Available online at www.sciencedirect.com

ScienceDirect

Control Engineering Practice 15 (2007) 1268–1279

CONTROL ENGINEERING PRACTICE

www.elsevier.com/locate/conengprac

Control relevant issues in semiconductor manufacturing: Overview with some new results

An-Jihh Su[a], Jyh-Cheng Jeng[a], Hsiao-Ping Huang[a], Cheng-Ching Yu[a,*], Shih-Yu Hung[b], Ching-Kong Chao[c]

[a]Department of Chemical Engineering, National Taiwan University, Taipei 106-17, Taiwan, ROC
[b]Department of Mechanical Engineering, Nan Kai Institute of Technology, Tsaotun, Nantou 542, Taiwan, ROC
[c]Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei 106-07, Taiwan, ROC

Received 24 July 2006; accepted 2 November 2006
Available online 29 December 2006

Abstract

The quality control of integrated circuit (IC) processing is becoming more and more important as the wafer becomes larger and the feature size shrinks. However, an advanced IC fabrication process consists of 300+ steps with scarce and usually difficult quality measurements. Thus product yield may not be realized until months into production while in-line measurements are available on the order of a millisecond. The series production nature and measurement setup lead to a unique process control problem. In this work, typical disturbances are explained and the possibility for inferential control is explored. This leads to a control architecture with multiple layers in a cascade structure. Next, the rapid thermal processing (RTP) is used to illustrate recipe generation and control structure design at the tool level. The resultant multivariable controller gives satisfactory setpoint tracking for a triangular-like temperature program. Effective delay in a feedback loop at the process level is also clarified which can be used to design a run-to-run controller or to prioritize the measurement queue for the metrology tool. In order to prolong the time between maintenance and to reduce rework, process trend monitoring of a tool is essential. Instead of using entire batch data, a key process variable is identified and an index is computed to capture dynamic behavior of the tool. An IC processing example is used to illustrate this approach and results clearly indicate that process trend is well predicted using the index-based time-series model. Finally, future research directions for improved semiconductor manufacturing are also described.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Semiconductor manufacturing; Process systems engineering; Virtual metrology; Wafer acceptance test; Rapid thermal processing; Control structure design

1. Introduction

The continuing miniaturization of integrated circuit (IC) components and the increasing number of functions and performance of single integrated circuit (IC) chip are the trend in the semiconductor industry. The quality control of the wafer is becoming more and more important as the wafer becomes larger (from 200 to 300 mm) and the feature size shrinks (from 350 to 90 nm). On the corporate level, improving yield is the only solution to remain competitive. Thus advanced equipment control and advanced process control (AEC/APC) have become standard practice in modern semiconductor manufacturing. Edgar et al. (2000) give a comprehensive review of the processes and control

Abbreviations: AEC, advanced equipment control; APC, advanced process control; ARIMA, auto-regressive integrated moving average; ARMA, auto-regressive moving average; CD, critical dimension; CPI, chemical process industry; d-EWMA, double exponentially-weighted moving average; EWMA, exponentially-weighted moving average; FB, feedback; FDC, fault detection and classification; FF, feedforward; IC, integrated circuit; KSI, key sensitive index; KST, key sensitive time slot (within a batch); KSV, key sensitive variable (may be computed from measurement); MPCA, multi-way principle component analysis; PM, preventive maintenance; R2R, run-to-run; RTA, rapid thermal annealing; RTP, rapid thermal processing; VM, virtual metrology; V-WAT, virtual WAT; WAT, wafer acceptance test

*Corresponding author. Tel.: +886 2 3365 1759; fax: +886 2 2362 3040.
E-mail address: ccyu@ntu.edu.tw (C.-C. Yu).

0967-0661/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.conengprac.2006.11.003

issues, Qin, Cherry, Good, Wang, and Harrison (2004) discuss the challenges in the IC industries, and Lewin, Lachman-Shalem, and Grosman (2006) explore process systems engineering (PSE) related issues in IC fabrication. Contrary to general understanding in chemical process industries (CPI), the AEC is generally concerned with keeping the equipment (unit operation in CPI terminology) in working condition and, in so doing, prolonging the time between maintenance and reducing rework. So, the AEC is synonymous with fault detection and classification (FDC) for the individual equipment. However, unlike chemical processes, an advanced IC fabrication may include 300–400 steps (or process units), and the success in a single step (equipment) certainly does not guarantee an acceptable wafer. The APC addresses the control issue from one step to another. Thus, feedforward (FF) and feedback (FB) control becomes important. The run-to-run (R2R) control is the typical element in the feedback loop, and controllers are integral-only (I-only) or double integrator (PI^2 control), and they are generally termed as exponentially weighted moving average (EWMA) or double EWMA (d-EWMA) algorithms, respectively. In chemical process control (CPC) terminology, the AEC can be viewed as the within batch control and fault detection and the APC is similar to batch-to-batch process control. The controllers used rarely go beyond PID types. One may wonder: "Why does such a hi-tech industry use seemingly low-tech control methodology?" The answer is quite simple: "We cannot fix (control) what we cannot detect (measure)." (Wang, 2004) However, the endeavor for yield improvement via improved process control can be seen throughout fabs worldwide. Currently, the AEC/APC symposium (Edgar, 2004; Hsu et al., 2004; Wang, 2004; Wu et al., 2005) is held in the USA, Europe, and Asia each year with hundreds of attendees to each conference, and they become the major events for APC division personnel from fabs worldwide. In fact, this is similar to the process control phenomena many witnessed in CPI 20 years ago. However, the approaches taken in the IC industries are quite different from those of the CPI for the following reasons: (1) scarce and sometimes difficult quality measurements, (2) multiple and iterative processing steps, (3) non-straightforward links between processing steps and product specification (e.g., for IC design), and (4) frequent tool maintenance.

In this paper, the process characteristics in IC fabrication are explained in Section 2 and opportunities in process control are explored. In Section 3, a specific tool, rapid thermal processing (RTP), is used to illustrate the tool level control problems. RTP is employed for various single-wafer thermal treatment processes including annealing, oxidation, cleaning, and chemical vapor deposition (Campbell & Knutson, 1992; Chao, Hung, & Yu, 2003a, b; Cho, Lee, Joo, & Lee, 2005; Gunawan, Jung, Seebauer, & Braatz, 2003, 2004; Huang, Yu, & Shen, 2000a, b; Huang, Liu, and Yu (2000c); Jung, Gunawan, Braatz, & Seebauer, 2003, 2004; Lee, Lee, Chin, Choi, & Lee (2001); Lord, 1988). Because wafers processed using RTP has the advantage of fast ramp-up and -down time as compared to conventional batch furnaces. The process level (or module-level) control, run-to-run control, with emphasis on metrology delay is also discussed. The preventive maintenance problem is studied in Section 4 via an industrial example followed by the conclusion.

