the ultimate gain value required for the process to be on the verge of instability, using the frequency approach. Let $q = \mathrm{e}^{j\omega}$; the ultimate gain ($K_\mathrm{F,u}$) and ultimate frequency (ω_u) of the forward transfer function ($G_\mathrm{C} G_\mathrm{P} q^{-D_\mathrm{eff}}$) can be found by solving the following two equations when G_C is chosen to be an EWMA controller

$$\arg(G_\mathrm{C} G_\mathrm{P} q^{-D_\mathrm{eff}})_{\omega=\omega_\mathrm{u}} = -\tan^{-1}\left(\frac{\sin \omega_\mathrm{u}}{1 - \cos \omega_\mathrm{u}}\right) - D_\mathrm{eff}\omega_\mathrm{u} = -\pi, \tag{7}$$

$$\left|G_\mathrm{C} G_\mathrm{P} q^{-D_\mathrm{eff}}\right|_{\omega=\omega_\mathrm{u}} = \frac{K_\mathrm{F,u}}{2\sin(\omega_\mathrm{u}/2)} = 1, \tag{8}$$

where ω is frequency. After some algebraic manipulation, the ultimate frequency is obtained as: $\omega_\mathrm{u} = \pi/(2D_\mathrm{eff} - 1)$, which, when substituted into Eq. (8), yields

$$K_\mathrm{F,u} = 2\sin\left(\frac{\pi}{2} \frac{1}{2D_\mathrm{eff} - 1}\right). \tag{9}$$

Fig. 7 shows the stable regions of the EWMA and double EWMA controller parameter space. The region above each curve is the stable region for the indicated value of ξ, the multiplicative model uncertainty parameter defined by:

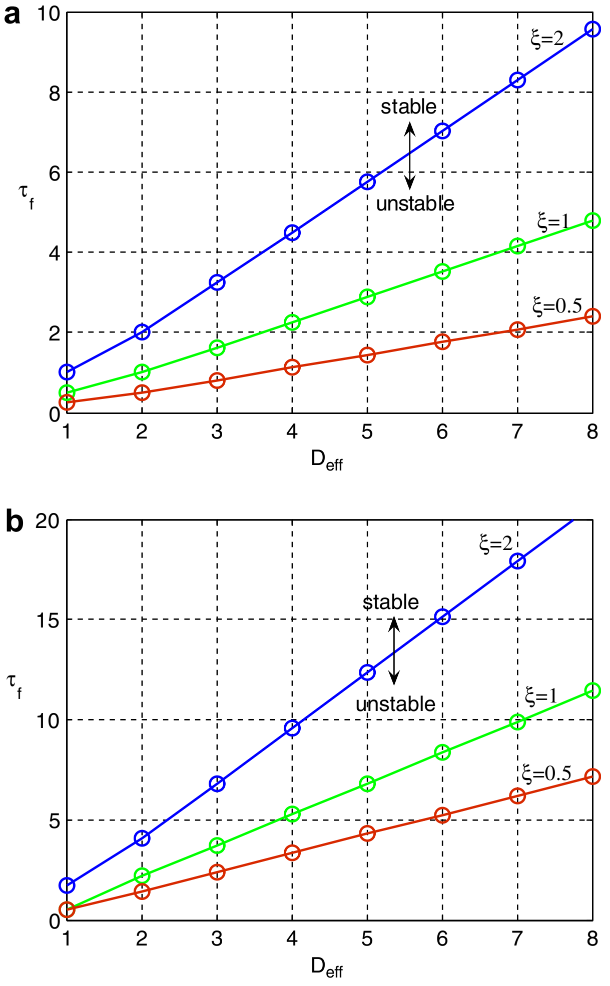

Fig. 7. Stable regions (above of each curve) for different $\xi(K_P/\widehat{K}_P)$ values (a) for the EWMA controller and (b) for the double-EWMA controller.

