ADCHEM 2006

International Symposium on Advanced Control of Chemical Processes
Gramado, Brazil – April 2-5, 2006

IFAC

PSE Relevant Issues in Semiconductor Manufacturing: Application to Rapid Thermal Processing

Cheng-Ching Yu[1†], An-Jhih Su[1], Jyh-Cheng Jeng[1], H. P. Huang[1], Shih-Yu Hung[2] and Ching-Kong Chao[2]

Dept. of Chem. Eng., National Taiwan University.[1], Taipei 106, TAIWAN
Dept. of Chem. Eng., National Taiwan University of Sci. Tech. [2], Taipei 106, TAIWAN

Abstract: The quality control of the wafer is becoming more and more important as the wafer becomes larger and the feature size shrinks. 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 possibility for inferential control is explored. This leads to a control architecture with multiple layers in a cascade structure. Next, 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. In order to reduce downtime, 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 the dynamic behavior of the tool. An RTP example is used to illustrate this approach and results clearly indicate that process trend is well predicted using the index-based time-series model.

1. INTRODUCTION

The continuing miniaturization of integrated circuit (IC) components and the increasing numbers of functions and performance of a 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 mm to 300 mm) and the feature size shrinks (from 350 nm to 90 nm). On the corporate level, improved yield is the only solution to remain competitiveness. Thus advanced equipment control and advanced process control (AEC/APC) have become a standard practice in modern semiconductor manufacturing. Edgar et al. (2000) give a comprehensive review in the processes and control issues, Qin et al. (2004) discuss the challenges in the IC industries, and Lewin et al. (2005) explore 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 unscheduled downtime. So, the AEC is synonymous with fault detection and classification (FDC) for the individual equipment. However, unlike chemical processes, an advanced IC fabrication process may include 300 steps (or process units), and 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). They are generally termed exponentially weighted moving average (EWMA) and double EWMA algorithms. 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 (Wang, 2004; Edgar, 2004; Wu et al., 2005) is held in the USA, Europe, and Asia each year with hundreds of attendees to each conference, and they have become the major events for APC division personnel from fabs worldwide. In fact, this is similar to the process control phenomena we 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., in terms of IC design), and (4) frequent tool

[†] to whom all correspondence should be addressed.
E-mail: ccyu@ntu.edu.tw
Fax: +886-2-2362-3040

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 and Knutson, 1992; Huang et al., 2000a,b,c; Chao et al., 2003a,b; Jung et al., 2003; Gunawan et al., 2004). 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, we mean 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 (Patel et al., 2000; Chen et al., 2005). Generally, some prior knowledge about the quality of the *incoming* wafers is available in semiconductor manufacturing processes. Thus, feedforward control or feed sequence arrangement can be devised to mitigate the incoming disturbance (Chen et al., 2005). A similar approach can be applied to the product-induced disturbance. The 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, we have seen almost weekly-based maintenance 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 and Guo, 2001).

2.2 Measurement

The product nature of IC makes the 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 we may not realize the yield 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 et al., 2000). The product yield is usually highly correlated to the WAT data. Generally, after each processing step, we have a quality measurement which is often denoted as the "metrology." Nano-scale nature makes the measurement (metrology) difficult and the measuring 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, we have 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 (Figure 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*. 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 in 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 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). Figure 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 several processing steps using the virtual WAT. The electrical properties of the product are controlled by adjusting metrology set points which subsequently affect the recipes in related tools. Figure 3 gives a detailed description of the control architecture for product quality control. It is clear that 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.

3. CONTROL OF RTP

Typically, wafer processing in a tool is described by a recipe which consists of on the order of ten 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. We will use rapid thermal processing (RTP) 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 increases 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. By varying the angle of the shield, an optimal shield configuration can be found to minimize the induced thermal stress (Young and 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 et al. (2000a,b,c).

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). 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 temperature programs shown in Fig. 5. 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 pn 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. We will use the spike annealing to illustrate thermal-stress-based temperature program generation with emphasis on the cooling curve.