2. Process characteristic

2.1. Disturbances

Similar to chemical process control, disturbance rejection is the major concern in semiconductor manufacturing. By disturbance rejection, it means maintaining the product quality in the face of process changes. Typical sources of process variations in IC fabrication include: (1) tool-induced disturbances which are generally known as process drift and/or process shift, (2) product-induced disturbance which typically comes from the IC foundry where high-mix products are manufactured, and (3) incoming disturbances which are often referred to as the variations which are a direct consequence of proceeding processing steps (Chen, Shiu, Yu, & Shen, 2005; Patel, Miller, Guinn, & Jenkins, 2000). Generally, some prior knowledge about the quality of the *incoming* wafers is available in semiconductor manufacturing processes. Thus, feedforward control (FF) or feed sequence arrangement can be devised to mitigate incoming disturbances (Chen et al., 2005). A similar approach can be applied to product-induced disturbance. Tool-induced disturbance is less frequently seen in chemical process control. Nano-scale-based operation generally requires an ultra-clean environment. A small contamination may lead to degraded tool performance. Thus, the weekly based maintenance is a norm in fabs as opposed to yearly based maintenance in chemical plants. It is never the less essential to maintain product quality under gradual degradation using feedback control (Chen & Guo, 2001; Qin, Scheid, & Riley, 2003; Sachs, Hu, & Ingolfsson, 1995).

2.2. Measurement

The product nature of IC makes quality measurement difficult, if not impossible. Unlike the product purity specification in chemical production, the product yield cannot be realized until the end of some 300+ processing steps. This implies the information of the yield will not be available until a *month* into production. The electrical performance of a wafer (die to be specific) cannot be tested till the end of the iteration for each metal layer. The electrical performance of a wafer is generally referred to as the wafer acceptance test (WAT) and the test results are available in the time-scale of a *week* (Fan, Guo, Chang, & Wei, 2000). The product yield is usually highly correlated to the WAT data. Generally, after each processing step, a quality measurement, generally denoted as the "metrology", becomes available. Nano-scale nature makes the measurement (metrology) difficult and the measuring

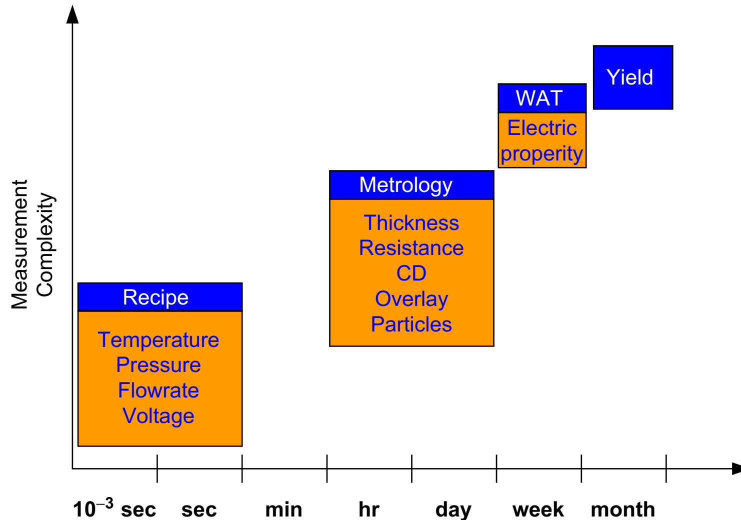

Fig. 1. Multiple-time-scale characteristic for the process and quality measurements.

station (metrology tool) expensive. The cost of a typical metrology tool is in the range of millions of dollars. This leads to a very different measurement setup as compared to chemical plants. That is: the metrology tool is *shared* by similar processing steps and only few of the wafers (1–4 wafers from each lot) are measured. The time-scale for a metrology measurement is in the order of hours to one day. This may result in delay problem if feedback control is installed. Typical metrology measurements include: thickness, resistance, critical dimension (CD), overlay, particles, etch rate, etc. Down to the tool level, the in-line measurements such as temperature, pressure, flow, current, etc., which are measured in the order of *milli-second to second*. Thus, quality/process variables are available on drastically different time scales and, obviously, the measurement complexity increases as one goes from the tool level to the product level (Fig. 1).

2.3. Control architecture

The ultimate goal of IC production is to improve the yield and, as pointed out earlier, process control is a means to achieve this. However, the process measurement setup in Fig. 1 reveals that effective control cannot be obtained without some type of inferential control (soft sensor in chemical engineering literature). The quality estimation can be further arranged into two tiers. One is at the tool level and the estimator is denoted as *virtual metrology* (Hsu et al., 2004). The other is at the product level which is generally called *virtual WAT* (Wu et al., 2005). Quality estimation is not unfamiliar to the chemical engineering community and it is often used to estimate product composition is a distillation column, molecular weight distribution in a polymerization reactor, etc. with certain degree of success. For example, in distillation, the relationship between product composition and tray temperature is governed by the thermodynamic equilibrium. Thus, a

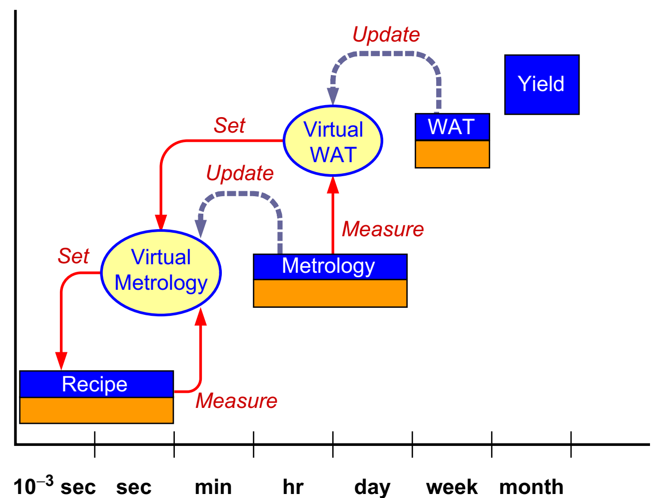

Fig. 2. Ideal control architecture for semiconductor manufacturing.

strong correlation between tray temperatures and composition can be established. However, the relationship between in-line measurements (e.g., temperature) and quality variable (e.g., sheet resistance) in semiconductor manufacturing is less obvious, especially when the tool is operated in a batch mode. A successful virtual metrology model relies on identifying key tool indices from the entire batch data. At the product quality level, few attempts have been made to relate end-of-line electrical properties to the metrology data over the entire process (Fan et al., 2000; Wu et al., 2005). Fig. 2 shows how the virtual metrology (VM) and virtual WAT (V-WAT) can be incorporated into the control architecture for improved yield management. Here, the estimated quality variable is maintained by changing the recipe (e.g., temperature set point) while the metrology model is updated when metrology data become available (e.g., via Kalman filtering). The electrical properties of a wafer can also be estimated at the completion of

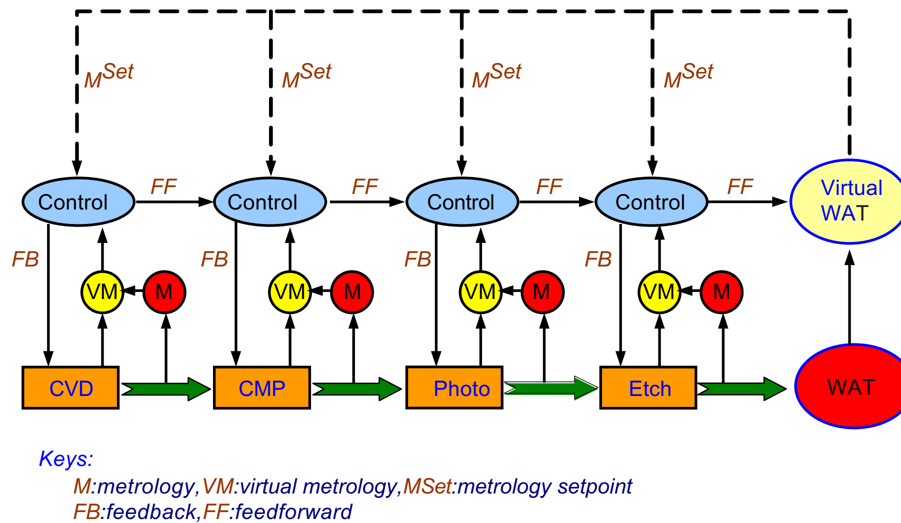

Fig. 3. Detailed feedforward/feedback control structure incorporated with virtual metrology and virtual WAT.

several processing steps using virtual WAT (Wu et al., 2005). Electrical properties of the product are controlled by adjusting metrology set points which subsequently affect the recipes in related tools. Fig. 3 gives a detailed description of the control architecture for product quality control. It is clear that the quality estimation (VM and V-WAT) plays a vital role in this framework. The series nature of the process flow leads to a feedforward/feedback (FF/FB) structure from a tool perspective provided with multiple layers of cascade control.

