DEC-07-1995 16:23 CRC PRESS-EDP DEPT. 407+994+3625 P.02/12

73

Control for Advanced Semiconductor Device Manufacturing: A Case History

T. Kailath, C. Schaper, Y. Cho, P. Gyugyi, S. Norman, P. Park, S. Boyd, G. Franklin, and K. Saraswat
Department of Electrical Engineering, Stanford University, Stanford, CA

M. Moslehi and C. Davis
Semiconductor Process and Design Center, Texas Instruments, Dallas, TX

73.1 Introduction 471
73.2 Modeling and Simulation 474
73.3 Performance Analysis 475
73.4 Models for Control 476
73.5 Control Design 480
73.6 Proof-of-Concept Testing 481
73.7 Technology Transfer to Industry 483
73.8 Conclusions 484
References 487

73.1 Introduction

Capital [1] costs for new integrated circuit (IC) fabrication lines are growing even more rapidly than had been expected even quite recently. Figure 73.1 was prepared in 1992, but a new Mitsubishi factory in Shoji, Japan, is reported to have cost $3 billion. Few companies can afford investments on this scale (and those that can perhaps prefer it that way). Moreover these factories are inflexible. New equipment and new standards, which account for roughly 3/4 of the total cost, are needed each time the device feature size is reduced, which has been happening about every 3 years. It takes about six years to bring a new technology on line. The very high development costs, the high operational costs (e.g., equipment down time is extremely expensive so maintenance is done on a regular schedule, whether it is needed or not), and the intense price competition compel a focus on high-volume low cost commodity lines, especially memories. Low volume, high product mix ASIC (application-specific integrated circuit) production does not fit well within the current manufacturing scenario.

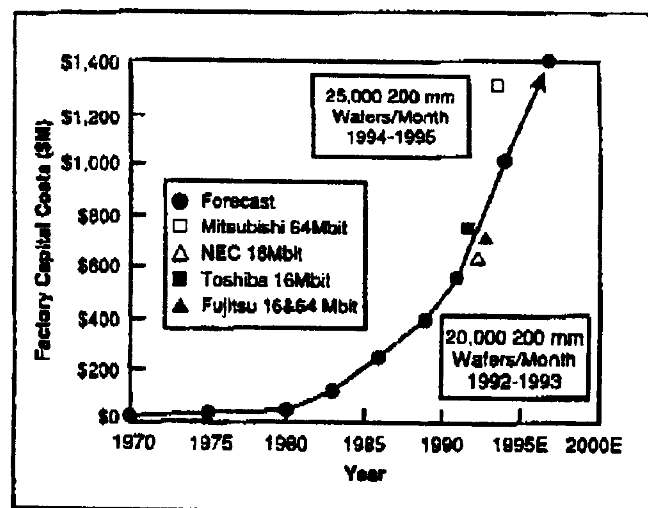


Figure 73.1 Capital cost for a new IC factory. (Source: *Texas Instruments Technical Journal*, 9(5), 8, 1992.)

In 1989, the Advanced Projects Research Agency(ARPA), Air Force Office of Scientific Research(AFOSR), and Texas Instruments (TI) joined in a $150 million cost-shared program called MMST (Microelectronics Manufacturing Science and Technology) to "establish and demonstrate (new) concepts for semiconductor device manufacture which will permit flexible, cost-effective manufacturing of application-specific logic integrated circuits in relatively low volume ... during the mid 1990s and beyond".

The approach taken by MMST was to seek fast cycle time by performing all single-wafer processing using highly instrumented flexible equipment with advanced process controls. The goal of the equipment design and operation was to quickly adapt the equipment trajectories to a wide variety of processing specifications and to quickly reduce the effects of manufacturing disturbances associated with small lot sizes (e.g., 1, 5 or 24 wafers)

[1]This research was supported by the Advanced Research Projects Agency of the Department of Defense, under Contract F49620-93-1-0085 monitored by the Air Force Office of Scientific Research.

0-8493-8570-9/96/$0.00+$.50
© 1996 by CRC Press, Inc.

DEC-07-1995 16:23 CRC PRESS-EDP DEPT. 407+994+3625 P.03/12

without the need for pilot wafers. Many other novel features were associated with MMST including a factory wide CIM (computer integrated manufacturing) computer system. The immediate target was a 1000-wafer demonstration (including demonstration of "bullet wafers" with three-day cycle times) of an all single-wafer factory by May 1993.

In order to achieve the MMST objectives, a flexible manufacturing tool was needed for the thermal processing steps associated with IC manufacturing. For a typical CMOS process flow, more than 15 different thermal processing steps are used, including chemical vapor deposition (CVD), annealing, and oxidation. The MMST program decided to investigate the use of Rapid Thermal Processing (RTP) tools to achieve these objectives.

TI awarded Professor K. Saraswat of Stanford's Center for Integrated Systems (CIS) a subcontract to study various aspects of RTP. About a year later, a group of us at Stanford's Information Systems Laboratory got involved in this project. Manufacturing was much in the news at that time. Professor L. Auslander, newly arrived at ARPA's Material Science Office, soon came to feel that the ideas and techniques of control, optimization, and signal processing needed to be more widely used in materials manufacturing and processing. He suggested that we explore these possibilities, and after some investigation, we decided to work with CIS on the problems of RTP.

RTP had been in the air for more than a decade, but for various reasons, its study was still in a research laboratory phase. Though there were several small companies making equipment for RTP, the technology still suffered from various limitations. One of these was an inability to achieve adequate temperature uniformity across the wafer during the rapid heating (e.g., 20°C to 1100°C in 20 seconds), hold (e.g., at 1100°C for 1–5 minutes), and rapid cooling phases.