Consider the RTP system shown in Fig. 4. The wafer thickness is assumed to be thin as compared to the radius of the wafer r_o, so we can regard this 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):

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

with boundary conditions given by

$$\frac{\partial T}{\partial r}=0\,,\quad \text{at r = 0} \qquad (2)$$

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

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 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:

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

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

$$\sigma_{r\theta} = 0 \quad (6)$$

where σ_{rr} and $\sigma_{\theta\theta}$ are the radial and tangential stress components, respectively. α and E denote 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 used which assumes that the wafer fails in shear when

$$S = \frac{\tau_{\max} \cdot F_s}{\tau_{yp}} > 1 \quad (7)$$

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:

$$\tau_{\max} = \frac{1}{2}\left|\sigma_{rr} - \sigma_{\theta\theta}\right| \quad (8)$$

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 (Hebb and Jensen, 1998) as:

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

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 $2.5\text{x}10^5$ Pa/s or its calculated value. If the result calculated from Equation (9) exceeds $3.1\text{x}10^8$ Pa, it is taken to be $3.1\text{x}10^8$ Pa which means that the wafer is at low temperature. From Equation (9) we know that the yield shear stress will be about 1.5 MPa when T= 1200°C at the beginning of the cooling process which is far less than 310 MPa at the room temperature T= 27°C. This simply indicates that, according to the failure criterion stated in Equation (7), a small temperature non-uniformity may induce material failure at high temperature. Since no analytical solution is available for the present problem, the numerical solutions are sought to the above governing equations. The calculation is carried out using a fully implicit finite difference method (Chao et al., 2003a).

Three scenarios are considered using the lamps radiative cooling condition: (1) fixed temperature-difference control 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 during the cooling process), (2) constant cooling-rate control scheme: The lamp's power decreases gradually at a constant rate of 10KW/m^2-s (by trial and error which ensures that the normalized maximum resolved stress is less than one during the cooling process), (3) maximum stress control scheme: The normalized maximum resolved stress is kept close to one until the lamp's power decreases to zero during the cooling process.

Chao et al. (2003a) show that the edge heat loss leads to large temperature gradient toward the wafer edge. Based on the maximum shear stress failure criterion, the results show that material failure always occurs at the edge of the wafer at the beginning of cooling processes. Furthermore, the maximum stress control scheme is shown to be more efficient because it can significantly reduce the required cooling time and thermal budgets. Thus, the conventional constant cooling-rate control scheme or linear temperature ramp-down scheme is not appropriate for the rapid thermal processor.

Fig. 6 shows, for the radiative-only cooling process, the tangential stress at the wafer edge is positive due to thermal shrinkage induced by the edge effect. On the other hand, the compressive tangential stress prevails at the central region of wafer. Since the tangential stress at the central region is far less than the tangential stress at the wafer edge. The wafer failure is dominated by the edge effect in the wafer and yield stress in shear. For the maximum stress control scheme, the lamp's power decreases dramatically during the cooling process. After five seconds have elapsed, the lamp's power for the fixed temperature-difference control scheme decreases gradually with a rate even smaller than the constant cooling-rate control scheme. The required cooling time for the maximum stress control scheme is only 18 sec from 1200°C to 600°C, compared to 30 sec 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. 7. This provides an attractive alternative for temperature program generation.

3.3 Control Structure Design

The state-of-the-art RTP typically consists of 7 lamp-heating zones with 7 temperature measurements, in addition to computed emissivity. Here we use a simple RTP model (Huang et al., 2000a) to illustrate the essential steps in the control structure design. This is an RTP system with 3 lamp-heating zones for a 200mm 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°C-1050°C with the duration of approximately 2 seconds.

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 power 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). Therefore, it becomes clear that the best temperature uniformity that can be achieved is the temperature profile minimizing the squares of temperature differences which is termed the *desired* temperature profile. Furthermore, the easiest way to maintain this profile is to keep the temperatures already at (or close to) set point (e.g., Fig. 5) under control. This can be interpreted as retaining the shape of the temperature profile by holding several key positions at the set point. If we have more zero-crossing temperatures than manipulated inputs, The next step is to check system interaction and inherent robustness using the structured singular value (SSV). 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 of 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. Because almost half of the batch cycle involves ramp-type setpoint trajectory, 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 1050ºC, the model gives the following process transfer function matrix: Note that the sampling rate (0.01 s) is so fast that, a continuous-time model is used here.

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

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.

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

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

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

We term this type of controller 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.

Figure 8 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. Moreover, the two important criteria, peak temperature and duration time over 1000ºC, are completely missed, even with PI^{2}D control. Table 1 summarized the spread of the peak temperature and duration time.