$\xi = K_P/\widehat{K}_P$. Note that larger D_{eff} and ξ values require larger τ_f values (slower desired closed-loop responses) to ensure stability. The implication is that for large effective delays and/or significant model uncertainty, closed-loop stability can only be achieved at the expense of closed-loop performance, which is in perfect keeping with practical experience (and robust control theory).

3. Control performance

3.1. Disturbance

In this study, we suppose that the process disturbance can be modeled as an *integrated-moving-average* (IMA) time series with a deterministic drift; i.e.

$$d_t - d_{t-1} = a_t - \theta \cdot a_{t-1} + k. \tag{10}$$

Here d_t is the disturbance time series, t is the discrete time index, a_t is a gaussian white noise sequence with zero mean and variance σ_a^2; θ is the moving average coefficient with $0 \leqslant \theta \leqslant 1$; k is the slope of the drift. Many industrial process disturbance data are known to be well-represented by this model [22]. In transfer function form, Eq. (10) is

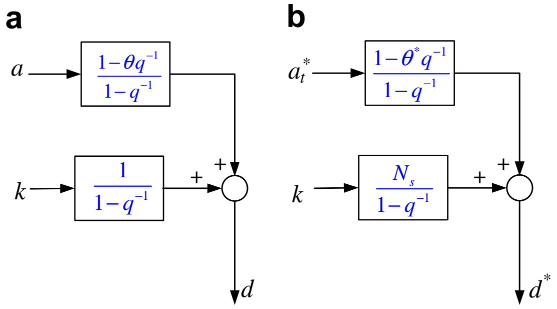
a
a
$\frac{1-\theta q^{-1}}{1-q^{-1}}$
k
$\frac{1}{1-q^{-1}}$
+
+
d
b
a_t^*
$\frac{1-\theta^* q^{-1}}{1-q^{-1}}$
k
$\frac{N_s}{1-q^{-1}}$
+
+
d^*

Fig. 8. The effect of sampling on an IMA time series (a) original sequence ($N_s = 1$) in production domain and (b) modified sequence in the sample-run domain, with sampling interval N_s.

$$d_t(q^{-1}) = \frac{1 - \theta q^{-1}}{1 - q^{-1}} a_t + \frac{k}{1 - q^{-1}}. \tag{11}$$

This disturbance can be treated as a combination of two signals: an IMA series and a ramp, as shown in Fig. 8a. When θ takes on the extreme value 0 or 1, the IMA series becomes, respectively, a random walk or a white noise process.

3.2. Effect of sampling intervals on disturbance model

In Section 2.2, we discussed the effect of sampling intervals on the effective delay in a control structure (Eq. (1)). Even though changing the sampling interval will not affect the nature of the disturbance, it is necessary to redefine the variables in the sampled-run domain (Box et. al. [1]). The relationship between the parameters in the original series (θ and σ_a^2) and in the sampled series (θ^* and $\sigma_{a^*}^2$) is:

$$\frac{N_s(1 - \theta)^2}{\theta} = \frac{(1 - \theta^*)^2}{\theta^*}, \tag{12}$$

$$\frac{\sigma_{a^*}^2}{\sigma_a^2} = \frac{\theta}{\theta^*}, \tag{13}$$

where N_s is the sampling interval $\sigma_{a^*}^2$ and θ^* are the parameters of the new sampled-run sequence as shown in Fig. 8b

$$d_t^* - d_{t-1}^* = a_t^* - \theta^* \cdot a_{t-1}^* + N_s \cdot k, \tag{14}$$

where a_t^* is $N(0, \sigma_{a^*}^2)$.

Having now introduced the process, the controller, the disturbance as well as the concept of the effective delay, we are now in a position to consider the issue of controller performance.