This chapter is a case history of how we successfully tackled this problem, using the particular "systems-way-of-thinking" very familiar to control engineers, but seemingly not known or used in semiconductor manufacturing. In a little over two years, we started with simple idealized mathematical models and ended with deployment of a control system during the May, 1993, MMST demonstration. The system was applied to eight different RTP machines conducting thirteen different thermal operations, over a temperature range of 450°C to 1100°C and pressures ranging from 10^{-3} to 1 atmosphere.

Our first step was to analyze the performance of available commercial equipment. Generally, a bank of linear lamps was used to heat the wafer (see Figure 73.2).

The conventional wisdom was that a uniform energy flux to the wafer was needed to achieve uniform wafer temperature distribution. However, experimentally it had been seen that this still resulted in substantial temperature nonuniformities, which led to crystal slip and misprocessing. To improve performance, various heuristic strategies were used by the equipment manufacturers, e.g., modification of the reactor through the addition of guard rings near the wafer edge to reflect more energy to the edge, modification of the lamp design by using multiple lamps with a fixed power ratio, and various types of reflector geometries. However, these modifications turned out to be satisfactory only for a narrow range of conditions.

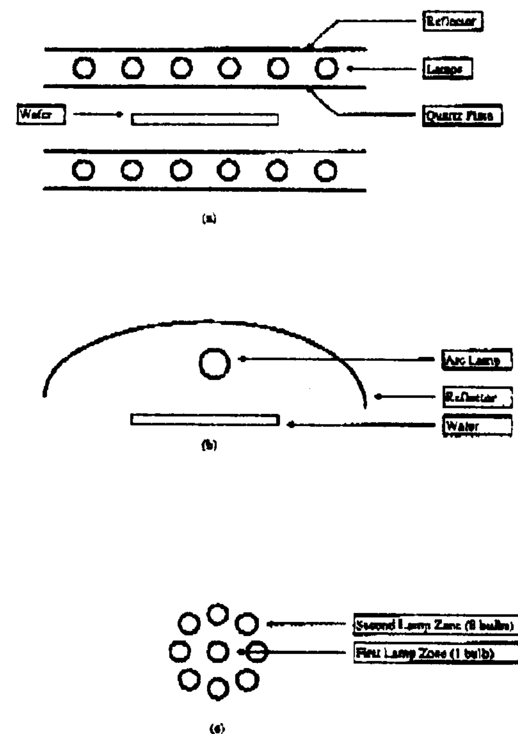


Figure 73.2 RTP lamp configurations: (a) bank of linear lamps, (b) single arc lamp, (c) two-zone lamp array.

The systems methodology suggests methods attempting to determine the performance limitations of RTP systems. To do this, we proceeded to develop a simple mathematical model, based on energy transfer relations that had been described in the literature. Computer simulations with this model indicated that conventional approaches trying to achieve uniform flux across the wafer would never work; there was always going to be a large temperature roll-off at the wafer edge (Figure 73.3). To improve performance, we decided to study the case where circularly symmetric rings of lamps were used to heat the wafer. With this configuration, two cases were considered: (1) a single power supply in a fixed power ratio, a strategy being used in the field and (2) independently controllable multiple power supplies (one for each ring of lamps). Both steady-state and dynamic studies indicated that it was necessary to use the (second) multivariable configuration to achieve wafer temperature uniformity within specifications. These modeling and analysis results are described in Sections 73.2 and 73.3, respectively.

The simulation results were presented to Texas Instruments, which had developed prototype RTP equipment for the MMST program with two concentric lamp zones, but operated in a scalar control mode using a fixed ratio between the two lamp zones. At our request, Texas Instruments modified the two zone lamp by adding a third zone and providing separate power supplies for each zone, allowing for multivariable control. The process engineers in the Center for Integrated Systems (CIS) at Stanford then evaluated the potential of multivariable control by their traditional so called "hand-tuning" methodology, which con-

DEC-07-1995 16:24 CRC PRESS-EDP DEPT. 407+994+3625 P.04/12

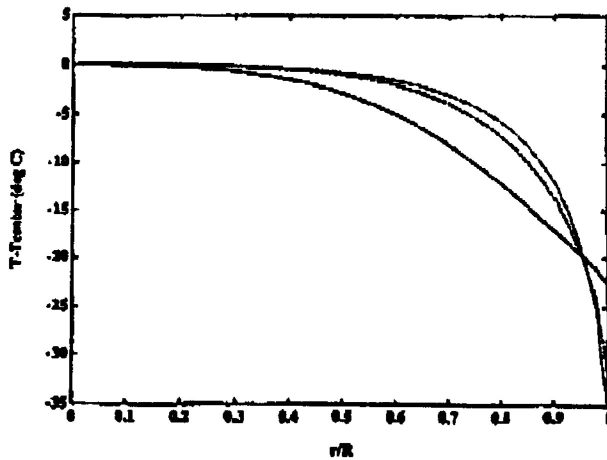


Figure 73.3 Nonuniformity in temperature induced by uniform energy flux impinging on the wafer top surface (center temperatures - solid line: 600°C; dashed line: 1000°C; dotted line: 1150°C.). R is the radius of the wafer, r is the radial distance from the center of the wafer.

sists of having experienced operators determining the settings of the three lamp powers by manual iterative adjustment based on the results of test wafers. Good results were achieved (see Figure 73.4), but it took 7–8 hours and a large number of wafers before the procedure converged. Of course, it had to be repeated the next day because of unavoidable changes in the ambient conditions or operating conditions. Clearly, an "automatic" control strategy was required.