If the peak temperature tracking and duration is the design criteria, the triangular temperature problem in Fig. 5A cannot be achieved with a realizable controller. Thus, a smooth temperature program is used instead as shown in Fig. 5B. The tabulated results in Table 1 also confirm this and the peak temperature spread is reduced to 6.9ºC as compared to 11.8ºC for triangular temperature program. Figure 9 shows the peak-temperature spread across the radial positions is reduced to 6.9ºC using the smooth temperature program for the RTP with 3 heating zones. The trend remains for wafer with different peak temperatures. The results presented here clearly indicate that the advanced control methodology can certainly be applied to semiconductor manufacturing at the tool level.

4. PROCESS MONITORING

Process monitoring and analysis is important in semiconductor manufacturing. Correct trend monitoring can be used to determine appropriate timing for preventive maintenance. In this work, instead of incorporating large number of trajectory data with variable batch time and possibly "missing" data for some process variables using multivariable statistic technique (e.g. MPCA), 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 time where the 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 the 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 an index for the 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. An IC 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 step 10 is identified as KST. Then, three important process variables are selected as possible 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. 10(a), $A_{\max}$ for some batches are abnormally greater than the average value. Since different products are usually processed with the same tool, Amax with particularly high values may result from different products. Thus, one product index, as shown in Fig. 10(b), is considered for the modification of $A_{\max}$ values. The result of modified $A_{\max}$, designated as $A'_{\max}$, is shown in Fig. 10(c) where all $A'_{\max}$ values follow the data trend. Consequently, an autoregressive moving average (ARMA) model of the following is built for $A'_{\max}$ based on measurements from 500 wafers.

$$\begin{aligned}&\left(1-1.744\,q^{-1}+0.776\,q^{-2}\right)A'_{\max}(t)\\&\quad=\left(1-1.346\,q^{-1}+0.476\,q^{-2}\right)e(t)\end{aligned} \tag{14}$$

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 the time series $A_{\max}(t)$ exhibits nonstationary behavior. For this reason, an autoregressive integrated moving average (ARIMA) model is then built to describe this behavior.

$$\begin{aligned}&\left(1+0.942\,q^{-1}\right)\nabla A'_{\max}(t)\\&\quad=\left(1+0.452\,q^{-1}-0.553\,q^{-2}\right)e(t)\end{aligned} \tag{15}$$

where $\nabla=\left(1-q^{-1}\right)$ 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. 11 where two abrupt changes are observed due to scheduled tool 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 $A'_{\max}$ more and more, while the forecast of ARIMA model keeps following the actual process. 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.

$$\mathrm{KSI}=\left|\mathrm{Residual}_{\mathrm{ARMA}}\text{-}\mathrm{Residual}_{\mathrm{ARIMA}}\right| \tag{16}$$

The computed KSI is shown in Fig. 12. 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 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

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, 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. In order to prolong the time between maintenance and to reduce unscheduled downtime, process trend monitoring of a tool is essential. Instead of using entire batch data, 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.

ACKNOWLEDGMENT

This work was supported in part by the National Science Council of Taiwan. We would like to thank Sunny Wu, B. H. Chen, J. S. Lin, Henry Lo, Jean Wang, C. H. Yu, and M. S. Liang of TSMC for continuing support and discussion. Long-term collaboration with Walters Shen of AMAT is also gratefully acknowledged. We also thank anonymous reviewers for thoughtful comments.

REFERENCES

Chao, C. K.; Hung, S. Y.; Yu, C. C. (2003a). Thermal Stress Analysis for Rapid Thermal Processor, *IEEE Trans. Semi. Manuf.* 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 J. Manufac. Sci. Eng.*, 125, 504.

Chen, A.; Guo, R. S. (2001). Age-based double EWMA controller and its application to CMP processes, *IEEE Trans. Semi. Manuf.*, 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, *Ind. Eng. Chem. Research*, 44, 4676.

Edgar, T. F. (2004). Multi-product Run-to-Run Control for High-Mix Fabs, *AEC/APC Symposium Asia,* HsinChu, Dec.

Edgar, T. F.; Butler, S. W.; Campbell, W. J.; Pfeiffer, C.; Bode, C.; Hwang, S. B.; Balakrishnan, K. S.; Hahn, J. (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 Trans. Semi. Manuf.*, 13, 344.

Gunawan, R.; Jung, M. Y. L.; Seebauer, E. G.; Braatz, R. D. (2004) Optimal Control of Rapid Thermal Annealing in a Semiconductor Process, *J. 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 Trans Semi. Manuf.*, 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, *J. Chin. Inst. Chem. Eng.*, 31, 585.

Huang, I.; Liu, H. H.; Yu, C. C. (2000c). Design for Control: Temperature Uniformity in Rapid Thermal Processor, *Korean J. Chem. Eng.*, 17, 111.

