Article

Batch Sequencing for Run-to-Run Control: Application to Chemical Mechanical Polishing

Yih-Hang Chen, An-Jhih Su, Sheng-Jyh Shiu, Cheng-Ching Yu, and Shih-Haur Shen

Ind. Eng. Chem. Res., **2005**, 44 (13), 4676-4686 • DOI: 10.1021/ie0491544

Downloaded from <http://pubs.acs.org> on November 28, 2008

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

[View the Full Text HTML](#)

ACS Publications
High quality. High impact.

Batch Sequencing for Run-to-Run Control: Application to Chemical Mechanical Polishing

Yih-Hang Chen,[†] An-Jhih Su,[†] Sheng-Jyh Shiu,[†] Cheng-Ching Yu,^{*,†} and Shih-Haur Shen[‡]

Department of Chemical Engineering, National Taiwan University, Taipei 106-17, Taiwan, and Applied Materials Taiwan Ltd., Hsin-Chu 300, Taiwan

This work compliments and extends the capability of the run-to-run (R2R) control by sequencing the incomings such that improved control performance can be achieved. Unlike chemical or mechanical systems, this is important for semiconductor manufacturing processes because some prior information about the incoming wafers is generally available. First, the limitation of a R2R control type of feedback system is explained. The frequency domain explanation is as follows: a negative feedback system is effective rejecting a *low*-frequency type of disturbance. From the feedback property, then the answer to the feed sequencing problem becomes clear: arrange the feed (which we are capable of doing) in such a way that it gives a low-frequency characteristic. Furthermore, a scalar load strength indicator is derived, and an engineering approach is taken to sort the incomings effectively. Two policies, from thin to thick (in terms of prethickness, policy L2R) and from thick to thin (policy R2L), are proposed, and they are shown to provide a better control performance over the conventional random feed policy (R). The feed sequencing problems are tested for systems with different dimensions, e.g., single input single output, single input multiple output, and multiple input multiple output systems, which include the model of an experimental chemical mechanical polishing process, and the results show that an improved control performance can be obtained simply by arranging the feeds.

1. Introduction

The general meaning of run-to-run (R2R) control is to implement control action between runs (or from batch to batch) of repetitive processes. Typically, R2R control is applied to processes such as robot systems, batch processes, and semiconductor manufacturing processes. This kind of control strategy uses the information generated from the previous run (e.g., error of output) to correct the process input such that the output at the next run will meet the product specification. In recent years, R2R control has received considerable attention in control of semiconductor manufacturing processes because of its relatively inexpensive cost of control as compared to the loss of unqualified products. As pointed out by Sachs and co-workers,^{1,2} R2R control combines the advantages of both statistical process control (SPC) and feedback control to overcome shifts and drifts during process operation. Typically, the disturbance was characterized statistically, followed by the design of the feedback controller. Many R2R control algorithms have been developed for different disturbance characteristics. They include the following: exponentially weighted moving average (EWMA)^{1,2} and double-EWMA algorithms,³ age-based double EWMA,⁴ generic cell controller,⁵ optimizing adaptive quality controller,⁶ self-tuning controller,⁷ and a Hammerstein model based generalized R2R control algorithm.⁸ Applications of R2R control are chemical mechanical polishing (CMP), chemical vapor deposition, plasma etching, and photolithography.^{2–6,9,10} Moyne et al.¹¹ give an updated summary in R2R control up to 2000, and Del Castillo¹² provides comprehensive references on SPC and R2R control.

In Patel et al.,⁹ R2R control of CMP of dielectric films, the following sources of process variations are clearly identified: (i) tool-induced, (ii) product-induced, and (iii) incoming disturbances. A two-tier approach was taken⁹ where, on the regulatory level, the local controllers compensate for tool-induced and incoming disturbances, while on the supervisory level the product-induced disturbance was rejected. Therefore, it is obvious that, unlike typical R2R control of chemical or robotic systems, some prior knowledge about the quality of the incoming wafers is available in semiconductor manufacturing processes. Instead of performance of R2R control on a *preset* feeding route, it is possible to *arrange* the feeding sequence such that a better control performance can be achieved. This is the objective if this work.

The remainder of this paper is organized as follows. In section 2, the strengths and limitations of a feedback system are explored and the feed sequencing problem is formulated mathematically. An engineering approach is taken to sequence the feed such that it represents a low-frequency type of disturbance into the feedback system. The effects of tuning, model mismatches, and time-varying parameters are explored, and the proposed approach is extended to multivariable systems [e.g., single input single output (SISO), single input multiple output (SIMO), and multiple input multiple output (MIMO) systems]. In section 3, a multizone CMP model is constructed from experiments and used to test the effectiveness of the proposed feeding policies via simulation. All simulations are carried out using MATLAB/Simulink. A lot of the 24 wafers with surface profiles generated from the experiment are used as the simulated feeds. Again, a significant improvement in the control performance is observed.

* To whom correspondence should be addressed. Fax: +886-2-3366-3037. E-mail: ccyu@ntu.edu.tw.

[†] National Taiwan University.

[‡] Applied Materials Taiwan Ltd.

Figure 1. Block diagram for typical R2R control.