However, the physics-based equations used to simulate the RTP were much too detailed and contained far too many uncertain parameters for control design. The two main characteristics of the simulation model were (1) the relationship between the heating zones and the wafer temperature distribution and (2) the nonlinearities (T^4) of radiant heat transfer. Two approaches were used to obtain a reduced-order model. The first used the physical relations as a basis in deriving a lower-order approximate form. The resulting model captured the important aspects of the interactions and the nonlinearities, but had a simpler structure and fewer unknown parameters. The second approach viewed the RTP system as a black box. A novel model identification procedure was developed and applied to obtain a state-space model of the RTP system. In addition to identifying the dynamics of the process, these models were also studied to assess potential difficulties in performance and control design. For example, the models demonstrated that the system gain and time constants changed by a factor of 10 over the temperature range of interest. Also, the models were used to improve the condition number of the equipment via a change in reflector design. The development of control models is described in Section 73.4.

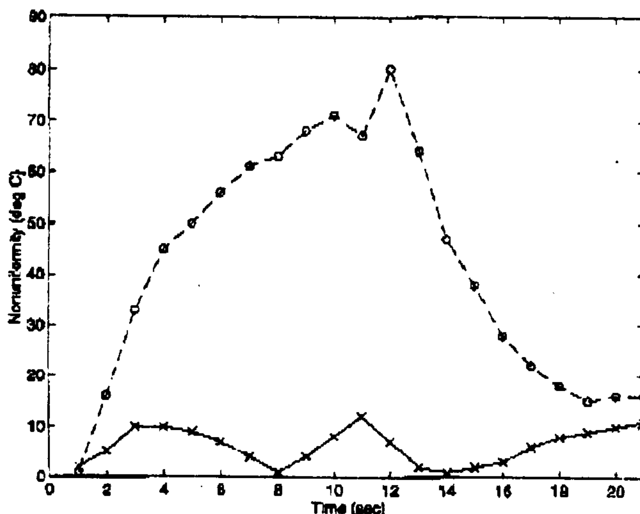
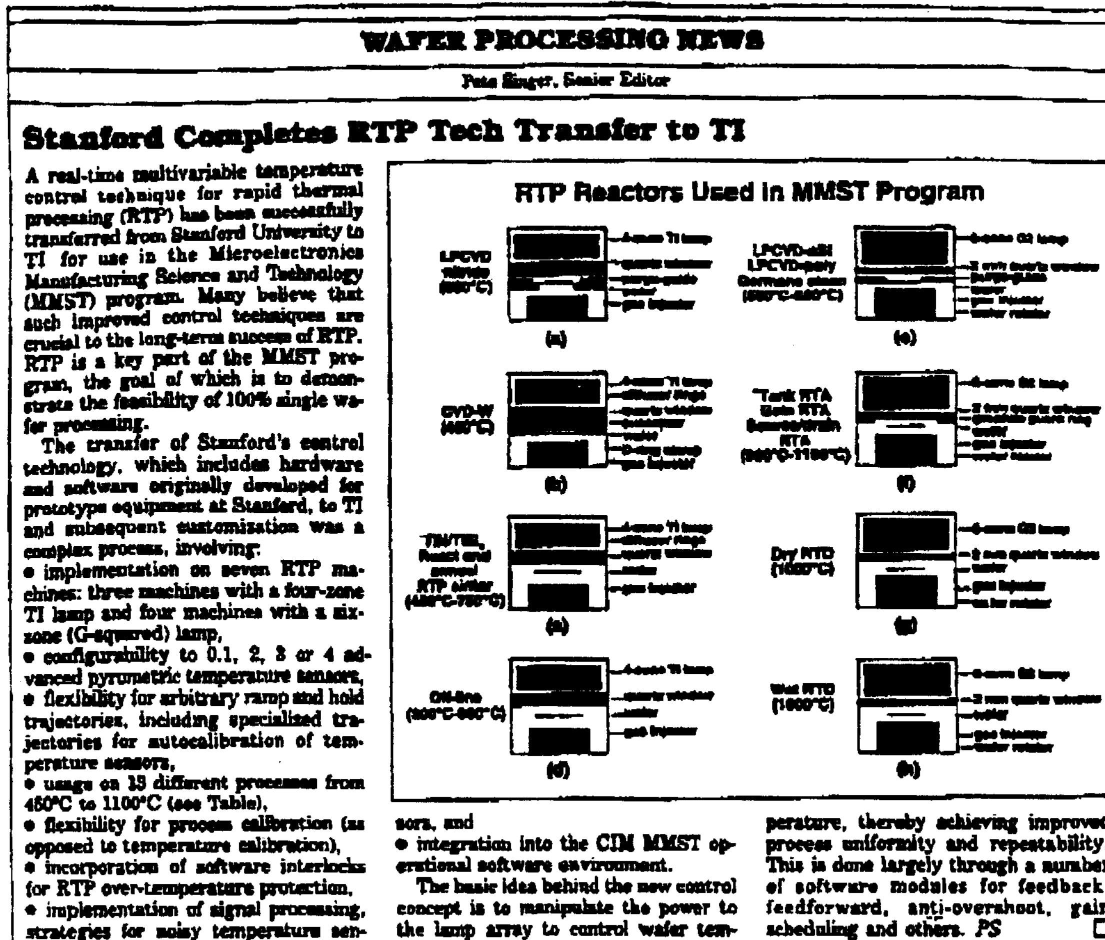


Figure 73.4 Temperature nonuniformity when the powers to the lamp were manually adjusted ("hand-tuning"). These nonuniformities correspond to a ramp and hold from nearly room temperature to 600°C at roughly 40°C/s. The upper curve (-o-) corresponds to scalar control (fixed power ratio to lamps). The lower curve (x-x) corresponds to multivariable control.