Jung, M. Y.; Gunawan, R.; Braatz, R. D.; Seebauer, E. G. (2003). Ramp-Rate Effects on Transient Enhanced Diffusion and Dopant Activation, *J. Electrochem. Soc.*, 150, G838.

Lewin, D. R.; Lachman-Shalem, S.; Grosman, B. (2005). More Process System Engineering (PSE) Applications in IC Manufacturing, I*FAC World Congress*, Prague, July.

Morari, M.; Zafiriou, E. (1989). *Robust Process Control*. Prentice-Hall, Englewood Cliff.

Nowinski, J. L. (1978). *Theory of Thermoelasticity with Application.* Sijthoff & Noordhoff.

Patel, N. S.; Miller, G. A.; Guinn, C.; Jenkins, S. T. (2000). Device dependent control of chemical-mechanical polishing of dielectric films, IEEE Trans. Semi. Manuf., 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.

Wang, T. (2004). Advanced Process Control Road Map and Challenges, *AEC/APC Symposium Asia*, HsinChu, Dec.

Wu, S.; Chen, P. H.; Lin, J. S.; Ko, F.; Lo, H.; Wang, J.; Yu, C. H.; Liang, M. S. (2005). Real-Time Device Performance Prediction for 90nm and Beyond, *AEC/APC Symposium USA*, Palm Spring, CA, Sept.

Young, G. L.; McDonald, K. A. (1990). Effect of Radiation Shield Angle on Temperature and Stress Profiles During Rapid Thermal Annealing, *IEEE Trans Semi. Manuf.,* 3, 176.

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

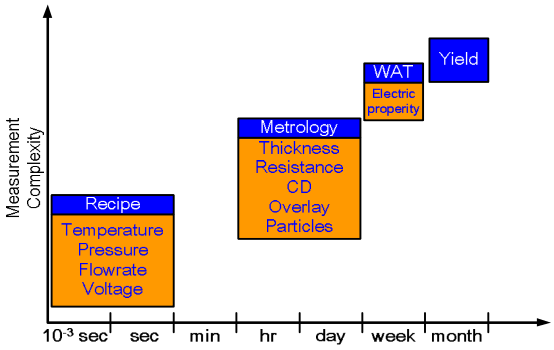

Figure 1. Measurement complexity and frequency

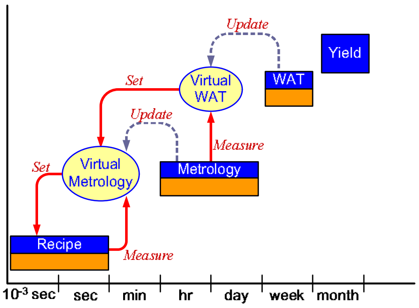

Figure 2. Structure of control action

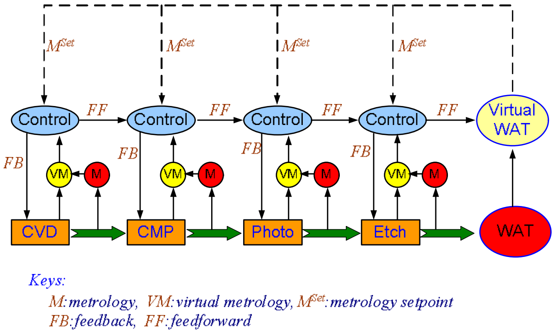

Figure 3. Fab-wide control schema

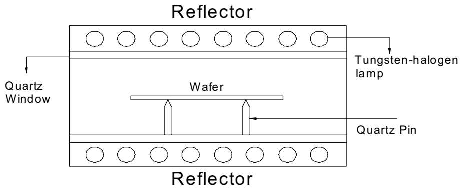

Figure 4. The physical model of RTP

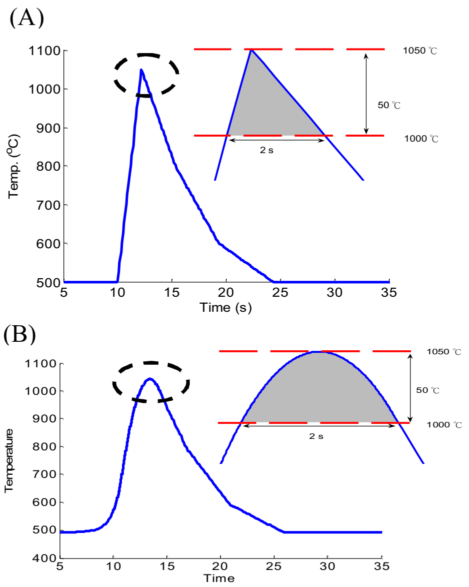

Figure 5. (A) triangular-like temperature program (B)smooth temperature program.