2. Feeding Sequence and R2R Control

2.1. Concept. 2.1.1. Feedback and Its Limitation.

In this section, the basic ideas behind R2R control and the capability of such a feedback structure are discussed. As pointed out by Sachs et al.,² R2R control is carried out via a feedback structure while being provided with statistical interpretations. On the implementation level, R2R control can be viewed as a supervisory control, which resets the set point of local controllers.⁹ Let us use the CMP process to illustrate the control architecture. The back pressure (p) and the platen speed (v) are controlled via local controllers. After a polishing run (possibly with a metrology delay if an ex situ measurement is employed), the error (e ; i.e., the difference between the post-thickness and the target) can be used to update the down force or the platen speed. Typically, an I, PI, or PI² (integral, proportional–integral, or proportional-plus-double-integral) controller is employed for the set-point adjustment. Figure 1 illustrates the control structure for R2R control and, in a CMP notation, y represents the post-thickness, y^{set} denotes the target, u , for example, is the backpressure, and u^{set} is the set point of u . Because the dynamics of the local control loop is much faster than the process time in each run (i.e., batch time), the local controller is treated as a part of the process. Therefore, R2R control can be visualized as a feedback system and, in the control notation, the process is denoted as G and the controller is K .

From the control perspective, R2R control is a discrete-time system; the z -transformed transfer functions are generally employed as shown in the block diagram of the feedback system (Figure 2A). Figure 2A describes a typical feedback system with output y , input u , set point (or target) y^{set} , load disturbance d , a delay element (z^{-1}) as a result of the measurement taken from the previous run, process G , and controller K . In semiconductor manufacturing, the process is typically taken as a steady-state gain (i.e., without a dynamic element).^{11,12} Actually, this is a reasonable assumption because the manipulated inputs respond almost instantaneously to set-point changes in the local loop and the speed of the response is much faster as compared to the dynamics of the processing time. For a feedback system depicted in Figure 2A, the sensitivity function S describes the return difference between an input signal and the resultant output signal. For example, the relationship between y and d can be expressed in terms of the sensitivity function:

$$S \equiv \frac{1}{1 + GK} = \frac{y}{d} \quad (1)$$

Consider a typical transfer function for CMP processes with $G(z) = K_p = 4000$ and an integral controller with $K(z) = K_c/(1 - z^{-1})$ and corresponding setting $K_c = 1/(10K_p)$. Figure 2B shows that the magnitude of the sensitivity function is less than unity at low frequency and reaches asymptotically toward 1 at high frequency. That implies that feedback control is effective in at-

Figure 2. Block diagram for a feedback system (A) and corresponding sensitivity function (B).

tenuating a low-frequency disturbance but can do little for a high-frequency one. Also note in Figure 2 that ω_N stands for the normalized frequency (i.e., $\omega_N = \omega T_s/\pi$, where T_s is the sampling time). Despite the fact that the bandwidth (e.g., the corner frequency and ~ 0.1 in Figure 2B) may vary with the controller parameter, the shape of the sensitivity function clearly indicates the inherent limitation of a feedback system. In terms of R2R control (a feedback system), this means that control will be effective if the load disturbance is of a low-frequency nature and will provide little disturbance rejection capability if the load change is of a high-frequency one. Fortunately, the process shifts (e.g., a step change) and drifts (e.g., a ramp down) belong to a low-frequency type of disturbance, as can be understood from the frequency responses of these signals. Therefore, R2R control is expected to perform well.^{11,12}

2.1.2. Problem Formulation. As mentioned earlier, in semiconductor manufacturing processes, we have some prior knowledge about the quality of incoming wafers (e.g., the thickness from prepolish measurement, which is denoted as the prethickness hereafter). The question then becomes, is it possible to achieve improved R2R control by rearranging the incomings? Let us use the prethickness variation in a CMP process as an example. Suppose, in the temporal mode, the average thickness of k incoming wafers is y_{avg} (e.g., 8000 Å), and we intend to remove all of the copper using a fixed removal rate (e.g., 4000 Å/min) provided with a given polish time (e.g., 2 min). For the sake of clarity, let us assume that R2R control is used to adjust the polish time according to deviation from the target. From a linear theory, the incomings (y_i) that deviate from the average prethickness (y_{avg}) can be treated as incoming disturbances (d_i). That is

$$y_i = y_{\text{avg}} + d_i \quad (2)$$

where y_i is the i th feed and d_i is the corresponding load disturbance. With the prethickness available, we can define a nominal feeding sequence as

$$\mathbf{d}_0 = [d_1, d_2, \dots, d_k]^T \quad (3)$$

where the superscript T denotes transpose. Note that the vector $\mathbf{d}_0$ defines the feed sequence in a temporal mode with a total of k incomings. Equations 2 and 3 indicate that the incomings are arranged into the sequence $y_1, y_2, \dots, y_k$ with corresponding load disturbances $d_1, d_2, \dots, d_k$. Any rearrangement of the feeding sequence can be expressed in terms of the permutation matrix $\mathbf{P}_j$ ¹⁴

$$\mathbf{d}_j^T = \mathbf{d}_0^T \mathbf{P}_j \quad (4)$$

where $\mathbf{P}_j (\in \mathbb{R}^{k \times k})$ is a permutation matrix (with single unity in each row and column zero elsewhere). For a system with k incomings, we have " $k!$ " possible feeding sequences (i.e., $k!$ permutation matrices for a $k \times k$ system¹⁴). It is clear that the number of possible feed arrangements grows exponentially and, for a lot of the 24 wafers, the number of possible feeding sequences is $24! = 6.2045 \times 10^{23}$.