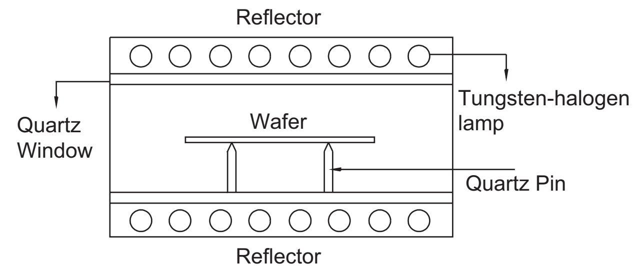

Fig. 4. Schematics of a simple RTP.

3. Control of RTP

Typically, wafer processing in a tool is described by a recipe which consists of on the order of 10 steps. These steps include: warm-up, temperature program, flow manipulation, cool-down, etc. Generally, very simple feedback control is used to ensure successful execution of the recipe. The rapid thermal processing (RTP) is used to illustrate the tool level control.

3.1. Process

RTP is an effective tool for various single-wafer thermal treatment processes. It permits processes to be accomplished with minimal dopant redistribution and uniform deposition quality with a smaller thermal budget. However, poor RTP system design can lead to significant temperature differences in the wafer. One of the main shortcoming that RTP must overcome is that of heating (or cooling) the wafers non-uniformly which results in material failure due to an increase in thermal stresses or serious warpage. The damage due to the presence of thermal stresses can represent a limit on the applicability of rapid thermal processing.

The temperature non-uniformity in the wafer is caused by three factors: edge effect, pattern effect, and heat source. The higher heat loss from the wafer edge has been found to result in a radial temperature gradient in the wafer. To improve the wafer temperature non-uniformity produced by the edge effect, several radiative shields can be placed at the edge of the wafer to reduce the heat loss from the wafer edge and reflect the radiative energy back into the wafer during the cooling process (Lord, 1988). By varying the angle of the shield, an optimal shield configuration can be found to minimize the induced thermal stress (Young & McDonald, 1990). Hebb and Jensen (1998) show that pattern-induced temperature non-uniformity can cause plastic deformation during a RTP cycle and the problem is exacerbated by single-side heating, increased processing temperature and ramp rate. Design and control of RTP to improve temperature uniformity was explored by Huang, Yu, and Shen (2000a, b) and Huang, Liu, and Yu (2000c).

A cross-sectional view of the furnace and wafer is shown in Fig. 4. A bank of tungsten halogen lamps provides the thermal radiative energy to the single silicon wafer through a transparent quartz window. Since quartz does not absorb light efficiently within the wavelength band of the lamps, it can be neglected in the thermal system. Let us assume the wafer is 200 mm in diameter held by three quartz pins and enclosed in a cylindrical chamber, where the chamber is axis-symmetric in geometry (Chao et al., 2003a, b; Huang

et al., 2000a, b). The chamber geometry is described in Huang et al. (2000a).

3.2. Recipe generation

The essential step in the RTP recipe, in addition to preparation steps, is the temperature program. Two types of temperature programs are often used in RTP: soak and spike temperature profiles. Consider the spike annealing of rapid thermal annealing (RTA). The post-implant annealing uses a lamp-based RTA with a program such as that shown in Fig. 5A. As pointed out by Jung et al. (2003), the ion-implantation technology is limited in part by transient-enhanced diffusion (TED) of dopants during RTA, often leading to significant spreading of the dopant profile. This may lead to defects in extremely shallow p–n junctions in electronic devices. Considerable efforts have been put forth to design a temperature program to produce the desired junction depth while maintaining low sheet resistance (Gunawan et al., 2004). A different approach is taken here. The spike annealing is used to illustrate thermal-stress-based temperature program generation with emphasis on the cooling curve.

Consider the RTP system in Fig. 4. The wafer thickness is assumed to be thin as compared to the radius of the wafer r_o, so it can be regarded as a one-dimensional plane-stress problem, that is, the temperature T is dependent on r only. The partial differential equations of the present thermoelastic problem can be written as (Nowinski, 1978):

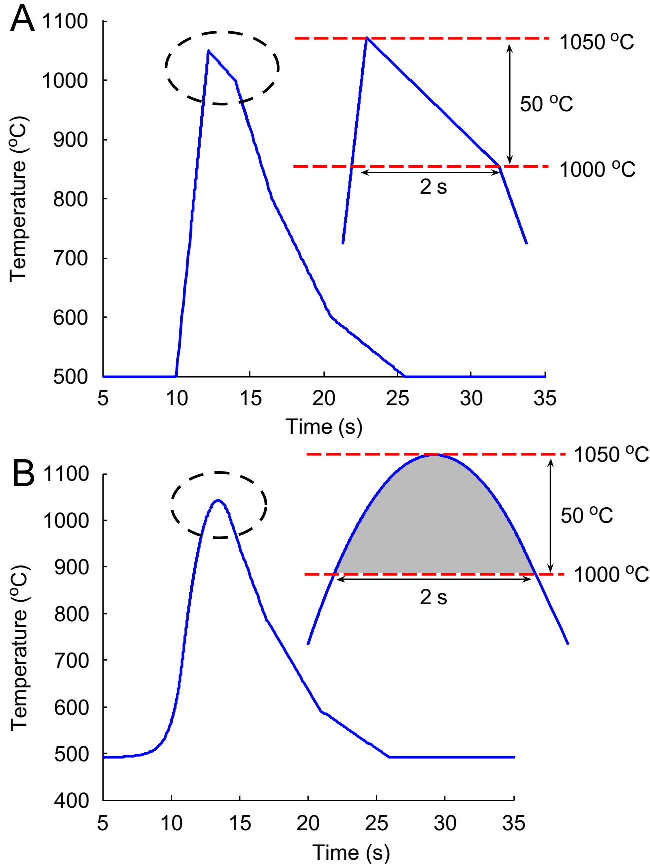

Fig. 5. Triangular (A) and smooth (B) temperature programs for spike annealing.

$$k\left(\frac{1}{r}\frac{\partial T}{\partial r}+\frac{\partial^2 T}{\partial r^2}\right)-q^{rad}-q^{conv}=\rho C_p\frac{\partial T}{\partial t} \tag{1}$$

with boundary conditions given by

$$\frac{\partial T}{\partial r}=0 \quad \text{at } r=0, \tag{2}$$

$$-k\frac{\partial T}{\partial r}=q_{edge} \quad \text{at } r=r_o, \tag{3}$$

where ρ, C_p and k are the density, specific heat capacity, and thermal conductivity of silicon, respectively. q^{rad} and q^{conv} represent the radiative and convective heat flux leaving a wafer surface per unit wafer volume, respectively. The quantity q_{edge} is the heat flux at the wafer edge that includes the heat loss of convection and radiation.