Using these models, a variety of control strategies was evaluated. The fundamental strategy was to use feedforward in combination with feedback control. Feedforward control was used to get close to the desired trajectory and feedback control was used to compensate for inevitable tracking errors. A feedback controller based on the Internal Model Control (IMC) design procedures was developed using the low-order physics-based model. An LQG feedback controller was developed using the black-box model. Gain scheduling was used to compensate for the nonlinearities. Optimization procedures were used to design the feedforward controller. Controller design is described in Section 73.5.

Our next step was to test the controller experimentally on the Stanford RTP system. After using step response and PRBS (Pseudo Random Binary Sequence) data to identify models of the process, the controllers were used to ramp up the wafer temperature from 20°C to 900°C at approximately 45°C/s, followed by a hold for 5 minutes at 900°C. For these experiments, the wafer temperature distribution was sensed by three thermocouples bonded to the wafer. The temperature nonuniformity present during the ramp was less than ±5°C from 400°C to the processing temperature and better than ±0.5°C on average during the hold. These proof-of-concept experiments are described in Section 73.6.

These results were presented to Texas Instruments, who were preparing their RTP systems for a 1000 wafer demonstration of the MMST concept. After upper level management review, it was decided that the Stanford temperature control system would be integrated within their RTP equipment. The technology transfer involved installing and testing the controller on eight different RTP machines conducting thirteen different thermal operations used in two full-flow 0.35 μm CMOS process technologies (see Figure 73.5 taken from an article appearing in a semiconductor manufacturing trade journal). More discussion concerning the



WAFER PROCESSING NEWS

Pete Singer, Senior Editor

Stanford Completes RTP Tech Transfer to TI

A real-time multivariable temperature control technique for rapid thermal processing (RTP) has been successfully transferred from Stanford University to TI for use in the Microelectronics Manufacturing Science and Technology (MMST) program. Many believe that such improved control techniques are crucial to the long-term success of RTP. RTP is a key part of the MMST program, the goal of which is to demonstrate the feasibility of 100% single wafer processing.

The transfer of Stanford's control technology, which includes hardware and software originally developed for prototype equipment at Stanford, to TI and subsequent customization was a complex process, involving:

- implementation on seven RTP machines: three machines with a four-zone TI lamp and four machines with a six-zone (G-squared) lamp,
- configurability to 0.1, 2, 3 or 4 advanced pyrometric temperature sensors,
- flexibility for arbitrary ramp and hold trajectories, including specialized trajectories for autocalibration of temperature sensors,
- usage on 13 different processes from 450°C to 1100°C (see Table),
- flexibility for process calibration (as opposed to temperature calibration),
- incorporation of software interlocks for RTP over-temperature protection,
- implementation of signal processing, strategies for noisy temperature sensors, and
- integration into the CIM MMST operational software environment.

The basic idea behind the new control concept is to manipulate the power to the lamp array to control wafer temperature, thereby achieving improved process uniformity and repeatability. This is done largely through a number of software modules for feedback, feedforward, anti-overshoot, gain scheduling and others. PS

Figure 73.5 Description of technology transfer in *Semiconductor International*, 16(7), 58, 1993.

technology transfer and results of the MMST demonstration is given in Section 73.7. Finally, some overview remarks are offered in Section 73.8.

73.2 Modeling and Simulation

Three alternative lamp configurations for rapidly heating a semiconductor wafer are shown in Figure 73.2. In Figure 73.2(a), linear lamps are arranged above and below the wafer. A single arc lamp is shown in Figure 73.2(b). Concentric rings of single bulbs are presented in Figure 73.2(c). These designs can be modified with guard rings around the wafer edge, specially designed reflectors, and diffusers placed on the quartz window. These additions allowed fine-tuning of the energy flux profile to the wafer to improve temperature uniformity.

To analyze the performance of these and related equipment designs, a simulator of the heat transfer effects was developed starting from physical relations for RTP available in the literature [1], [2]. The model was derived from a set of PDE's describing the radiative, conductive and convective energy transport effects. The basic expression is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(kr\frac{\partial T}{\partial r}\right)+\frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k\frac{\partial T}{\partial \theta}\right)+\frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right)=\rho C_p\frac{\partial T}{\partial t} \tag{73.1}$$

where T is temperature, k is thermal conductivity, ρ is density, and C_p is specific heat. Both k and C_p are temperature dependent. The boundary conditions are given by

$$\begin{aligned} k\frac{\partial T}{\partial r} &= q_{edge}(\theta, z), r = R, \\ k\frac{\partial T}{\partial z} &= q_{bottom}(r, \theta), z = 0, \text{ and} \\ k\frac{\partial T}{\partial z} &= q_{top}(r, \theta), z = Z, \end{aligned}$$

where q_{edge}, q_{bottom}, and q_{top} are heat flow per unit area into the wafer edge, bottom, and top, respectively, via radiative and convective heat transfer mechanisms, Z is the thickness of the wafer, and R is the radius of the wafer. These terms coupled the effects of the lamp heating zones to the wafer.