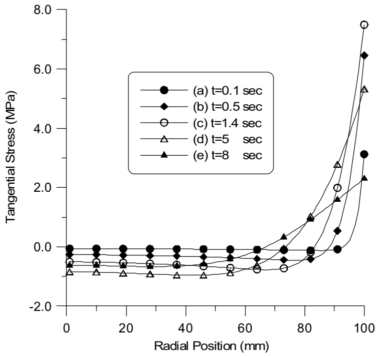

Figure 6. The tangential stress distribution on wafer for the room temperature cooling.

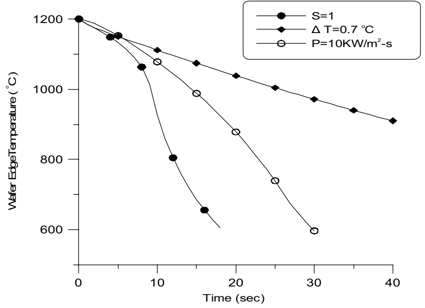

Figure 7. The temperature variation at wafer edge under three different control schemes

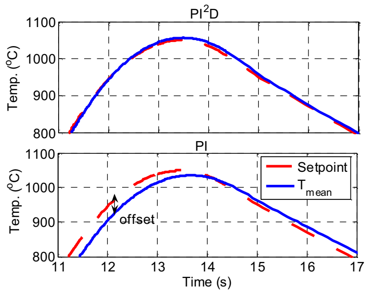

Figure 8. control results of PI ans PI^{2}D for smooth temperature program.

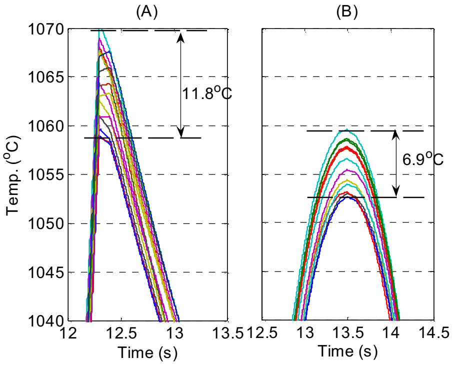

Figure 9. Spread of the peak temperature for (A) triangular-like (B) smooth temperature program.

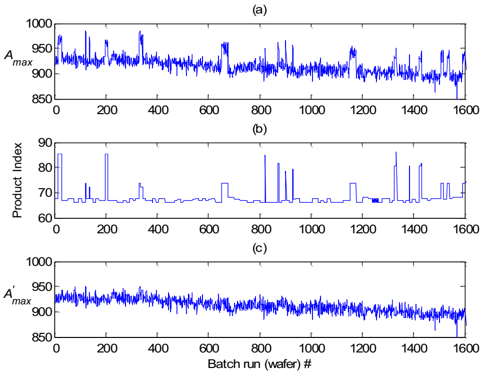

Figure 10. KSV and product index (a) KSV (b) product index (c) modified KSV

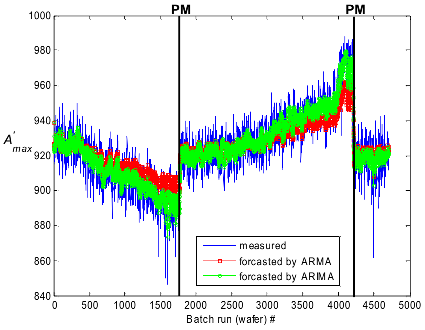

Fig. 11. Comparison of ARMA and ARIMA prediction as compared to the true measurement.

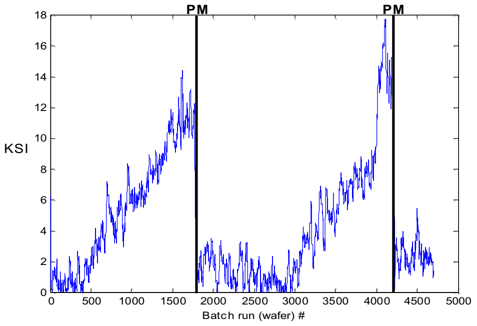

Figure 12. KSI for process trend monitoring