To evaluate the control performance, a performance measure is defined first. A 2-norm-based objective function is employed here. It can be expressed in either the time domain or the frequency domain:¹³

$$\|e\|_2^2 = \int_0^\infty e(t)^2 dt = \frac{1}{\pi} \int_0^\infty |e(i\omega)|^2 d\omega \quad (5)$$

The frequency domain approach is taken to facilitate analysis. From the block diagram in Figure 2A, the relationship between the control error and load disturbance can be expressed as

$$e(i\omega) = \frac{d(i\omega)}{1 + GK(i\omega)} \quad (6)$$

Note that d in eq 6 represents a specific sequence of d_i 's that shapes the frequency content of a load disturbance. Here, d is defined by a permutation matrix $\mathbf{P}_j$ (eq 4), and for notational convenience, the subscript j is dropped hereafter. In other words, from the time domain perspective, $\mathbf{P}_j$ describes the shape of load disturbances, e.g., step, ramp up, ramp down, random, etc. With the definitions of all possible feed sequences and performance measurement, the optimal feed sequencing problem can be formulated as an optimization problem. That is, find a permutation matrix such that the 2-norm of control error can be minimized. Mathematically, we have

$$\min_{\mathbf{P}_j} \|e\|_2^2 = \min_{\mathbf{P}_j} \frac{1}{\pi} \int_0^\infty \left| \frac{d(i\omega)}{1 + GK(i\omega)} \right|^2 d\omega \quad (7)$$

However, for a system with k incoming wafers, the number of search spaces is " $k!$ ", and it becomes impractical to solve the optimization problem by evaluating the integrands for $k > 10$. Some engineering insight is needed for the feed sequencing problem.

2.1.3. Heuristic Approach. As mentioned earlier, R2R control, a feedback system, is effective in eliminating low-frequency disturbances (Figure 2B) and, therefore, a better R2R control performance can be obtained if the incomings can be arranged in such a manner that the feeding sequence shows a low-frequency characteristic. Let us use a generic system to illustrate the effect of the feeding sequence. Consider a CMP example where the prethickness of the incomings follows a normal distribution

$$y_t = y_{\text{avg}} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \quad (8)$$

Figure 3. Three feeding policies based on a normal distribution: L2R (thin to thick), R2L (thick to thin), and R (random).

where the subscript t is the temporal index that denotes the t th feed and ϵ_t is a normally distributed error with zero mean and variance σ^2 . That implies (eq 2) that

$$d_t = \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \quad (9)$$

On the basis of engineering feasibility, three feeding policies are devised. The first one is to arrange the feeds from the left side of a normal distribution to the right (left to right according to the distribution as shown in Figure 3), which is denoted as policy L2R, the second one is from right to left (denoted as policy R2L), and the third one is to arrange the feed randomly (denoted as the policy R, as shown in Figure 3). In terms of the prethickness in the CMP, the policy L2R means from thin to thick, policy R2L implies from thick to thin, and policy R stands for doing nothing. One may devise much more sophisticated feeding policies, but it should be emphasized that these three policies are relatively easy to implement. The next question then becomes, what are the frequency contents of these feeding policies? The power spectrum density (PSD) is often used to evaluate the frequency content of a time series. Let us take 500 data from a normal distribution. The first column of Figure 4 shows the feeding sequences for these three policies, where policies L2R and R2L indicate smooth changes among runs while policy R shows a typical random behavior of the incomings. After Fourier transformation, the PSDs in Figure 4 clearly reveal that policies L2R and R2L indicate low-frequency characteristics while policy R displays the same strength over the entire frequency range (Figure 4). From the analyses of the frequency content of load disturbances (Figure 4) and the effective frequency range of a feedback system (Figure 2B), it is expected that policies L2R and R2L will give much a better control performance than policy R. Note that, by a control performance, we mean the 2-norm of error using the same control algorithm with the same tuning parameters except for different feeding sequences.

2.2. SISO Systems. The following linear SISO example mimics R2R control of the prethickness variation in a CMP process. For 500 incoming wafers, the average copper thickness to be removed is $y_{\text{avg}} = 8000 \text{ \AA}$ with a fixed removal rate $RR = 4000 \text{ \AA/min}$ [$G(z) = K_p = 4000$], and this corresponds to a polish time of 2 min ($u = 2$). Without R2R control, the polishing process will be carried out accordingly and, obviously, errors will result if the prethicknesses are different from the average one. This is referred to as the open-loop error (e_{OL}). If the prethickness of the incomings follows a normal distribution (eqs 8 and 9), the square of the open-loop error is simply the variance of the incoming thickness variation (i.e., $\|e_{\text{OL}}\|_2^2 = \sigma^2$). In the context of feedback control, system variables, e.g., y and u , typically are expressed in terms of deviation variables (i.e., deviation from the nominal steady state). Thus, R2R control of the simple

Figure 4. Three feeding policies and corresponding PSDs for the SISO example with 500 runs generated from a normal distribution.

CMP process is translated into a feedback system with a static process transfer function $[G(z) = 4000]$, a unit delay in the feedback loop (z^{-1}), and a controller K (Figure 2A). For each wafer, the deviation from the average thickness is treated as load disturbance d_i (eqs 2, 8, and 9) and the sequence of the load change ($\mathbf{d}$) is defined by the permutation matrix $\mathbf{P}_j$ (eq 4). For the gain process plus unit delay, if the zeroth-order Pade approximation is employed, the internal model control principle gives an integral-only controller. That is

$$K(z) = K_c / (1 - z^{-1}) \quad (10)$$

with a tuning constant K_c , where here K_c is set to $1/6$ of K_p . In the time domain expression, the control action is computed according to

$$u_t - u_{t-1} = \Delta u_t = K_c e_t \quad (11)$$

where the subscript t is the time index. Consider the case of 500 incoming wafers and d_i following a normal distribution with a variance of 1. Let us use this simple example to illustrate the importance of the feeding sequence. All simulations are carried out using MATLAB/Simulink. A comparison is made between the open-loop error and closed-loop error (under R2R feedback control) for three feeding policies.