Approximations were made to the general energy balance assuming axisymmetry and neglecting axial temperature gradients. The heating effects in RTP were developed by discretizing the wafer into concentric annular elements. Within each annular

DEC-07-1995 16:25 CRC PRESS-EDP DEPT. 407+994+3625 P.06/12

wafer element, the temperature was assumed uniform [2]. The resulting model was given by a set of nonlinear vector differential equations:

$$\begin{aligned} C\dot{T} &= \mathbf{K}^{rad}T^4 + \mathbf{K}^{cond}T + \mathbf{K}^{conv}(T - T_{gas}) \\ &\quad + \mathbf{F}P + q^{wall} + q^{dist} \end{aligned} \quad (73.2)$$

where

$$\begin{aligned} T &= [T_1\ T_2\ \ldots\ T_N]^T \\ T^4 &= \left[T_1^4\ T_2^4\ \ldots\ T_N^4\right]^T \\ P &= [P_1\ P_2\ \ldots\ P_M]^T \end{aligned}$$

where N denotes the number of wafer elements and M denotes the number of radiant heating zones; $\mathbf{K}^{rad}$ is a full matrix describing the radiation emission characteristics of the wafer, $\mathbf{K}^{cond}$ is a tridiagonal matrix describing the conductive heat transfer effects across the wafer, $\mathbf{K}^{conv}$ is a diagonal matrix describing the convective heat transfer effects from the wafer to the surrounding gas, $\mathbf{F}$ is a full matrix quantifying the fraction of energy leaving each lamp zone that radiates onto the wafer surface, q^{dist} is a vector of disturbances, q^{wall} is a vector of energy flux leaving the chamber walls and radiating onto the wafer surface, and $\mathbf{C}$ is a diagonal matrix relating the heat flux to temperature transients. More details can be found in [2] and [3].

73.3 Performance Analysis

We first used the model to analyze the case of uniform energy flux impinging on the wafer surface. In Figure 73.3, the temperature profile induced by a uniform input energy flux is shown for the cases where the center portion of the wafer was specified to be at either 600°C, 1000°C, or 1150°C. A roll-off in temperature is seen in the plots for all cases because the edge of the wafer required a different amount of energy flux than the interior due to differences in surface area. Conduction effects within the wafer helped to smooth the temperature profile. These results qualitatively agreed with those reported in the literature where, for example, sliplines at the wafer edge were seen because of the large temperature gradients induced by the uniform energy flux conditions.

We then analyzed the multiple concentric lamp zone arrangement of Figure 73.2(c) to assess the capability of achieving uniform temperature distribution during steady-state and transients. We considered each of four lamp zones to be manipulated independently. The optimal lamp powers were determined to minimize the peak temperature difference across the wafer at a steady-state condition,

$$\max_{0 \le r \le R} \left|T^{ss}(r, P) - T^{set}\right| \quad (73.3)$$

where T^{set} is the desired wafer temperature and $T^{ss}(r, P)$ is the steady-state temperature at radius r with the constant lamp power vector P, subject to the constraint that each entry P_j of P satisfies $0 \le P_j \le P_j^{max}$. Using the finite difference model, the objective function of Equation 73.3 was approximated as

$$\max_i \left|T_i^{ss}(P) - T^{set}\right| = \left|\left|T^{ss}(P) - T^{set}\right|\right|_\infty \quad (73.4)$$

where $T_i^{ss}(P)$ is the steady-state temperature of element i with constant lamp power vector P and T^{set} is a vector with all entries equal to T^{set}. A two-step numerical optimization procedure was then employed in which two minimax error problems were solved to determine the set of lamp powers that minimize Equation 73.3 [4] and [2]. In Figure 73.6, the temperature deviation about the set points of 650°C, 1000°C, and 1150°C is shown. The deviation is less than ±1°C, much better than for the case of uniform energy flux.

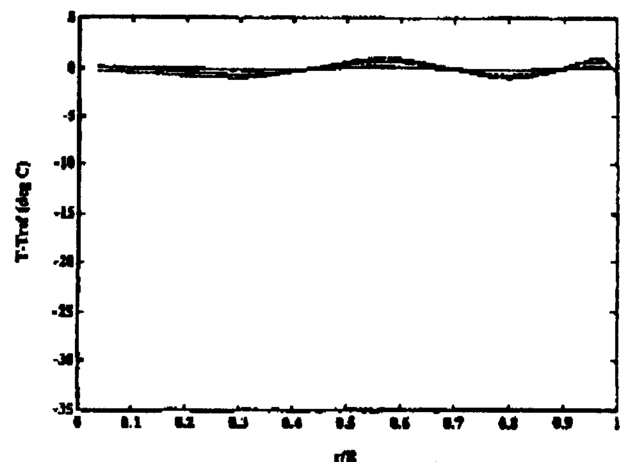


Figure 73.6 Optimal temperature profiles using a multizone RTP system (center temperatures - solid line: 600°C; dashed line: 1000°C; dotted line: 1150°C).

In addition, an analysis of the transient performance was conducted because a significant fraction of the processing and the potential for crystal slip occurs during the ramps made to wafer temperature. We compared a multivariable lamp control strategy and a scalar lamp control strategy. Industry, at that time, employed a scalar control strategy. For the scalar case, the lamps were held in a fixed ratio of power while total power was allowed to vary. We selected the optimization criterion of minimizing

$$\max_{t_0 \le t \le t_f} \left|\left|T(t) - T^{ref}(t)\right|\right|_\infty \quad (73.5)$$

which denotes the largest temperature error from the specified trajectory $T^{ref}(t)$ at any point on the wafer at any time between an initial time T_0 and a final time t_f. The reference temperature trajectory was selected as a ramp from 600°C to 1150°C in 5 seconds. The optimization was carried out with respect to the power to the four lamp zones, in the case of the multilamp configuration, or to the total power for a fixed ratio that was optimal only at a 1000°C steady-state condition. The temperature at the center of the wafer matched the desired temperature trajectory almost exactly for both the multivariable and scalar control

DEC-07-1995 16:26 CRC PRESS-EDP DEPT. 407+994+3625 P.07/12

cases. However, the peak temperature difference across the wafer was much less for the multivariable case compared to the scalar (fixed-ratio) case as shown in Figure 73.7.