3.3. Achievable performance

Under the condition that the setpoint is unchanged, the disturbance is the only input signal to the control system, so that the relationship between process output y^* and disturbance d^* in Fig. 5a is

$$y_t^* = \frac{1}{1 + G_C \cdot G_P \cdot q^{-D_{\text{eff}}}} \cdot d_t^*. \tag{15}$$

For an IMA disturbance, substituting Eq. (11) into Eq. (15) gives

$$\begin{aligned} y_t^* &= \frac{1}{1+G_{\mathrm{C}}\cdot G_{\mathrm{P}}\cdot q^{-D_{\mathrm{eff}}}}\left(\frac{1-\theta^* q^{-1}}{1-q^{-1}}\cdot a_t^* + \frac{N_{\mathrm{s}}k}{1-q^{-1}}\right) \\ &= S_{\mathrm{a}}(q^{-1})\cdot a_t^* + S_k(q^{-1})\cdot N_{\mathrm{s}}k. \end{aligned} \tag{16}$$

Note that the effect of the drift signal k for the EWMA controller will result in an *offset* of magnitude $\frac{N_{\mathrm{s}}k}{K_{\mathrm{F}}}$ from the target; with the double EWMA controller, this will not be the case because under these conditions, S_k will be zero at steady-state (from the final value theorem). By applying *Parseval's theorem* to Eq. (16), we can compute the *asymptotic variance* as the control performance from the power spectrum

$$\sigma_{y^*}^2 = \frac{1}{2\pi}\int_0^{2\pi} |S_{\mathrm{a}}(\omega)|^2\,\mathrm{d}\omega\cdot\sigma_{\mathrm{a}^*}^2. \tag{17}$$

The ratio of $\sigma_{y^*}^2$, the *sampled-run* process output variance, to σ_{a}^2, the *original* variance of the white noise disturbance, can be obtained, from Eqs. (13) and (17) – a ratio we will use as our performance index. The controller parameter may now be determined by minimizing this performance index, i.e.

$$\min_{\tau_{\mathrm{f}}}\left\{\frac{\sigma_{y^*}^2}{\sigma_{\mathrm{a}}^2}\right\}. \tag{18}$$

To illustrate, consider a system using the EWMA controller for which

$$S_{\mathrm{a}}(q^{-1}) = \frac{1-\theta^* q^{-1}}{1-q^{-1}+K_{\mathrm{F}}q^{-D_{\mathrm{eff}}}}. \tag{19}$$

The control performance with $D_{\mathrm{eff}} = 1$ can be calculated from Eq. (17) as

$$\sigma_{y^*}^2 = \frac{2K_{\mathrm{F}}\theta^* + (1-\theta^*)^2}{(2-K_{\mathrm{F}})K_{\mathrm{F}}}\sigma_{\mathrm{a}^*}^2. \tag{20}$$

Eq. (20) is the same as that reported in [2,3], and optimal achievable control performance (achieved when $\sigma_{y^*}^2 = \sigma_{\mathrm{a}^*}^2$) is obtained with $K_{\mathrm{F}} = 1-\theta^*$. This result also indicates that the EWMA controller is a minimum variance controller when then the effective delay is 1. Fig. 9 shows the achievable optimal control performance for different effective delays and sampling intervals. When the disturbance is a white noise sequence ($\theta = 1$), the achievable optimum control performance is $\sigma_{\mathrm{a}^*}^2$. This implies that a purely white noise disturbance cannot be eliminated with feedback control; in fact any feedback control action taken will only amplify this signal. This, of course, is the well-known result from classical SPC that the best way to deal with a white noise disturbance is to take no control action ($K_{\mathrm{F}} = 0$); alternatively, that the performance of a process subject to only white noise disturbance cannot be improved by taking control action of any sort.

The same analyses (Fig. 9) can also be applied to the double EWMA controller and, as will be shown later in Section 4.2, it can also be extended for determining the tolerable effective delay or sampling intervals.