For policy R, R2R offers little improvement over the open-loop case, as shown in Figure 5. The error ratio (the closed-loop error over the open-loop error, $\|e_{CL}\|_2^2 / \|e_{OL}\|_2^2$) for the random feed policy is close to 1, and practically no improvement is made for feedback control. This is within one's expectation because policy R excites the entire frequency range uniformly (Figure 5). On the other hand, policies L2R and R2L provide significant improvement over the open-loop case, as can be seen from the error ratios of 4–5%, a 95% improve-

Figure 5. Control performance (in terms of the error ratio in 2-norm) of three feeding policies for the SISO example with 500 runs.

Figure 6. Block diagrams of the original feedback system (A) and SVD-based design for a nonsquare multivariable system (B).

ment! The results also coincide with one's intuition, as shown in the PSDs of these two feeding sequences. The behavior of the manipulated variables (u) for the L2R and R2L cases may raise some concern for possible instability (Figure 5). However, as one compares the shape of the load disturbance with that of the manipulated variable in Figure 5, it becomes clear that the manipulated variable counteracts with the load change and, unless the load change becomes unstable (e.g., the prethickness grows to infinity), the feedback system will remain stable.

One may ask whether the difference in the nominal performance is the result of controller tuning. Again, this can be explained from the PSD of the feed sequences. R2R control gives virtually no performance improvement for policy R as a result of the inherent limitation imposed on the feedback system. This situation holds regardless of controller settings. As for the R2L policy (similarly for the L2R policy), the PSD is a monotonically decreasing function of frequency and, therefore, a tighter controller tuning (i.e., increasing K_c) will result in a better control performance. The same result can be found when model mismatches arise.

2.3. Nonsquare Systems. Up to now, we have addressed the feed sequencing problem for SISO systems. Unfortunately, most of the semiconductor manufacturing processes are of the distributed parameter nature. Thus, we have a nonsquare multivariable system. Two possible cases are explored. One is the system with a single manipulated variable, a SIMO system, and the other is the MIMO system, systems with multiple manipulated variables.

2.3.1. SIMO Systems. Let us use a simple static example to illustrate the effects of the feeding sequence in R2R control. Consider a nonsquare linear process with m outputs and one input.

$$\mathbf{y} = \mathbf{K}_p \mathbf{u} \quad (12)$$

where $\mathbf{y}$ ($\in \mathcal{R}^{m \times 1}$) is the process output with the entry y_i , $\mathbf{u}$ ($\in \mathcal{R}^1$) is the process input, and $\mathbf{K}_p$ ($\in \mathcal{R}^{m \times 1}$) consists of the model parameters. Note that the spatial mode is discussed here. If eq 12 describes a Cu polishing process, y_i 's can be viewed as the copper thickness from the center (y_1) toward the edge (y_m), K_{pi} 's denote removal rates across the radial position, and u denotes the polish time. It should be emphasized here that, for a multivariable system under feed sequencing, two modes (and corresponding indices) should be distinguished. One is the temporal mode with the subscript t (eq 2), and the other is the spatial mode with the subscript i (e.g., eq 12). Assuming that the t th incoming wafers can be described by

$$\mathbf{y}_t = \mathbf{y}_{\text{avg}} + \epsilon_t = \begin{bmatrix} y_{\text{avg},1} + \epsilon_{t1} \\ \vdots \\ y_{\text{avg},m} + \epsilon_{tm} \end{bmatrix}, \quad \epsilon_{ti} \sim N(0, \sigma_i^2) \quad (13)$$

where $y_{\text{avg},i}$ is the average thickness in the i th wafer position ($i = 1, \dots, m$) for all incoming wafers ($t = 1, \dots, k$) and ϵ_{ti} is a normally distributed error with zero mean and variances σ_i^2 , eq 13 describes a typical surface profile of an incoming wafer.

Figure 7. Control performance (in terms of the error ratio in 2-norm) of feed sequencing according to the load strength indicator (d^*) of two feeding policies for the simple 2×1 example.

Because we are more interested in the planarization of the wafer surface, variations across the radial position should also be treated as load disturbances for the control system to overcome. Thus, the average wafer surface across the radial position (in a spatial mode) is used. That is

$$\bar{y} = \sum_{i=1}^m y_{\text{avg},i} \quad (14)$$

where $\bar{y}$ is the average wafer thickness in the temporal and spatial modes. Following eq 2, we have

$$\mathbf{y}_t = \bar{y} + \mathbf{d}_t = \begin{bmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix} + \begin{bmatrix} (y_{\text{avg},1} - \bar{y}) + \epsilon_{t1} \\ \vdots \\ (y_{\text{avg},m} - \bar{y}) + \epsilon_{tm} \end{bmatrix} \quad (15)$$

The disturbance vector for the t th incoming can be

further divided into two parts:

$$\mathbf{d}_t = \mathbf{d}^s + \mathbf{d}_t^t = \mathbf{d}^s + \epsilon_t = \begin{bmatrix} (y_{\text{avg},1} - \bar{y}) \\ \vdots \\ (y_{\text{avg},m} - \bar{y}) \end{bmatrix} + \begin{bmatrix} \epsilon_{t1} \\ \vdots \\ \epsilon_{tm} \end{bmatrix} \quad (16)$$

Here, $\mathbf{d}^s$ represents the disturbance in the spatial mode (will not change with the feeding sequence), and $\mathbf{d}_t^t$ denotes the disturbance in the temporal mode. It then becomes clear that the feeds should be sequenced based on the temporal part of the load disturbance ($\mathbf{d}_t^t = \epsilon_t$) because the deviation in the spatial mode (e.g., $y_{\text{avg},i} - \bar{y}$) comes into the system as a steplike disturbance, a low-frequency type of disturbance, which can be taken care of by the R2R feedback controller. The appendix offers an alternative approach to validating that the feed should be sequenced according to the temporal part of the disturbance.

Figure 8. Schematics of a three-zone CMP.