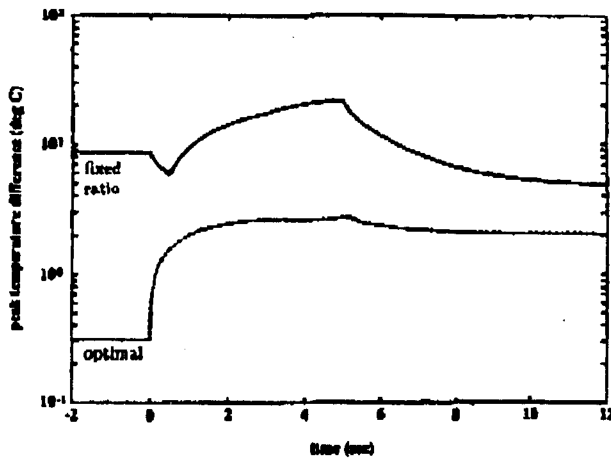


Figure 73.7 Peak temperature nonuniformity during ramp.

For the case of the fixed-ratio lamps, the peak temperature difference was more than 20°C during the transient and the multivariable case resulted in a temperature deviation of about 2°C. The simulator suggested that this nonuniformity in temperature for the fixed-ratio case would result in crystal slip as shown in Figure 73.8 which shows the normalized maximum resolved stress

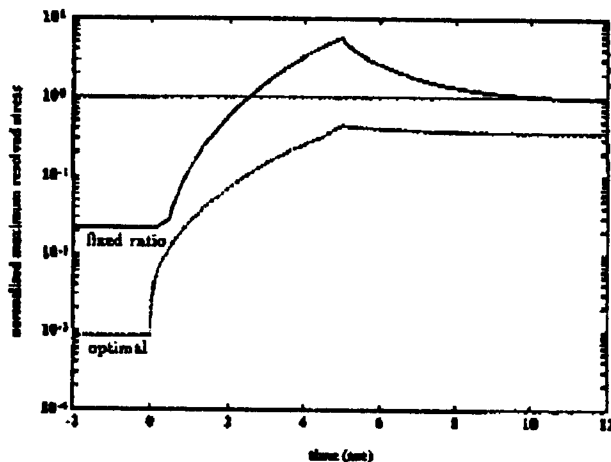


Figure 73.8 Normalized maximum resolved stress during ramp.

(based on simulation) as a function of time. No slip was present in the multivariable case. This analysis of the transient performance concluded that RTP systems configured with multiple independently controllable lamps can substantially outperform any existing scalar RTP system: for the same temperature schedule, much smaller stress and temperature variation across the wafer was achieved; and for the same specifications for stress and temperature variation across the wafer, much faster rise times can be achieved [2].

At the time of these simulations, prototype RTP equipment was being developed at Texas Instruments for implementation in the MMST program. TI had developed an RTP system with two concentric lamp zones. Their system at that time was operated in a scalar control mode with a fixed ratio between the two lamp zones. Upon presenting the above results, the two zone lamp was modified by adding a third zone and providing separate power supplies for each zone. This configuration allowed multivariable control. A resulting three-zone RTP lamp was then donated by TI to Stanford University. The chronology of this technology transfer is shown in Figure 73.9.

CHRONOLOGY OF TECHNOLOGY TRANSFER FROM STANFORD TO TEXAS INSTRUMENTS

(1/90 - 8/90)	Modeling of heat transfer for RTP Optimization and simulation of performance limits Comparison of multizone lamp configurations
(9/90 - 5/91)	Development and simulation of controllers
(6/91 - 3/92)	Experimental demonstration on Stanford RTM
(4/92 - 12/92)	Transfer and customization on 8 RTP's, 13 different processes at TI
(1/93 - 5/93)	Usage for 1,000 wafer MMST marathon demo

Figure 73.9 Chronology of the technology transfer to Texas Instruments.

A schematic of the Stanford RTP system and a picture of the three-zone arrangement are shown in Figures 73.10 and 73.11, respectively.

"Hand-tuning" procedures were used to evaluate the performance of the RTP equipment at Stanford quickly. In this approach, the power settings to the lamp were manually manipulated in real-time to achieve a desirable temperature response. In Figure 73.4, open-loop, hand-tuned results are shown for scalar control (i.e., fixed power ratio) and multivariable control as well as the error during the transient. Clearly, this comparison demonstrated that multivariable control was preferred to the scalar control method [5]. However, the hand-tuning approach was a trial and error procedure that was time-consuming and resulted in sub-optimal performance. An automated real-time control strategy is described in the following sections.

73.4 Models for Control

Two approaches were evaluated to develop a model for control design. In the first approach, the nonlinear physical model presented earlier was used to produce a reduced-order version. An energy balance equation on the ith annular element can be ex-

DEC-07-1995 16:27 CRC PRESS-EDP DEPT. 407+554+3625 P.08/11

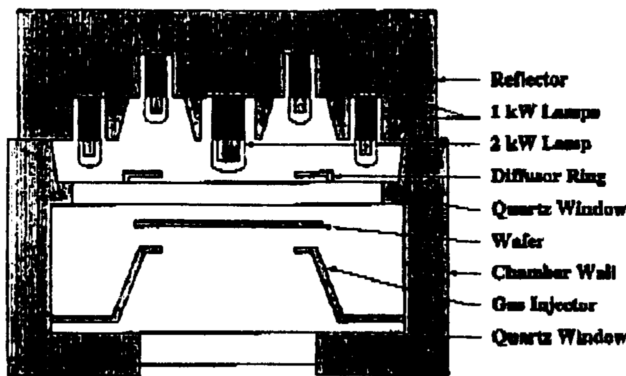


Figure 73.10 Schematic of the rapid thermal processor.