Once the temperature profile has been obtained, the components of stresses are obtained as

$$\sigma_{rr}=\alpha E\left(\frac{1}{r_o^2}\int_0^{r_o}T(\eta)\eta\,d\eta-\frac{1}{r^2}\int_0^{r}T(\eta)\eta\,d\eta\right), \tag{4}$$

$$\sigma_{\theta\theta}=\alpha E\left(-T+\frac{1}{r_o^2}\int_0^{r_o}T(\eta)\eta\,d\eta+\frac{1}{r^2}\int_0^{r}T(\eta)\eta\,d\eta\right), \tag{5}$$

$$\sigma_{r\theta}=0, \tag{6}$$

where σ_{rr} and $\sigma_{\theta\theta}$ are the radial and tangential stress components, respectively. α and E denotes the linear thermal expansion coefficient and Young's modulus, respectively. Since the obtained temperature profile is expressed in a discrete manner, the stresses in Eqs. (4) and (5) are determined by a trapezoidal integration technique.

In the present study, the maximum shear stress failure criterion is assumed that the wafer fails in shear when

$$S=\frac{\tau_{\max}\cdot F_s}{\tau_{yp}}>1, \tag{7}$$

where S is the normalized maximum resolved stress, F_S is the safety factor which is usually taken to be 2 and the maximum shear stress is calculated using Mohr's circle as (Boley & Weiner, 1960):

$$\tau_{\max}=\frac{1}{2}|\sigma_{rr}-\sigma_{\theta\theta}|. \tag{8}$$

At high temperature, silicon behaves like a viscous material. The yield stress in shear can be expressed in terms of the temperature and the maximum shear stress rate (Fan & Qiu, 1997; Hebb & Jensen, 1998; Lord, 1988) as

$$\tau_{yp}=23.17\exp(16.1-0.00916T)\left(\frac{d\tau}{dt}\right)^{0.4}, \tag{9}$$

where the stress unit is in Pascal and the temperature unit is in degree Celsius. The stress rate $d\tau/dt$ is taken to be the larger of 2.5×10^5 Pa/s or its calculated value. If the result calculated from Eq. (9) exceeds 3.1×10^8 Pa, it is taken to be 3.1×10^8 Pa which means that the wafer is at low temperature. From Eq. (9) indicates that the yield shear stress will be about 1.5 MPa when $T = 1000\,^{\circ}$C at the beginning of the cooling process which is far less than 310 MPa at the room temperature $T = 27\,^{\circ}$C. This simply indicates that, according to the failure criterion stated in Eq. (7), a small temperature non-uniformity may induce material failure at high temperature. Because no analytical solution is available for the present problem, the numerical solutions are sought to the above governing equations. The calculations are carried out by using a fully implicit finite difference method (Chao et al., 2003a).

Three cooling scenarios are considered: (1) *Fixed temperature-difference scheme*: the maximum temperature difference within a wafer is fixed to 0.7 °C (by trial and error such that the normalized maximum resolved stress is less than one), (2) *Constant cooling-rate scheme*: the lamp powers decrease gradually at a constant rate of 10 KW/m^2 s (by trial and error which ensures that the normalized maximum resolved stress is less than one), (3) *Maximum stress scheme*: the normalized maximum resolved stress is kept right below one until the lamp powers become zero.

Based on the maximum shear stress failure criterion, the calculated results show that material failure always occurs at the edge of the wafer at the beginning of cooling processes. Furthermore, the maximum stress scheme is shown to be more efficient because it can significantly reduce the required cooling time and thermal budget. Thus, the conventional constant cooling-rate control scheme or linear temperature ramp-down scheme is not appropriate for RTP.

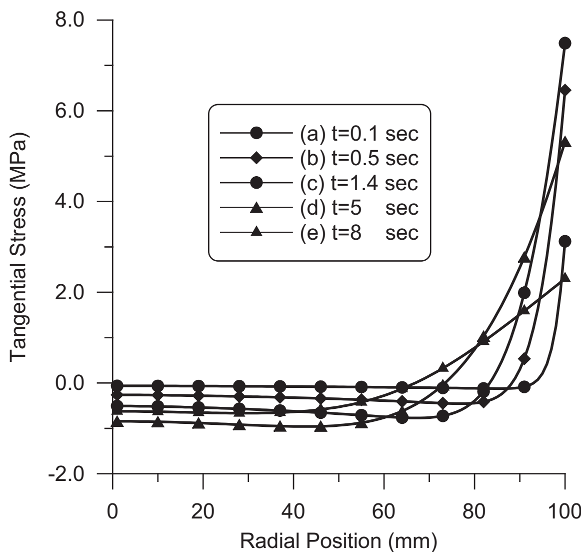

Fig. 6. The tangential stress distribution on wafer for the room temperature (top and bottom of the oven) cooling condition.

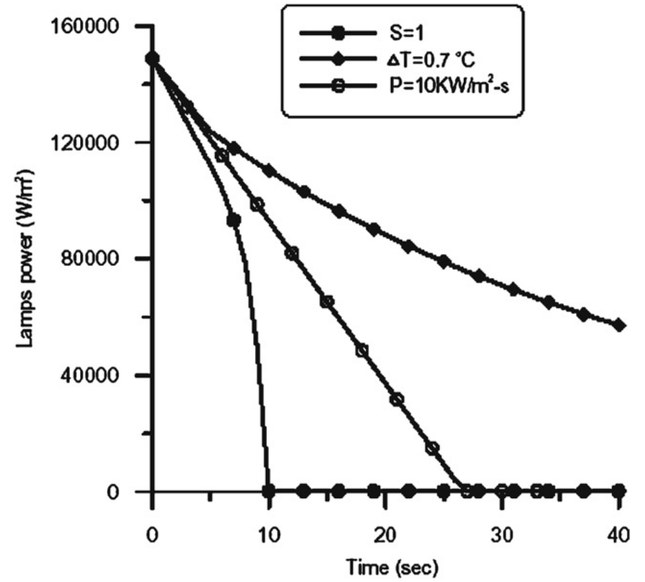

Fig. 7. Variation of the lamp's power for three different control schemes.

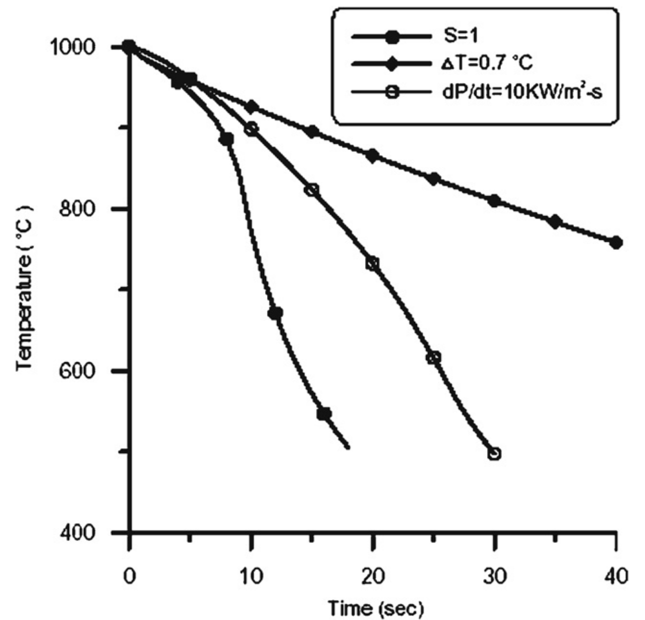

Fig. 8. The temperature variation at wafer edge under three different control schemes.