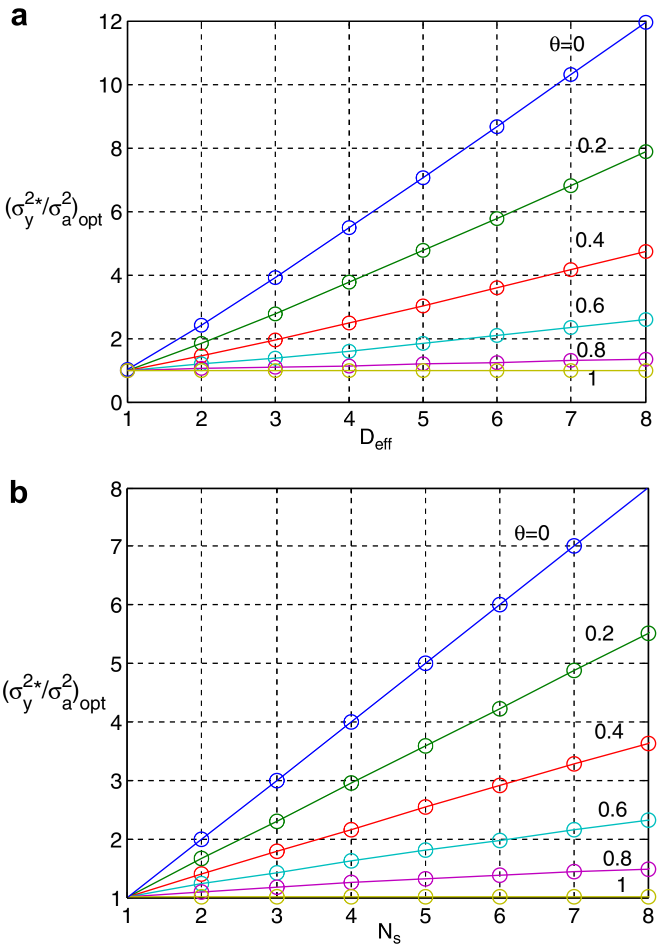

Fig. 9. Achievable optimal control performance using EWMA controller: (a) for different effective delays with $N_{\mathrm{s}} = 1$; (b) for different sampling intervals with $D_{\mathrm{eff}} = 1$.

4. Measurement strategy

Determining the appropriate number of metrology tools to use for a specific manufacturing problem is still a challenging problem. On the one hand, by increasing the number of metrology tools, one can reduce the loading of metrology tools and consequently achieve a higher sampling rate and throughput; however, the additional costs associated with such a decision can be significant. Box and Luceño [3] and Box and Kramer [5] propose a deadband control structure for determining the sampling interval, based on a minimum-cost scheme including adjustment cost, sampling cost, and quality lost. However, they do not consider the effects of sampling on stability – a very important issue in semiconductor manufacturing as we have shown already in this paper. Metrology strategies must be designed carefully and the sampling rate should be optimized to ensure acceptable control performance based on minimizing the cost of operation. However, defining and quantifying the economic impact of such decisions is a very complicated issue in semiconductor manufacturing. In what follows we consider control performance as taking priority over economics. This is not a naïve, unrealistic assumption; there are in fact critical steps in semiconductor manufacturing where the cost of quality lost is unquestionably more substantial than the cost of metrology tools. But we recognize that control performance will not always take priority over economics across the board.

4.1. Effect of sampling intervals and metrology delay

Having characterized in Section 3 the achievable control performance for given sampling intervals and effective delays in general, we now present two specific examples here to illustrate how these results can be used to determine appropriate sampling strategies.