This will ome out etter in he final Version ⇒

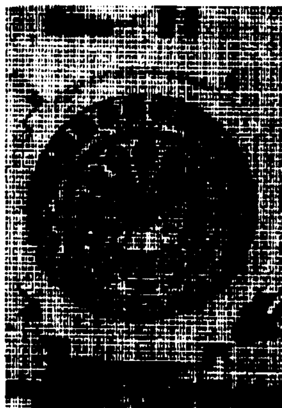


Figure 73.11 Picture of the Stanford three-zone RTM lamp.

pressed as [3] and [6]

$$\rho V_i C_p \frac{dT_i}{dt} = -\epsilon\sigma A_i \sum_{j=1}^{N} D_{i,j} T_j^4 - h_i A_i (T_i - T_{gas}) + q_i^{cond} + q_i^{wall} + q_i^{dist} + \epsilon \sum_{j=1}^{M} F_{i,j} P_j \quad (73.6)$$

where ρ is density, V_i is the volume of the annular element, C_p is heat capacity, T_i is temperature, ϵ is total emissivity, σ is the Stefan-Boltzmann constant, A_i is the surface area of the annular element, $D_{i,j}$ is a lumped parameter denoting the energy transfer due to reflections and emission, h_i is a convective heat transfer coefficient, q_i^{cond} is heat transfer due to conduction, $F_{i,j}$ is a view factor that represents the fraction of energy received by the ith annular element from the jth lamp zone, and P_j is the power from the jth lamp zone.

To develop a simpler model, the temperature distribution of the wafer was considered nearly uniform and much greater than that of the water-cooled chamber walls. With these approximations, q_i^{cond} and q_i^{wall} were negligible. In addition, the term accounting for radiative energy transport due to reflections can be simplified by analyzing the expansion,

$$\sum_{j=1}^{N} D_{i,j} T_j^4 = \sum_{j=1}^{N} D_{i,j} \quad (73.7)$$

$$\left(T_i^4 + 4T_i^3\delta_{i,j} + 6T_i^2\delta_{i,j}^2 + 4T_i\delta_{i,j}^3 + \delta_{i,j}^4\right),$$

where $\delta_{i,j} = T_j - T_i$. After eliminating the terms involving $\delta_{i,j}$ (since $T_i >> \delta_{i,j}$), the resulting model was,

$$\rho V C_p \frac{dT_i}{dt} = -\epsilon\sigma A_i T_i^4 \sum_{j=1}^{N} D_{i,j} - h_i A_i (T_i - T_{ambient}) + \epsilon \sum_{j=1}^{M} F_{i,j} P_j. \quad (73.8)$$

It was noted that Equation 73.8 was interactive because each lamp zone affects the temperature of each annular element and noninteractive because the annular elements did not affect one another.

The nonlinear model given by Equation 73.8 was then linearized about an operating point $(\bar{T}_i, \bar{P}_i)$,

$$\rho V C_p \frac{d\tilde{T}_i}{dt} = -\left[4\epsilon\sigma A_i \bar{T}_i^3 \sum_{j=1}^{N} D_{i,j} + h_i A_i\right] \tilde{T}_i + \epsilon \sum_{j=1}^{M} F_{i,j} \tilde{P}_j, \quad (73.9)$$

where the deviation variables are defined as $\tilde{T}_i = T_i - \bar{T}_i$ and $\tilde{P}_i = P_i - \bar{P}_i$. This equation can be expressed more conveniently as

$$\tau_i \frac{d\tilde{T}_i}{dt} = -\tilde{T}_i + \sum_{j=1}^{M} K_{i,j} \tilde{P}_j, \quad (73.10)$$

where the gain and time-constant are given by

$$K_{i,j} = \frac{\epsilon F_{i,j}}{4\epsilon\sigma A_i \bar{T}_i^3 \sum_{j=1}^{N} D_{i,j} + h_i A_i} \quad (73.11)$$

$$\tau_i = \frac{\rho V C_p}{4\epsilon\sigma A_i \bar{T}_i^3 \sum_{j=1}^{N} D_{i,j} + h_i A_i} \quad (73.12)$$

From Equation 73.11, the gain decreases as $\bar{T}$ was increased. Larger changes in the lamp power were required at higher $\bar{T}$ to achieve an equivalent rise in temperature. In addition, from Equation 73.12, the time constant decreases as $\bar{T}$ is increased. Thus, the wafer temperature responded faster to changes in the lamp power at higher $\bar{T}$. The nonlinearities due to temperature were substantial, as the time constant and gain vary by a factor of 10 over the temperature range associated with RTP.

The identification scheme to estimate τ_i and **K** from experimental data is described in [7], [8]. A sequence of lamp power values was sent to the RTP system. This sequence was known as a recipe. The recipe was formulated so that reasonable spatial temperature uniformity was maintained at all instants in order to satisfy the approximation used in the development of the low-order model. The eigenvalues of the system were estimated at various temperature using a procedure employing the TLS ESPRIT algorithm [9]. After the eigenvalues were estimated, the

amplitude of the step response was estimated. This was difficult because of the temperature drift induced by the window heating; however, a least-squares technique can be employed. The gain of the system and view factors were then identified using a least-squares algorithm again. The results are shown in Figure 73.12 and 73.13 for the estimation of the effects of temperature on the gain and time constant, respectively.

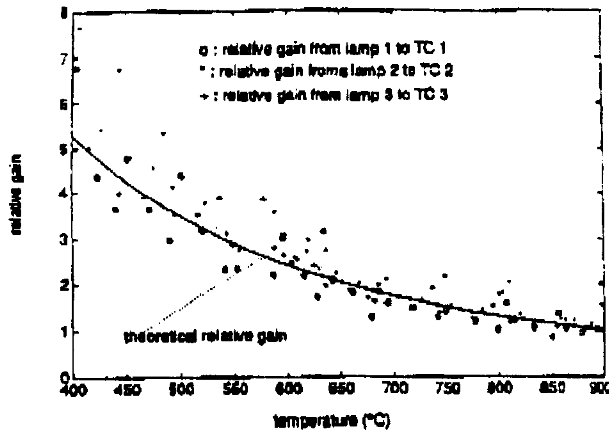


Figure 73.12 Gain of the system relative to that at 900°C and comparison with theory.