Fig. 6 shows that the tangential stress at the wafer edge is positive due to thermal shrinkage induced by the edge effect. On the other hand, compressive tangential stress prevails at the central region of wafer. Fig. 7 shows lamp powers decrease dramatically during the cooling process for the maximum stress control scheme. After 5 s have elapsed, lamp powers for the fixed temperature-difference scheme decrease gradually with a rate even smaller than the constant cooling-rate scheme. The required cooling time for the maximum stress control scheme is only 18 s from 1000 to 500 °C, compared to 30 s for the constant cooling-rate control scheme, and, moreover, it is only one-fifth of

the required time for the constant temperature-difference scheme as shown in Fig. 8. This provides an attractive alternative for temperature program generation.

3.3. Control structure design

The state-of-the-art RTP typically consists of seven lamp-heating zones with seven temperature measurements, in addition to computed emissivity (emissivity is calculated with the combination of measurements and equations). Here a simple RTP model (Huang et al., 2000a) is used to illustrate the essential steps in the control structure design. This is an RTP system with three lamp-heating zones for a 200 mm wafer. Once a temperature program becomes available (Fig. 5A), the design procedure consists of the following steps: (1) selection of temperature measurements, (2) controller design, and, possibly, (3) temperature program modification. Spike annealing is considered here. The control objective is to maintain temperature uniformity, especially around the peak temperature. The focus of the program is the temperature range of 1000–1050 °C with the duration of ~2 s.

The temperature profile along the radial position plays an important role for the measurement selection. The RTP system uses a linear combination of *three* lamp powers to match the desired intensity. Notice that each lamp ring has an intensity profile similar to the normal distribution (e.g., Fig. 5). The optimal temperature uniformity corresponds to a unique lamp powers combination. The desired temperature profile is a nonlinear function in r and it crosses the temperature set point several times. The profile is similar to a high-order polynomial: $T - T^{set} = \prod(r - z_i)$ where T^{set} is the temperature set point, n is the number of set point crossings and z_i denotes the location of the set point crossing (zero of the polynomial). Note that the best temperature uniformity that can be achieved is obtained by minimizing the square error between the set point and temperature profile. This is termed the *desired* temperature profile hereafter. Furthermore, the easiest way to maintain this profile is to keep the temperatures already at (or close to) set point under control. This can be interpreted as retaining the shape of the temperature profile by holding several key positions at the set point. If the zero-crossing temperatures is greater than manipulated inputs, the next step is to check system interaction and inherent robustness using the structured singular value (SSV; Morari & Zafiriou, 1989). Therefore, the temperature measurement selection criterion can be summarized as follows (Huang et al., 2000c):

(1) Identity the set point crossing locations for the desired temperature profile.
(2) Prefer the approximately equal-spaced rule for placing temperature measurements on these locations.
(3) Check for system robustness, and if the SSV is not acceptable, go back to step 2.

The procedure suggests control T_3, T_{17}, and T_{29} out of 30 zones in the radial position.

Once the control structure is determined, the next step is to design a multivariable temperature controller. The conventional PID controller is preferred for its simplicity and transparency. But almost half of the batch cycle involves ramp-type setpoint trajectory (e.g., ramp-up and cool down), so the IMC design principle of Morari and Zafiriou (1989) is employed (Huang et al., 2000a) and type-2 system is considered. For the RTP operated at 1050 °C, the model gives the following process transfer function matrix:

$$G(s) = K \cdot \mathrm{diag}(1/(\tau_i s + 1)), \tag{10}$$

where K is the steady-state gain matrix and τ_i is the time constant. Following the design procedure of Huang et al. (2000a), it leads to a diagonal PID type of controller with a static decoupler. Moreover, the diagonal controller has *double* integrators.

$$C(s) = K^{-1} diag(K_{ii}), \tag{11}$$

where K_{ii} is the diagonal PID type of controller.

$$K_{ii} = K_{c,i}\left(1 + \frac{1}{\tau_{I,i}s} + \tau_{D,i}s\right)\frac{1}{s}. \tag{12}$$

It is denoted as PI^{2}D controller, hereafter. The controller parameters can be expressed in terms of IMC filter time constant τ_f.

$$K_{c,i} = \frac{\tau_i + 2\tau_f}{\tau_f^2}, \tau_{I,i} = \tau_i + 2\tau_f, \tau_{D,i} = \frac{2\tau_i\tau_f}{\tau_i + 2\tau_f}. \tag{13}$$

Therefore, once the closed-loop time constant τ_f is set, the tuning constants for the PI^{2}D controller can be determined immediately. In this work it is taken as 1/4 of the open-loop time constant.

Fig. 9A clearly indicates the advantage of PI^{2}D control, derived from type-2 disturbance, over PI control, derived from type-1 disturbance, in which significant offsets are observed in ramp-up and ramp-down periods. If a PID controller is implemented, there will be still offsets because that a single integrator cannot deal with a ramp input signal. And the manipulated variable changes more violent than PI controller. Moreover, the two important criteria, peak temperature and duration time over 1000 °C, are completely missed, even with PI^{2}D control. Table 1 summarized the spread of the peak temperature and duration time. It should be emphasized here that computing cost of controller is critical for RTP control, because the state-of-art RTP system has an extremely fast ramp-up rate (~250 °C/s) and the sampling time for RTP is relatively short (~0.01 s). Instead of using Model Predictive Control (MPC; Dassau, Grosman, & Lewin, 2006), an inverse-based is designed here and it is implemented in a PID form. In doing this, effective control can be achieved with much efficient computing cost.

If the peak temperature tracking and duration is the design criteria, the triangular temperature problem in

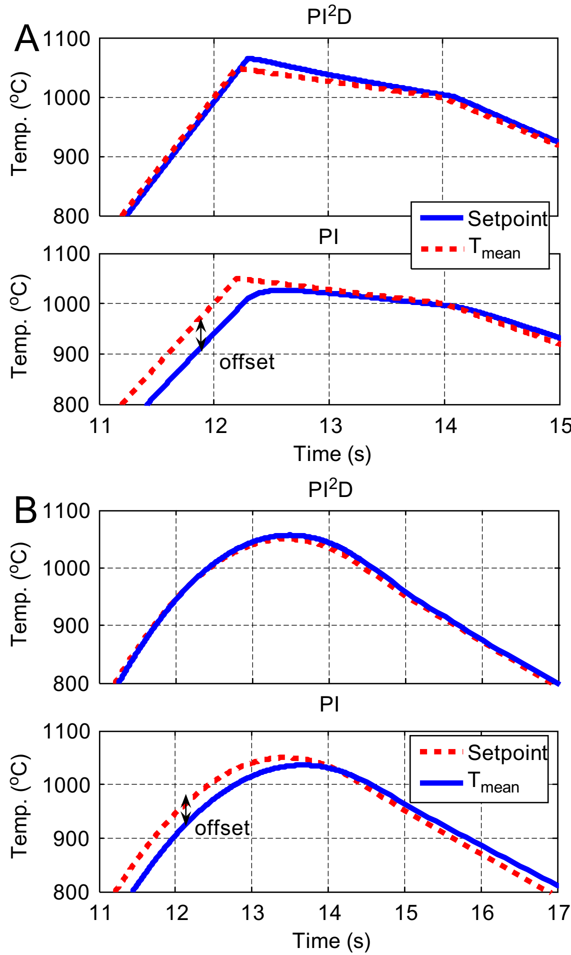

Fig. 9. Setpoint tracking for triangular (A) and smooth (B) temperature program using PI and PI^{2}D control.