With the temporal part of the load variable ($\mathbf{d}_t^t$) defined, how can we arrange the feed based on a vector? The block diagram in Figure 6B presents an attractive solution, and that is reducing the dimension to 1 according to the singular value decomposition (SVD). Thus, one obtains

$$d_t^* = \mathbf{U}^T \mathbf{d}_t^t = \mathbf{U}^T \boldsymbol{\epsilon}_t = [u_{11} \dots u_{1m}] \begin{bmatrix} \epsilon_{t1} \\ \vdots \\ \epsilon_{tm} \end{bmatrix} = \sum_{i=1}^m u_{1i} \epsilon_{ti} \quad (17)$$

Here d_t^* is a scalar that can be used as an indicator of the strength of the load change in the temporal mode.

Without loss of generality, consider a 2×1 case of $K_{p1} = K_{p2} = 4000$, and ϵ_{ti} is a normally distributed error with zero mean and a variances of 1. Let us further assume that $y_{\text{avg},1} = y_{\text{avg},2}$. Again, policies R and R2L are used to illustrate the effects of the feed sequencing (based on d_t^*) on the control performance. R2R control is designed using the SVD (Figure 6B), and the controller gain (K_c) is set to $1/6$ of K_p . Figure 7 reveals that, similar to the SISO case, the error ratio of policy R stays close to 1 while policy R2L gives almost 50% improvement over the uncontrolled case. It is observed that the margin of improvement, despite being significant, is much less than that of the SISO example ($\sim 95\%$ in Figure 5). The reason is that, for a nonsquare system, it is not possible to keep both outputs at set points with only one manipulated variable and a certain degree of control error will remain, regardless of the feed arrangement and/or complexity of the controllers. Figure 7 also shows that the load strength indicator (d_t^*) indeed shows a low-frequency characteristic for policy R2L as opposed to the random behavior for policy R. This is exactly why the feedback mechanism of R2R control will be in effect and a 50% improvement can be made.

2.3.2. MIMO Systems. The ongoing analysis deals with a system with only single input, and the structure of the process (SIMO) leads to a scalar load strength indicator. However, in semiconductor manufacturing, systems with multiple manipulated inputs are sometimes encountered (i.e., $\mathbf{K}_p \in \mathcal{R}^{m \times n}$ with $m \geq n > 1$), how can we generate a load strength indicator in this case? Again, the SVD offers some light in this direction. Consider a general (square or nonsquare) multivariable

process $\mathbf{K}_p (\in \mathcal{R}^{m \times n}$ with $m \geq n$) with the temporal part of the load vector $\mathbf{d}_t^t (\in \mathcal{R}^{m \times 1})$. Unlike the SIMO case, the product of $\mathbf{U}^T \mathbf{d}_t^t$ is an $n \times 1$ vector:

$$\mathbf{U}^T \mathbf{d}_t^t = \mathbf{U}^T \boldsymbol{\epsilon}_t = \begin{bmatrix} u_{11}\epsilon_{t1} + u_{21}\epsilon_{t2} + \dots + u_{m1}\epsilon_{tm} \\ \vdots \\ u_{1n}\epsilon_{t1} + u_{2n}\epsilon_{t2} + \dots + u_{mn}\epsilon_{tm} \end{bmatrix} \quad (18)$$

Note that $\mathbf{U}$ is an orthonormal matrix where the strength along the principle direction is not explicitly taken into consideration and it is described by the singular values in the diagonal $\boldsymbol{\Sigma}$ matrix ($\in \mathcal{R}^{n \times n}$).^{14,15} Taking the degree of stretching into consideration, we arrive at the following scalar load strength indicator:

$$d_t^* = \text{sum}(\boldsymbol{\Sigma} \mathbf{U}^T \mathbf{d}_t^t) = \sum_{j=1}^n \sigma_j \sum_{i=1}^m u_{ij} \epsilon_{ti} \quad (19)$$

where $\text{sum}(\cdot)$ is the summation of all elements in the vector ($\cdot$). Note that d_t^* in eq 19 is the load strength indicator for a general $m \times n$ system and eq 16 is simply a special case for $n = 1$ (note that the nonzero value of σ_1 will not affect the result of sorting).

With the load strength indicator (d_t^*) available, the incoming wafers can be arranged in the following steps:

S1. Find the average thickness across the radial position ($y_{\text{avg},i}$, $i = 1, \dots, m$) for all incomings ($t = 1, \dots, k$) and corresponding temporal deviations from the average ones ($\mathbf{d}_t^t = \boldsymbol{\epsilon}_t$) (eq 16).

S2. Transform the vector-valued $\boldsymbol{\epsilon}_t$ into the scalar load strength indicator (d_t^*) for the t th incoming feed (eq 19).

S3. Sequence the feed according to d_t^* for all incomings ($t = 1, \dots, k$) using policy R2L or L2R, whichever has a smaller absolute value of d_1^* (the first incoming is practically uncontrolled in the R2R framework).

These three steps complete the engineering feed sequencing for effective R2R control. Also note that for special cases, e.g., SISO systems, the step S2 can be bypassed and the procedure can be simplified significantly.