Example 4.1. Consider a stand-alone metrology tool that is shared by eight process tools. For simplicity, assume that for lot-to-lot control, the metrology delay due to queuing is eight runs; and that the effective delay is also eight because each lot is sampled. If a double-EWMA controller is implemented in the process tools, and the disturbance is identified as an IMA process with $\theta = 0.6$, we wish to investigate how the control performance is improved by introducing an additional metrology tool. If the new metrology unit shares half of the measurement loading, the metrology delay will be reduced to four. Fig. 10 shows the effect of different D_{eff} values on the control performance under these circumstances. It indicates that a 39% improvement in the variance can be achieved when D_{eff} is reduced from 8 to 4. The margins of improvement is 29% when D_{eff} is halved from 4 to 2 and 21% when D_{eff} is further halved from 2 to 1. Note that such information can be used to quantify and justify the return on proposed investment on metrology tools.

As mentioned in Section 1, an alternative strategy for dealing with the effect of metrology delay is to implement virtual metrology (i.e., a soft sensor for quality). But since such soft sensor models are never perfect, the precision and accuracy of virtual metrology must be taken into consideration just as one would real sensors. If this soft sensor “measurement noise” is too large, (indicative of poor precision of the virtual metrology component), the resulting performance may not be acceptable, despite the advantage of a much shorter effective delay.

Example 4.2. Consider a wafer-to-wafer control process tool with an integrated metrology tool in which measurement time is negligible ($N_{\mathrm{m}} = 1$). The sampling interval is chosen to be four ($N_{\mathrm{s}} = 4$) for a specific production rate, and the controller and disturbance are the same as in the previous example. If the same throughput is maintained, adding one more metrology unit to this equipment increases the sampling rate to $N_{\mathrm{s}} = 2$. This strategy results in a 26% improvement in optimal performance as shown in Fig. 11.

Usually, control performance is more important than throughput; consequently the more interesting issue is how to increase throughput under the guarantee of performance. This is discussed next.

4.2. Maximum tolerable effective delay and sampling intervals

Before designing the measurement strategy, it is necessary to have available the following information about the process and control system:

- The process characteristics (e.g., gain).
- The disturbance characteristics (e.g., IMA or drift).
- The appropriate controller (and controller tuning parameters).
- The desired specification limit.

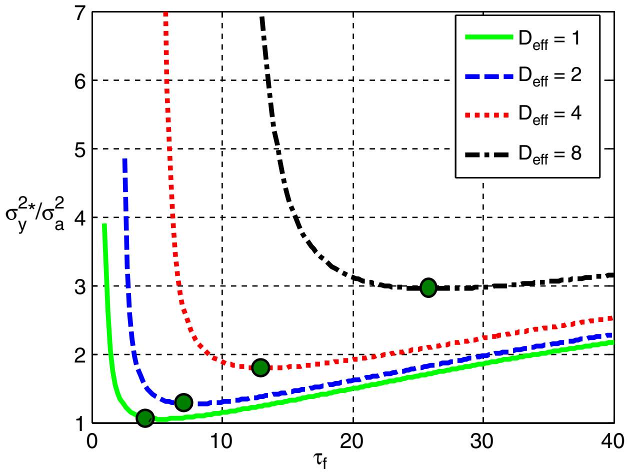

Fig. 10. The effect of D_{eff} on the control performance in Example 4.1 (circles indicate optimal control performance).

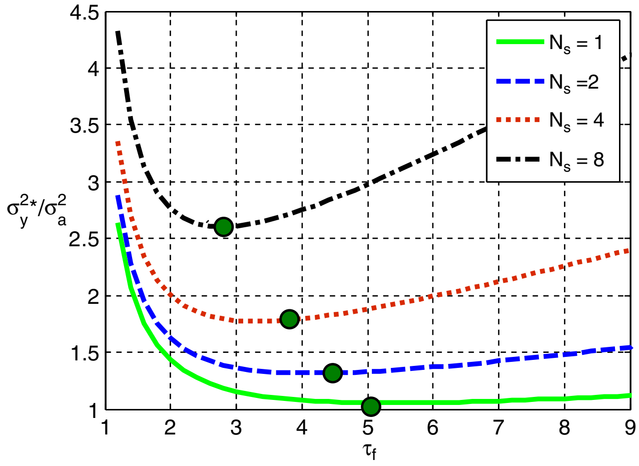

Fig. 11. The effect of N_{s} on the control performance in Example 4.2 (circles indicate optimal control performance).