Table 1
Control performance of different types of temperature programs

	Triangular	Smooth
Mean of peak temp. (°C)	1066.1	1056.9
Range of peak temp. (°C)	11.8	6.9
Std. dev. of peak temp. (°C)	3.8	2.2
Mean of duration (s)	2.08	2.08
Range of duration (s)	0.155	0.086
Std. dev. of duration (s)	0.039	0.027

Fig. 5A cannot be achieved with a realizable controller. Thus, a smooth temperature program is used instead as shown in Fig. 5B. The results clearly indicate that satisfactory peak temperature tracking can be obtained while keeping the duration nearly constant (Fig. 9B). The tabulated results in Table 1 also confirm this and the peak temperature spread is reduced to 6.9 °C as compared to 11.8 °C for triangular temperature program. Fig. 10 shows the transient responses of wafer temperatures across the

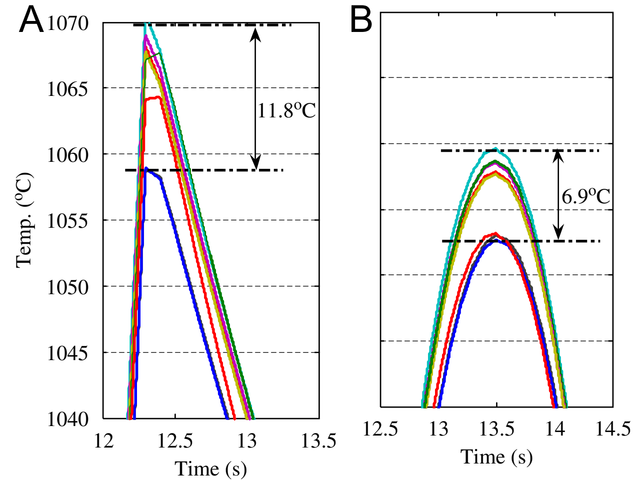

Fig. 10. Temperature spreads for triangular (A) and smooth (B) temperature program using PI^{2}D control.

radial positions. The results presented here clearly indicate that the advanced control methodology can certainly be applied to semiconductor manufacturing at the tool level. However, it should be emphasized here that a single step success does not guarantee a yield improvement.

3.4. Run-to-run control

Run-to-run (R2R) control is becoming a standard practice in critical processing steps. However, the iterative process configuration and shared metrology tool practice may complicate the process dynamics for this discrete-event system. It is well known in the control literature that time-delay places limit to achievable control performance. Thus, the effective delay unit (D_{eff}) in a feedback loop should be correctly computed and it is generally not equal to the time needed for metrology measurement. The following analyses can be applied to either lot-to-lot control or wafer-to-wafer control. First, the term "runs" corresponds to wafers in the case of wafer-to-wafer control and lots in the case of lot-to-lot control. Sampling interval, N_s, means one out of N_s runs will be measured or the next N_sth run will be measured after a measured run. Thus, N_s is an integer and has a minimum value of unity. Metrology delay, N_m, denotes the metrology data from a measured run will be available for feedback to the next N_mth run. Because of this definition, the metrology delay should not only include the measurement time, but also the queue time and material transportation time. Thus, the minimum value of N_m is unity and N_m is also an integer. When the sampling interval (N_s) and metrology delay (N_m) are given, it can be shown that the effective delay unit becomes:

$$D_{eff} = round_{up}(N_m/N_s). \quad (14)$$

$round_{up}(\cdot)$ means round the number up to the next integer. Eq. (14) is useful in designing R2R control using metrology measurement or to prioritize measurement queue.

4. Process monitoring

Process monitoring and analysis are important in semiconductor manufacturing (Wise, Gallagher, Butler, White, & Barna, 1999). Correct trend monitoring can be used to determine appropriate timing for preventive maintenance. In this section, instead of incorporating large number of trajectory data with variable batch time and possibly "missing" data for some process variables using multivariable statistic technique (Kourti, 2003; Kourti, Nomikos, & MacGregor, 1995; Louwerse & Smilde, 2000; Nomikos & MacGregor, 1994), a key sensitive index (KSI) based approach is proposed for batch process trend monitoring. From process insight or the experience of the process operator, a certain period time within a batch where measurements have significant effect on product quality, the key sensitive time-slot (KST), is identified. Next, based on the KST, possible key sensitive process variables (KSV) are chosen. The KSV may not be measured values themselves in KST, but some quantity, such as area, slope, maximum, etc., computed from the raw measurements. Once a KSV is computed for each batch (wafer-to-wafer) under normal operation, its autocorrelation function is calculated as the batch process progresses. If significant autocorrelation is found, a time-series model is established for the selected KSV, if not, a different KSV is sought. With the time-series model, the process trend can thus be forecasted and then index for process operating status (key sensitive index, KSI) is defined and computed. By monitoring the KSI, possible maintenance action can therefore be called for, whenever necessary. This provides dynamical capability for process trend monitoring while maintaining the simplicity of single-variate analysis. A rapid thermal processing example is used to illustrate the KSI-based approach.

In the manufacturing of semiconductor, IC is processed through the recipes which comprise a sequence of different treatments (steps). In general, only some steps are critically related to the product quality so that the processing intervals corresponding to these critical steps are the aforementioned KST. In this example, the recipe comprises 11 steps where the processing time from step 6 to 10 is identified as KST. Then, three important process variables are selected as possible candidates for computing KSV. From correlation analysis, only the maximum of one variable (say, variable A) in KST shows significant autocorrelation and, hence, this maximum value, $A_{\max}$, is chosen as KSV. However, as shown in Fig. 11A, $A_{\max}$ for some batches is abnormally greater than the average value. Since different products are produced with the same tool, $A_{\max}$ with particularly high values may be the consequence of different product specification. Thus, a product index, as shown in Fig. 11B, is employed to modify $A_{\max}$ values. The result of modified $A_{\max}$, designated as $A'_{\max}$, is shown in Fig. 11C. It clearly shows that the variance of $A'_{\max}$ is much smaller and the process trend is easier to predict. Consequently, an autoregressive moving average (ARMA) model of the following is built for $A'_{\max}$ based on measurements from 500 wafers.