3. Nonsquare CMP Example

3.1. Multizone CMP. In recent years, CMP has become an important technology because there exists the requirement of global wafer planarity for multilevel interconnect devices in integrated circuit fabrication. CMP also gives the advantages of defect reduction, wide windows for etching and lithography, and yield improvement.^{16–18} Despite recent advances in CMP, some manufacturing concerns associated with successful implementation of CMP remain to be overcome.^{16–23} In theory, CMP can achieve global planarity, but the problem of within wafer nonuniformity (WIWNU) still remains as one of the major operation concerns. WIWNU indicates the variation in the surface thickness across the wafer radial position, especially on the edge. Besides, in the surface profile of wafers produced from an electrochemical plating (ECP) process, it appears that the metallic layer is thicker on the edge area. Thus, a new type of CMP, multizone CMP, is implemented.²⁴ Multizone CMP is expected to reduce WIWNU and to achieve a wider process window. Unlike the typical single-zone configuration, the wafer carrier is divided

Figure 9. Columns of the steady-state gain matrix $\mathbf{K}_p$ (A) and columns of the output orthonormal matrix $\mathbf{U}$ (B) for the three-zone CMP example.

into three zones in the radial position and different pressures can be applied to each zone (Figure 8).

For a wafer carrier with a radius of 150 mm, the first zone covers 0–130 mm, the second zone ranges from 130 to 140 mm, and the third zone covers 140 mm and beyond. Next, the well-known Preston's equation²⁵ is used to model the polishing process. It models the material removal as a linear function of pressure and rotation speed.

$$RR = K \cdot p \cdot v \quad (20)$$

where K is the Preston constant, p is pressure, and v is the rotation speed. The Preston equation can be extended to multizone CMP in a straightforward manner. For the multizone system, the relationship for the removal rate at the i th radial position can be expressed as

$$RR_i = \sum_j K_{ij} p_j v = \sum_j K_{p,ij} p_j \quad (21)$$

where K_{ij} is the local Preston constant describing the effect of pressure from the j th zone on the i th radial position, p_j denotes the pressure of the j th zone, and v is again the rotation speed. Because the rotation speed is fixed throughout all runs, v is absorbed into $K_{p,ij}$ for subsequent development.

The next task is to find all of the constants to describe the multizone CMP system. In the modeling phase, two two-level factorial designs are carried out on all three pressures (20 combinations). For the Applied Materials' Mirra polisher using a 300-mm Titan Profiler head, the copper thickness across the radial position is measured. The measurement is converted to 59 data points along the radial position. Then, the approximated removal rate at each radial position can be obtained by dividing the amount of copper removal with the total polish time, which is set at 59 s in each test. In a control notation, the multizone polishing model can be written as

$$\mathbf{y} = \mathbf{K}_p \mathbf{u} \quad (22)$$

Figure 10. Surface profiles of 24 wafers (a lot) after ECP and right before CMP.

Figure 11. Control performance (in terms of the error ratio in 2-norm) of feed sequencing according to the load strength indicator (d^*) of two feeding policies for the three-zone CMP under multivariable R2R control (a 59×3 system).

where $\mathbf{y}$ is the removal rate vector ($\in \mathcal{R}^{59 \times 1}$), $\mathbf{K}_p$ is a steady-state gain matrix ($\in \mathcal{R}^{59 \times 3}$) describing local Preston constants, and $\mathbf{u}$ is a vector of input pressures ($\in \mathcal{R}^{3 \times 1}$). Given experimental data, the $\mathbf{K}_p$ matrix was simply determined using the least-squares regression. The regression result versus wafer radial position is plotted in Figure 9. The effects of input pressures on each zone appear to be quite reasonable. The pressure applied on zone 1, u_1 , shows a significant effect on the removal rate in the range of 0–110 mm, and the influence degrades gradually toward the wafer edge as

can be seen from the numerical values of the first column of $\mathbf{K}_p$ (i.e., $\mathbf{K}_{p1}$). The pressure applied on the second zone, u_2 , shows a large value of the Preston constant around 135 mm, and its influence on the removal rate diminishes toward both ends as shown in $\mathbf{K}_{p2}$. As expected, the input pressure of the third zone shows little effect on the removal rate until 135 mm, and then the influence grows linearly toward to the edge.

Next, we perform the SVD on the $\mathbf{K}_p$ matrix. This leads to a 59×3 $\mathbf{U}$ matrix, a 3×3 $\mathbf{V}^T$ matrix, and a 3

$\times 3$ diagonal matrix Σ with the following singular values: $\sigma_1 = 264$, $\sigma_2 = 31$, and $\sigma_3 = 16$. Figure 9 also reveals that the columns of the $\mathbf{U}$ matrix also take shapes qualitatively similar to that of $\mathbf{K}_p$.

To test the effectiveness of different feeding policies, experimentally obtained surface profiles of 24 wafers (a lot) are used for R2R control, as shown in Figure 10. These surface profiles were obtained from Cu deposition via the ECP process, which is carried out with Applied Materials' iECP and the copper thickness varying from 8500 to 12 000 Å. All copper thickness measurements were made with the i-Scan sensor. In the following, a simulation study is carried out to evaluate the R2R control performance using the profiles in Figure 10.

3.2. Control. Multivariable R2R control is explored. This corresponds to a 59×3 multivariable control problem. Following the SVD, the controller is arranged according to the block diagram in Figure 6B. Tuning constants of the integral-only controllers for all three diagonal elements in $\mathbf{K}_{\text{diag}}$ are set to $1/6$. Following the procedure in section 2.3, two feeding policies are examined. The feed sequencing procedure is as follows: (1) perform SVD on $\mathbf{K}_p$, (2) compute $y_{\text{avg},i}$ ($i = 1, \dots, 59$) across the radial position for all 24 incoming wafers ($t = 1, \dots, 24$), (3) find the corresponding temporal deviations from the average ones ($\epsilon_t = \mathbf{y}_t - \mathbf{y}_{\text{avg}}$) for each incoming (eq 13), (4) compute the load strength indicator d_t^* (eq 19), and (5) sort the feed according to d_t^* . The 24 wafers are sorted accordingly for the R2L policy and perturbed randomly for policy R. The result, Figure 11, indicates that a 35% improvement in the control performance can be achieved by arranging these 24 incoming wafers properly (policy R2L), while the random feed policy (policy R) practically offers no improvement as compared to the open-loop error.