We use a simple example to demonstrate how to determine the tolerable sampling interval.

Example 4.3. Consider a CMP process with a perfect model ($\xi = 1$), subject to an IMA disturbance with parameters $\theta = 0.4$ and $\sigma_{a_o}^2 = 24$ Å. The specification limits are defined as 100 ± 25 Å. If the process is capable of achieving three sigma quality (only 1% of products out of spec under normal condition), so that $3\sigma_{\text{spec}} = 25$ Å, the threshold value for the performance index is obtained as

$$\frac{\sigma_{\text{spec}}^2}{\sigma_{\text{a}}^2} = \frac{(25/3)^2}{24} = 2.87. \tag{21}$$

If wafer-to-wafer control is implemented with an integrated metrology tool, by applying this threshold value to Fig. 9, we reach two conclusions: (1) the effective delay cannot be greater than 4 if we sample every wafer; (2) the sampling intervals cannot be greater than 5 if $D_{\text{eff}} = 1$.

The first conclusion not only guarantees control performance but also stability. If the tool cannot satisfy this condition, an additional metrology tool or wafer flow mechanism will have to be considered. The second conclusion could be applied to a CMP tool that can process two lots simultaneously, but with only one integrated metrology tool for quality measurement. In this tool, two processing equipments sharing one metrology tool may result in a queue build up inside the tool. Due to the design, it is impossible to add more metrology tools within the tool. However, the preceding analysis of tolerable sampling interval suggests that one can make the sampling interval larger while achieving acceptable control performance. Applying this strategy will therefore substantially increase the throughput for this equipment.

Note that the optimal performance index in Fig. 9 is for optimal controller tuning obtained under the assumption of a perfect process model. In practice, the model will never be perfect, and we must take model uncertainty into consideration. By considering plant/model mismatch in the range $0.8 < \xi < 1.2$, is the maximum tolerable sampling interval still valid? Fig. 12 shows the achievable control performance as a function of the open-loop gain, K_{F}, for different sampling intervals, N_{s}. The entire curve for $N_{\text{s}} = 6$ lies completely above the threshold value in Eq. (21), this strategy should therefore not be implemented because of specification violation. The circle on the curve marks the location of the optimal control performance and corresponding controller settings, $K_{\text{F,opt}}$. Because K_{F} is proportional to ξ, we use squares to locate the control performance at $0.8K_{\text{F,opt}}$ and $1.2K_{\text{F,opt}}$, and the results indicate that robust performance can be maintained for ±20% steady-state gain variations. Actually, the curve crosses the threshold line at $0.61K_{\text{F,opt}}$ and $1.42K_{\text{F,opt}}$. This implies that, for $N_{\text{s}} = 5$, robust performance can be achieved for $0.61 < \xi < 1.42$. The maximum tolerable effective delay can also be determined in a similar fashion by analyzing control performance over different D_{eff} values.

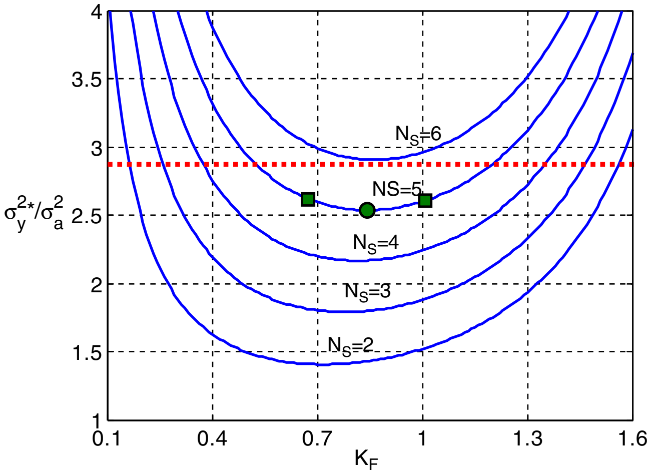

Fig. 12. Control performance as a function of forward loop gain $K_{\text{F}} = (K_{\text{P}}/\widehat{K}_{\text{P}})/\tau_{\text{f}}$ for different sampling intervals N_{s} (circle indicates optimal control performance).