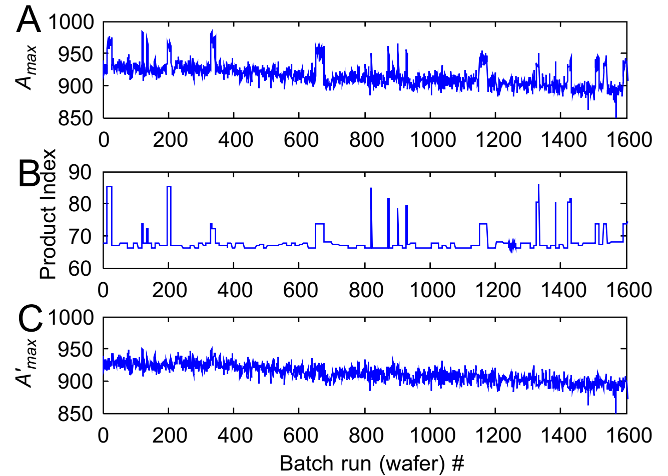

Fig. 11. Key sensitive variable (KSV) without considering the product index (A), product index itself (B), and considering the product index (C).

$$(1-1.744q^{-1}+0.776q^{-2})A'_{\max}(t) = (1-1.346q^{-1}+0.476q^{-2})e(t), \quad (15)$$

where q^{-1} is the backward shift operator and $e(t)$ is white noise. It is found that one root of the autoregressive polynomial is close to unity, which means that the time series of $A'_{\max}$ exhibits non-stationary behavior. For this reason, an autoregressive integrated moving average (ARIMA) model is then built as the following to describe this behavior.

$$(1+0.942q^{-1})\nabla A'_{\max}(t) = (1+0.452q^{-1}-0.553q^{-2})e(t), \quad (16)$$

where $\nabla = (1-q^{-1})$. These two time-series models are then used for forecasting the values of $A'_{\max}$ as the batch process progresses. The result is shown in Fig. 12 where two abrupt changes are observed due to the scheduled tool preventive maintenance (PM). Initially, both the forecasts of ARMA and ARIMA models can follow the process trend well. However, as the batch process progresses, the forecast of ARMA model starts to deviate from the actual $A'_{\max}$ and gets worse toward the end of each period (right before PM), while the forecast of ARIMA model maintains acceptable tracking. This phenomenon disappears after PM and then can be observed again as the batch process progresses. In order to capture the drifting behavior of this batch process, the KSI is thus defined as the absolute value of difference between residuals of these two models.

$$KSI = |Residual_{ARMA} - Residual_{ARIMA}|. \quad (17)$$

The computed KSI is shown in Fig. 13. The results clearly indicate that the process trend can be realized using the proposed KSI and tool maintenance is required once this KSI is greater than a prescribed limit. Therefore, this

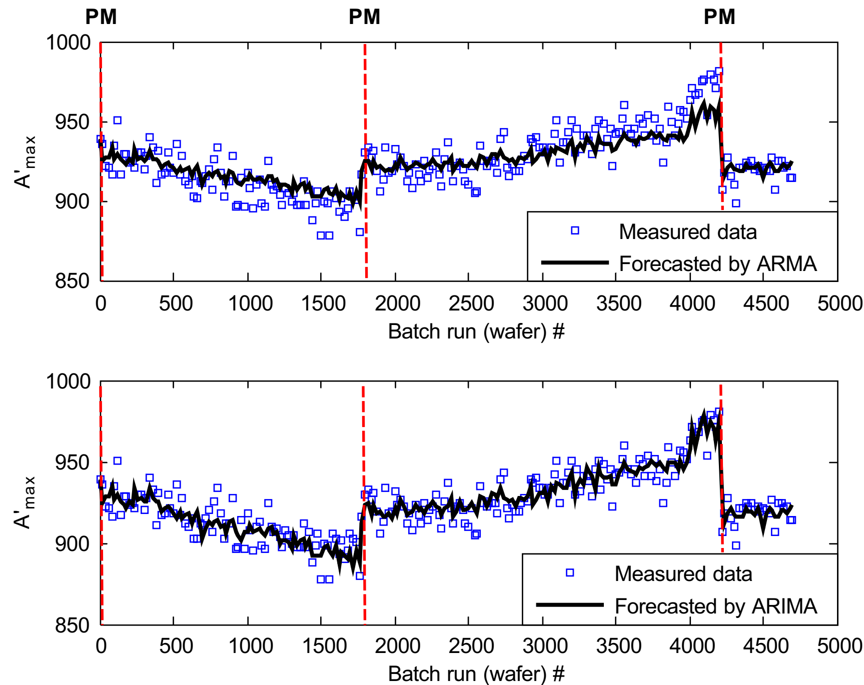

Fig. 12. Comparison of ARMA and ARIMA prediction as compared to the true measurement (only one out of 20 measurements shown for better resolution; the dash lines indicating the time for preventive maintenance, PM).

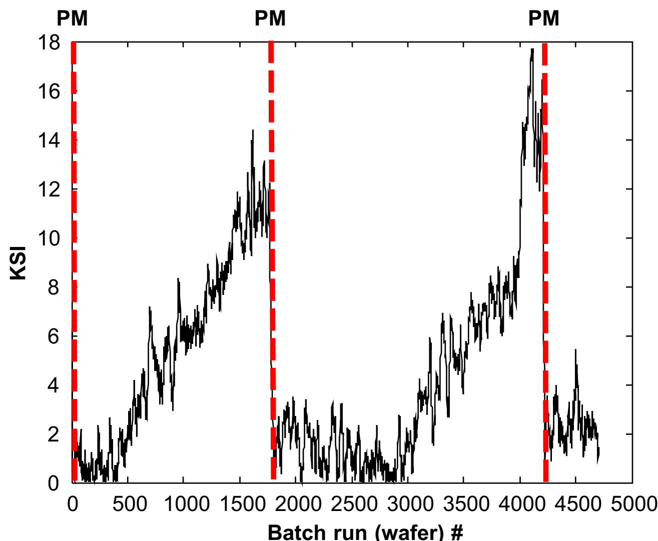

Fig. 13. The resultant key sensitive index (KSI) and corresponding maintenance time (PM as indicated by dash lines).

KSI-based approach not only can be used for batch process trend monitoring, but also it is helpful for the engineers to decide when to call for tool maintenance.

5. Conclusion and challenges

An advanced IC fabrication consists of 300+ steps with scarce and usually difficult quality measurements. The series production nature and measurement setup lead to a unique process control problem. In this work, typical disturbances in semiconductor manufacturing are explained and the necessity of quality estimation is outlined. This leads to a control architecture with multiple layers in cascade structure. Next, rapid thermal processing (RTP) is used to illustrate the recipe generation and control structure design at the tool level. The resultant multivariable controller gives satisfactory setpoint tracking for a triangular-like temperature program. Effective delay in a feedback loop at the process level is also clarified which can be used to design a run-to-run controller or to prioritize the measurement queue for the metrology tool. In order to prolong the uptime and to reduce rework, process trend monitoring of a tool is essential. Instead of using entire batch data, a key process variable is identified and an index is computed to capture dynamic behavior of the tool. An IC processing example is used to illustrate this approach and results clearly indicate that the process trend is well predicted using the index-based time-series model.

Despite endeavors for improved semiconductor manufacturing, challenges remain. At the tool level, new sensors are needed for continuing improvement in *within-wafer non-uniformity* (WIWNU) and, to achieve this goal, new actuators should be incorporated to maintain the uniformity. From control perspective, this leads to a distributed parameter systems at the equipment level. At the *wafer-to-wafer* level, continuing improvement on APC is necessary. Key issues in controller design include: (1) handling variable model parameters and (2) handling mixed products. Unlike traditional process industries, estimation plays a vital part for improved process yield.

For tool level estimation (virtual metrology), key process parameters should be identified for metrology estimation. If such a variable does not exist, new sensor should be added to obtain VM. Moreover, the model should be able to accommodate parts replacement and frequent preventive maintenance. At the *module level* (virtual WAT), it is even more important to identify key parameters in essential tools, because so many data are available for estimation and only few of them play critical role in determining electrical properties. Certainly, the use of VM and V-WAT to ensure successful IC fabrication is also important. This is especially true for a series production structure with 300+ steps.