4. Conclusion

In semiconductor manufacturing processes, particularly CMP, we have some prior knowledge about the quality of incoming wafers (because they come directly from the previous processing step). In this work, we intend to answer the following question: Is it possible to achieve improved R2R control by rearranging the incomings? First, the characteristics of a feedback system are studied, and its limitation is also explained in the frequency domain. That is, a negative feedback system is effective at rejecting a low-frequency type of disturbance, and this is exactly why R2R control was shown to be effective at overcoming shifts/drifts often encountered in semiconductor manufacturing processes. This work strengthens the capability of R2R control by sequencing the incoming such that improved control performance can be achieved. From the feedback property, the answer to the feed sequencing problem becomes obvious: arrange the feed (i.e., load disturbance) in such a way that it gives a low-frequency characteristic. Furthermore, a scalar load strength indicator is proposed and an engineering approach is taken to sort the incomings effectively (as opposed to exhausting all $k!$ possibilities). Two policies are proposed: from thin to thick (in terms of the prethickness, policy L2R) and from thick to thin (policy R2L). They are shown to provide significant improvement in the control performance, while the typically used random feed policy (R) shows practically no reduction in the closed-loop error over the open-loop one. The issues of controller tuning, model mismatches, and time-varying (slow-drifting)

parameters are also explored, and the analyses again reveal the robust performance of the proposed feeding policies. The feed sequencing problems are tested for systems with different dimensions, e.g., SISO, SIMO, and MIMO systems, which include the model of an experimental CMP process. Again, the results show that a significant improvement in the control performance can be obtained by arranging the feed properly. Finally, it is important to recognize that this improvement is achieved by extremely simple means: rearranging the incoming as a low-frequency type of disturbance.

Acknowledgment

This work was supported in part by the National Science Council of Taiwan under grant NSC 93-2214-E002-026.

Appendix: Load Strength Indicator for Multivariable Systems Based on Achievable Performance

Another approach to finding the correct feed sequencing measure is to take the capability of the least-squares-based controller into account. The achievable performance is used here, and without further information, let us assume that the desired removed thickness is simply $y_{\text{avg},i}$. From the achievable performance for a nonsquare system, we have

$$\mathbf{E}_{\min} = \mathbf{y}^{\text{set}} - \mathbf{y}_{\text{best}} = (\mathbf{I} - \mathbf{K}_p \mathbf{K}_p^+) \mathbf{y}^{\text{set}} = (\mathbf{I} - \mathbf{U} \mathbf{U}^T) \mathbf{y}^{\text{set}} \quad (\text{A1})$$

From eqs 14 and 16, we have

$$\mathbf{y}_t = \mathbf{y}_{\text{best}} + (\mathbf{y}_{\text{avg}} - \mathbf{y}_{\text{best}}) + \epsilon_t = \mathbf{y}_{\text{best}} + (\mathbf{I} - \mathbf{U} \mathbf{U}^T) \mathbf{y}_{\text{avg}} + \epsilon_t \quad (\text{A2})$$

The second and third factors on the right-hand side of eq A2 can be viewed as the load changes. Thus, one obtains

$$\mathbf{d}_t = (\mathbf{I} - \mathbf{U} \mathbf{U}^T) \mathbf{y}_{\text{avg}} + \epsilon_t \quad (\text{A3})$$

From the definition of the load strength indicator, we arrive at

$$d_t^* = \mathbf{U}^T \mathbf{d}_t = \mathbf{U}^T (\mathbf{I} - \mathbf{U} \mathbf{U}^T) \mathbf{y}_{\text{avg}} + \mathbf{U}^T \epsilon_t = \mathbf{U}^T \epsilon_t \quad (\text{A4})$$

This implies sequencing of the feed according to ϵ_t provided with the $\mathbf{U}$ matrix.

Notations

CMP = chemical mechanical polishing
 d = load variable
 $\mathbf{d}$ = vector of the load variable (temporal or spatial mode)
 $\mathbf{d}_0$ = vector of the load variable (temporal mode)
 $\mathbf{d}^s$ = spatial part of the load disturbance
 $\mathbf{d}_t^t$ = temporal part of the load disturbance
 d_t^* = transformed load variable (load strength indicator)
 $\mathbf{G}$ = process transfer function
 e = error
 e^* = transformed error
 e_{CL} = closed-loop (under control) error
 $\mathbf{E}_{\min}$ = least-squares error for a nonsquare system
 e_{OL} = open-loop (without control) error
 I = integral-only controller
 k = number of feeds

K = controller transfer function
 K_c = controller gain
 $\mathbf{K}_{\text{diag}}$ = diagonal controller under SVD
 $\mathbf{K}_p$ = steady-state gain matrix
 $\mathbf{K}_{pi}$ = i th column of $\mathbf{K}_p$
 K_p^+ = pseudo-inverse of K_p
 L2R = left-to-right (based on a normal distribution) policy
 m = number of outputs
 n = number of inputs
 N = normal distribution
 p = pressure (for CMP)
 PI = proportional–integral controller
 PI² = proportional-plus-double-integral controller
 PSD = power spectrum density
 R2L = right-to-left (based on a normal distribution) policy
 R2R = run to run
 RR = removal rate
 S = sensitivity function
 SVD = singular value decomposition
 u = input variable
 $\mathbf{U}$ = output orthonormal matrix (from SVD)
 v = rotation speed (for CMP)
 $\mathbf{V}$ = input orthonormal matrix (from SVD)
 y = output variable
 $y_{\text{avg},i}$ = average value of the output in temporal mode at the i th position
 y_{best} = achievable performance ($E_{\text{min}} = y^{\text{set}} - y_{\text{best}}$)
 y^{set} = set point of y
 z = z -transformation variable