4.3. Measurement priority in queue for a metrology tool

Although integrated metrology tools can improve overall process equipment productivity, capital investment considerations almost always dictate that stand-alone metrology tools are preferred because they can measure lots from several pieces of equipment, and one stand-alone metrology tool is less expensive than several integrated metrology tools. Especially in a stable process step, integrated metrology tools are usually not needed. When in a cluster of stand-alone metrology tools the loading of metrology tools increases suddenly, or one metrology tool shuts down, how should we prioritize the queue for measurements?

Let us define an index, $D_{\text{eff},i}$, to represent the current effective delay value for tool i. $D_{\text{eff},i}$ should be updated every N_{s} runs and reset to zero when metrology data is fed back. From the previous section, it is possible to define the maximum tolerable effective delay for each process tool as $D_{\max,i}$, and then define the difference between these two values as a reference index

$$C_i = D_{\max,i} - D_{\text{eff},i}. \tag{22}$$

We should sort the queue based on this reference index, C_i, and priority should be given to the lot with the smallest C_i. If this value reaches zero and the measurement still cannot be made, to ensure product quality, the equipment should be put on hold until the measurement is completed.

5. Conclusions

We have proposed a procedure for designing metrology strategies from the point of view of process control, including considerations for controller derivation, stability, and control performance. These strategies can help manufacturers determine the appropriate number of metrology units, if performance is the top priority. We also investigated strategies for determining measurement priority. These results are helpful in understanding the influence of sampling interval and metrology delay, and for determining appropriate sampling strategies when implementing run-to-run control.

Nevertheless, economics is often more important for a fab manager who is usually more concerned with answering such questions as: "should the fab purchase integrated metrology tools for each piece of equipment, or several stand-alone metrology tools to be shared by all of the equipment?" "Which will give a better return on investment?" A larger number of metrology tools could provide a tighter process window and result in better yield, but it will also increase capital costs. Combining economics and process control to determine suitable metrology strategies will be the focus of future work.

Acknowledgements

We thank National Science Council for the support of the collaborated research. A.J.S. thanks TSMC for financial support. Comments and suggestions from Jeffery Ward and David S.H. Wong are also appreciated. B.A.O. acknowledges the William L. Friend professorship for facilitating this collaboration.

References

[1] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, Time Series Analysis: Forecasting and Control, third ed., Prentice Hall, 1994.

[2] E. Del Castillo, Statistical Process Adjustment for Quality Control, John Wiley and Sons, New York, 2002.

[3] G.E.P. Box, A. Luceño, Statistical Control: By Monitoring and Feedback Adjustment, John Wiley and Sons, 1997.

[4] S.J. Qin, G. Cherry, R. Good, J. Wang, C.A. Harrison, Semiconductor manufacturing process control and monitoring: a fab-wide framework, J. Process Contr. 16 (2006) 179–191.

[5] G.E.P. Box, T. Kramer, Statistical process monitoring and feedback adjustment – a discussion, Technometrics 34 (1992) 251–285.

[6] P. Jula, C.J. Spanos, R.C. Leachman, Comparing the economic impact of alternative metrology methods in semiconductor manufacturing, IEEE Trans. Semicond. Manuf. 15 (4) (2002) 454–463.

[7] R.K. Nurani, R. Aeklla, A.J. Strojwas, In-line defect sampling methodology in yield management: an integrated framework, IEEE Trans. Semicond. Manuf. 9 (4) (1996) 506–517.