Acknowledgment

This work was supported in part by the National Science Council of Taiwan. The continuing support and discussion with Sunny Wu, B.H. Chen, J.S. Lin, Henry Lo, Jean Wang, C.H. Yu, and M.S. Liang of TSMC are appreciated. Long-term collaboration with Walters Shen of AMAT is also gratefully acknowledged. Inputs from Jeff Ward is also acknowledged.

References

Boley, B. A., & Weiner, J. H. (1960). *Theory of thermal stresses*. New York: Wiley.

Campbell, S. A., & Knutson, K. L. (1992). Transient effects in rapid thermal processing. *IEEE Transactions on Semiconductor Manufacturing*, *5*, 302.

Chao, C. K., Hung, S. Y., & Yu, C. C. (2003a). Thermal stress analysis for rapid thermal processor. *IEEE Transactions on Semiconductor Manufacturing*, *13*, 335.

Chao, C. K., Hung, S. Y., & Yu, C. C. (2003b). Effect of lamp radius on thermal stresses for rapid thermal processing system. *ASME Journal of Manufacturing Science and Engineering*, *125*, 504.

Chen, A., & Guo, R. S. (2001). Age-based double EWMA controller and its application to CMP processes. *IEEE Transactions on Semiconductor Manufacturing*, *14*, 11.

Chen, Y. H., Shiu, S. J., Yu, C. C., & Shen, S. H. (2005). Batch sequencing for run-to-run control: Application to chemical mechanical polishing. *Industrial and Engineering Chemistry Research*, *44*, 4676.

Cho, M., Lee, Y., Joo, S., & Lee, K. S. (2005). Semi-empirical model-based multivariable iterative learning control of an RTP system. *IEEE Transactions on Semiconductor Manufacturing.*, *18*, 430.

Dassau, E., Grosman, B., & Lewin, D. R. (2006). Modeling and temperature control of rapid thermal processing. *Computers and Chemical Engineering*, *30*, 686.

Edgar, T.F. (2004). Multi-product run-to-run control for high-mix fabs. *AEC/APC symposium*. HsinChu, Asia, December.

Edgar, T. F., Butler, S. W., Campbell, W. J., Pfeiffer, C., Bode, C., Hwang, S. B., et al. (2000). Automatic control of microelectronics manufacturing: Practices, challenges and possibilities. *Automatica*, *36*, 1567.

Fan, C. M., Guo, R. S., Chang, S. C., & Wei, C. S. (2000). SHEWMA: An end-of-line SPC scheme using wafer acceptance test data. *IEEE Transactions on Semiconductor Manufacturing*, *13*, 344.

Fan, Y. H., & Qiu, T. (1997). Analyses of thermal stresses and control schemes for fast temperature ramps of batch furnaces. *IEEE Transactions on Semiconductor Manufacturing*, *10*, 433.

Gunawan, R., Jung, M. Y. L., Seebauer, E. G., & Braatz, R. D. (2003). A maximum a posteriori estimation of transient enhanced diffusion energetics. *American Institute of Chemical Engineers*, *49*, 2114.

Gunawan, R., Jung, M. Y. L., Seebauer, E. G., & Braatz, R. D. (2004). Optimal control of rapid thermal annealing in a semiconductor process. *Journal of Process Control*, *14*, 423.

Hebb, J. P., & Jensen, K. F. (1998). The effect of patterns on thermal stress during rapid thermal processing of silicon wafers. *IEEE Transactions on Semiconductor Manufacturing*, *11*, 99.

Huang, C. J., Yu, C. C., & Shen, S. H. (2000a). Selection of measurement location for the control of rapid thermal processor. *Automatica*, *36*, 705.

Huang, C. J., Yu, C. C., & Shen, S. H. (2000b). Identification and nonlinear control for rapid thermal processor. *Journal of Chinese Institute of Chemical Engineers*, *31*, 585.

Huang, I., Liu, H. H., & Yu, C. C. (2000c). Design for control: Temperature uniformity in rapid thermal processor. *Korean Journal of Chemical Engineering*, *17*, 111.

Hsu, C.W., Hung, M.Y., Chen, B.C., Cheng, Y.J., Hsu, C., & Lai, J.H. (2004). The challenges of implementing APC in foundry. *AEC/APC symposium*. USA XVI, Westminster, CO, September.

Jung, M. Y., Gunawan, R., Braatz, R. D., & Seebauer, E. G. (2003). Ramp-rate effects on transient enhanced diffusion and dopant activation. *Journal of the Electrochemical Society*, *150*, G838.

Jung, M. Y., Gunawan, R., Braatz, R. D., & Seebauer, E. G. (2004). A simplified picture for transient enhanced diffusion of boron in silicon. *Journal of the Electrochemical Society*, *151*, G1.

Kourti, T. (2003). Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-up and grade transitions. *Journal of Chemometrics*, *17*, 93.

Kourti, T., Nomikos, P., & MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control*, *5*, 277.

Lee, K. S., Lee, J., Chin, I., Choi, J., & Lee, J. H. (2001). Control of wafer temperature uniformity in rapid thermal processing using an optimal iterative learning control technique. *Industrial and Engineering Chemistry Research*, *40*, 1661.

Lewin, D.R., Lachman-Shalem, S., & Grosman, B. (2006). The role of process system engineering (PSE) Applications in integrated circuit (IC) manufacturing, Control Engineering Practice (doi:10.1016/j.conengprac.2006.04.003).

Lord, H. A. (1988). Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven. *IEEE Transactions on Semiconductor Manufacturing*, *1*, 105.

Louwerse, D. J., & Smilde, A. K. (2000). Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science*, *55*, 1255.

Morari, M., & Zafiriou, E. (1989). *Robust process control*. Englewood Cliffs, NJ: Prentice-Hall.

Nomikos, P., & MacGregor, J. F. (1994). Monitoring of batch processes using multi-way principle component analysis. *American Institute of Chemical Engineers*, *40*, 1361.

Nowinski, J. L. (1978). *Theory of thermoelasticity with application*. Sijthoff and Noordhoff.

Patel, N. S., Miller, G. A., Guinn, C., & Jenkins, S. T. (2000). Device dependent control of chemical-mechanical polishing of dielectric films. *IEEE Transactions on Semiconductor Manufacturing*, *13*, 331.

Qin, S.J., Cherry, G., Good, R., Wang, J., & Harrison, C.A. (2004). *Control and monitoring of semiconductor manufacturing processes: Challenges and opportunities*. DYCOPS-7, Boston, July.

Qin, S. J., Scheid, G. W., & Riley, T. J. (2003). Adaptive run-to-run control and monitoring for a rapid thermal processor. *Journal of Vacuum Science and Technology B*, *21*, 301–310.

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

Wang, T. (2004). Advanced process control road map and challenges. *AEC/APC symposium*. HsinChu, Asia, December.

Wise, B. M., Gallagher, N. B., Butler, S. W., White, D. D., & Barna, G. G. (1999). A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in semiconductor etch process. *Journal of Chemometrics*, *13*, 379.

Wu, S., Chen, P.H., Lin, J.S., Ko, F., Lo, H., Wang, J., et al. (2005). Real-time device performance prediction for 90 nm and beyond. *AEC/APC symposium*. Palm Spring, CA, USA, September.

Young, G. L., & McDonald, K. A. (1990). Effect of radiation shield angle on temperature and stress profiles during rapid thermal annealing. *IEEE Transactions on Semiconductor Manufacturing*, *3*, 176.