Greek Letters

ϵ = deviation from the mean value
 σ = singular value
 ω = frequency
 ω_N = normalized frequency

Literature Cited

- (1) Ingolfsson, A.; Sachs, E. Stability and sensitivity of an EWMA controller. *J. Qual. Technol.* **1993**, *25*, 271–287.
- (2) Sachs, E.; Hu, A.; Ingolfsson, A. Run by run process control: combining SPC and feedback control. *IEEE Trans. Semicond. Manuf.* **1995**, *8*, 26–43.
- (3) Butler, S. W.; Stefani, J. A. Supervisory run-to-run control of polysilicon gate etch using in situ ellipsometry. *IEEE Trans. Semicond. Manuf.* **1994**, *7*, 193–201.
- (4) Chen, A.; Guo, R. S. Age-based double EWMA controller and its application to CMP processes. *IEEE Trans. Semicond. Manuf.* **2001**, *14*, 11–19.
- (5) Boning, D. S.; Moyne, W. P.; Smith, T. H.; Moyne, J.; Telfeyan, R.; Hurwitz, A.; Shellman, S.; Taylor, J. Run by run control of chemical-mechanical polishing. *IEEE Trans. Compon. Packag. Manuf. Technol. Part C* **1996**, *19*, 307–314.
- (6) Qin, S. J.; Scheid, G. W.; Riley, T. J. Adaptive run-to-run control and monitoring for a rapid thermal processor. *J. Vac. Sci. Technol. B* **2003**, *21*, 301–310.
- (7) Del Castillo, E.; Hurwitz, A. M. Run-to-run process control: literature review and extensions. *J. Qual. Technol.* **1997**, *29*, 184–196.
- (8) Nagrath, D.; Bequette, B. W.; Cramer, S. M. Evolutionary operation and control of chromatographic processes. *AIChE J.* **2003**, *49*, 82–95.
- (9) Patel, N. S.; Miller, G. A.; Guinn, C.; Jenkins, S. T. Device dependent control of chemical-mechanical polishing of dielectric films. *IEEE Trans. Semicond. Manuf.* **2000**, *13*, 331–343.
- (10) Ljung, L.; Soderstrom, T. *Theory and Practice of Recursive Identification*; MIT Press: Cambridge, MA, 1983.
- (11) Moyne, J.; Del Castillo, E.; Hurwitz, A. M. *Run-to-Run Control in Semiconductor Manufacturing*; CRC Press: Boca Raton, FL, 2001.
- (12) Del Castillo, E. *Statistical Process Adjustment for Quality Control*; Wiley: New York, 2002.
- (13) Skogestad, S.; Postlethwaite, I. *Multivariable Feedback Control*; John Wiley & Sons: Chichester, U.K., 1996.
- (14) Strang, G. *Linear Algebra and Its Applications*, 3rd ed.; Harcourt Brace Jonovich: San Diego, CA, 1988.
- (15) Steward, G. W. *Introduction to Matrix Computations*; Academic Press: New York, 1973.
- (16) Edgar, T. F.; Butler, S. W.; Campbell, W. J.; Pfeiffer, C.; Bode, C.; Hwang, S. B.; Balakrishnan, K. S.; Hahn, J. Automatic control of microelectronics manufacturing: practices, challenges and possibilities. *Automatica* **2000**, *36*, 1567–1603.
- (17) Kaufman, F. B.; Thompson, D. B.; Broadie, R. E.; Jaso, M. A.; Guthrie, W. L.; Pearson, D. J.; Small, M. B. Chemical-mechanical polishing for fabricating patterned W metal features as chip interconnects. *J. Electrochem. Soc.* **1991**, *138*, 3460–3465.
- (18) Yao, C. H.; Feke, D. L.; Robinson, K. M.; Meikle, S. The influence of feature-scale surface geometry on CMP processes. *J. Electrochem. Soc.* **2000**, *147*, 3094–3099.
- (19) Chen, C. Y.; Yu, C. C.; Shen, S. H.; Ho, M. Operational aspects of chemical mechanical polishing: polish pad profile optimization. *J. Electrochem. Soc.* **2000**, *147*, 3922–3930.
- (20) Chiu, J. B.; Yu, C. C.; Shen, S. H. Application of soft landing to the process control of chemical mechanical polishing. *Microelectron. Eng.* **2003**, *65*, 345–356.
- (21) Kao, Y. C.; Yu, C. C.; Shen, S. H. Robust operation of copper chemical mechanical polishing. *Microelectron. Eng.* **2003**, *65*, 61–75.
- (22) Prasad, S.; Loh, W.; Kapoor, A.; Chang, E.; Stein, B.; Boning, D.; Chung, J. Statistical metrology for characterizing CMP processes. *Microelectron. Eng.* **1997**, *33*, 231–240.
- (23) Chiu, J. B.; Su, A. J.; Yu, C. C.; Shen, S. H. Planarization strategy of Cu CMP: interaction between of plated copper thickness and removal rate. *J. Electrochem. Soc.* **2004**, *151*, G217–G222.
- (24) Shiu, S. J.; Yu, C. C.; Shen, S. H. Multivariable Control of Multi-Zone Chemical Mechanical Polishing. *J. Vac. Sci. Technol. B* **2004**, *22*, 1679–1687.
- (25) Preston, F. W. The theory and design of plate glass polishing machines. *J. Soc. Glass Technol.* **1927**, *11*, 214–256.

Received for review September 5, 2004
 Revised manuscript received February 4, 2005
 Accepted April 15, 2005

IE0491544