[8] R. Good, S.J Qin, Stability analysis of double EWMA run-to-run control with metrology delay, in: Proceedings of American Control Conference, Anchorage, AK May B-20, 2002.

[9] R. Good, S.J. Qin, On the stability of MIMO EWMA run-to-run controllers with metrology delay, IEEE Trans. Semicond. Manuf. 19 (1) (2006) 78–86.

[10] S.T. Tseng, N.J. Hsu, Sample-size determination for achieving asymptotic stability of a double EWMA control scheme, IEEE Trans. Semicond. Manuf. 18 (1) (2005) 104–111.

[11] A. Chen, R.S. Guo, Age-based double EWMA controller and its application to CMP processes, IEEE Trans. Semicond. Manuf. 14 (2001) 11–19.

[12] A. Infolfsson, E. Sachs, Stability and Sensitivity of an EWMA controller, J. Quality Technol. 25 (4) (1993) 271–287.

[13] M. Morari, E. Zafiriou, Robust Process Control, Prentice Hall, 1989.

[14] S. Wu, P.H. Chen, J.S. Lin, F. Ko, H. Lo, J. Wang, C.H. Yu, M.S. Liang, Real-time Device Performance Prediction for 90 nm and Beyond, in: AEC/APC Symposium USA, Palm Spring, CA, September 2005.

[15] Y.H. Chen, A.J. Su, S.J. Shiu, C.C. Yu, S.H. Shen, Batch sequencing for run-to-run control: application to chemical mechanical polishing, Ind. Eng. Chem. Res. 44 (2005) 4676–4686.

[16] C.M. Fan, R.S. Guo, S.C. Chang, C.S. Wei, SHEWMA: an end-of-line SPC scheme using wafer acceptance test data, IEEE Trans. Semi. Manuf. 13 (2000) 344–358.

[17] T.F. Edgar, S.W. Butler, W.J. Campbell, C. Pfeiffer, C. Bode, S.B. Hwang, K.S. Balakrishnan, J. Hahn, Automatic control of microelectronics manufacturing: practices, challenges and possibilities, Automatica 36 (2000) 1567.

[18] C.C. Yu, A.J. Su, J.C. Jeng, H.P. Huang, S.Y. Hung, C.K. Chao, Control relevant issues in semiconductor manufacturing: Overview with some new results, Control Eng. Pract. 15 (2007) 1268–1279.

[19] S.H. Lu, P.R. Kumar, Distributed scheduling based on due dates and buffer priorities, IEEE Trans. Automat. Contr. 36 (1991) 1406–1416.

[20] M. Quirk, J. Serda, Semiconductor Manufacturing Technology, Prentice Hall, 2001.

[21] B.A. Ogunnaike, W.H. Ray, Process Dynamics, Modeling and Control, Oxford University Press, NY, 1994.

[22] J.F. MacGregor, Optimal choice of sampling interval for process control, Technometrics 18 (1976) 151.

[23] A. Khan, J. Moyne, D. Tilbury, Fab-wide virtual metrology and feedback control, AEC/APC Symposium Asia 2006, Taipei, Taiwan, November 2006.

[24] K. Lensing, B. Stirton, Integrated metrology and wafer-level control, Semicond. Int. 29 (6) (2006) 44–54.

[25] T. Edgar, Run to run control, sampling, and performance monitoring for high mix fabs, IFAC Workshop on Advanced Process Control for Semiconductor Manufacturing, Singapore, December 2006.

[26] J. Qin, Fab-wide control of electrical parameters, IFAC Workshop on Advanced Process Control for Semiconductor Manufacturing, Singapore, December 2006.

[27] J. Moyne, A methodology for roi analysis of run-to-run control solutions, IFAC Workshop on Advanced Process Control for Semiconductor Manufacturing, Singapore, December 2006.