The model was expressed in discrete-time format for use in designing a control system. Using the zero-order hold to describe the input sequence, the discrete-time expression of Equation 73.10 was given by

$$\tilde{T}(z) = \Gamma(z)K\tilde{P}(z) \tag{73.13}$$

where z denotes the z-transform,

$$\Gamma(z) = diag\left[\frac{(1 - e^{-\Delta t/\tau_i})z^{-1}}{1 - e^{-\Delta t/\tau_i}z^{-1}}\right] \tag{73.14}$$

and Δt denotes the sampling time. The system model was inherently stable since the poles lie within the unit circle for all operating temperatures.

In order to obtain a complete description of the system, this relationship was combined with models describing sensor dynamics and lamp dynamics [4]. The sensor and lamp dynamics can be described by detailed models. However, for the purpose of model-based control system design, it was only necessary to approximate these dynamics as a simple time-delay relation,

$$\tilde{T}_{m,i}(t) = \tilde{T}_i(t - \theta). \tag{73.15}$$

The measured temperature at time t was denoted by $T_{m,i}(t)$, and the time delay was denoted by θ. The resulting model expressed in z-transform notation was given by

$$\tilde{T}_m(z) = z^{-d}\Gamma(z)K\tilde{P}(z) \tag{73.16}$$

where $d = \theta/\Delta t$ was rounded to the nearest integer.

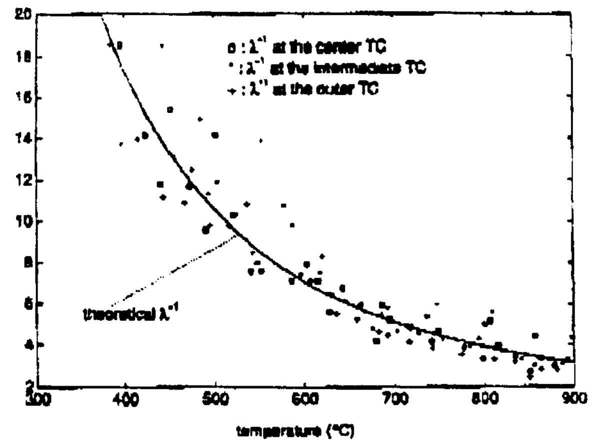


Figure 73.13 Time constant of the process model as a function of temperature and comparison with theory.

The power was supplied to the lamp filaments by sending a 0–10 volt signal from the computer control system to the power supplies that drive the lamps. The relation between the voltage signal applied to the supplies and the power sent to the lamps was not necessarily linear. Consequently, it was important to model that nonlinearity, if possible, so that it could be accounted for by the control system. It was possible to determine the nonlinearity with a transducer installed on the power supply to measure the average power sent to the lamps. By applying a known voltage to the power supplies and then recording the average power output, the desired relationship can be determined. This function can be described by a polynomial and then inverted to remove the nonlinearity from the loop because the model is linear with respect to radiative power from the lamps (see Equation 73.16). We noted that the average power to the lamps may not equal the radiative power from the lamps. The offset was due to heating losses within the bulb filament. However, this offset was implicitly incorporated in the model when the gain matrix, K, was determined from experimental data.

A second strategy that considered the RTP system as a black box was employed to identify a linear model of the process [11], [12]. Among numerous alternatives, an ARX model was used to describe the system,

$$\begin{aligned} T(t) - T_{ss} &= \sum_{k=1}^{n_a} A_k(T(t-k) - T_{ss}) \\ &\quad + \sum_{k=1}^{n_b} B_k(P(t-k) - P_{ss}) + n(t), \end{aligned} \tag{73.17}$$

where T is an $l \times 1$ vector describing temperature, P is an $M \times 1$ vector describing percent of maximum zone power, A_k is an $l \times l$ matrix, B_k is an $l \times M$ matrix, and l is the number of sensors on the wafer measuring temperature. The steady-state temperature and power were denoted by T_{ss} and P_{ss}, respectively. Because the steady-state temperature was difficult to determine accurately because of drift, a slight modification to the model was made. Let $T_{ss} = \hat{T} + \Delta\hat{T}$. The least squares problem for

model identification can then be formulated as

$$min_{A_i,B_i} \sum_{t=1}^{N} \left\| (T(t) - \hat{T}) - \sum_{k=1}^{n_a} A_k(T(t-k) - \hat{T}) - \sum_{k=1}^{n_b} B_k(P(t-k) - P_{ss}) - T_{bias} \right\|_2^2 \qquad (73.18)$$

where $T_{bias} = (I - \sum_{k=1}^{n_a} A_k)\Delta\hat{T}$.

The strategy for estimating the unknown model parameters utilized PRBS (pseudo-random binary sequence) to excite the system to obtain the necessary input-output data. The mean temperature that this excitation produced is designated as $\hat{T}$. Some other issues that were accounted for during model identification included the use of a data subsampling method so that the ARX model could span a longer time interval and observe a larger change of the temperature. Subsampling was needed because the data collection rate was 10 Hz and over that interval the temperature changed very little. Consequently, the least-squares formulation, as an identification method, may contain inherent error sources due to the effect of the measurement sensor noise (which was presumed to be Gaussian distributed) and the quantization noise (which was presumed to be uniformly distributed with quantization level of 0.5°C).

With the black box approach, the model order needed to be selected. The criterion used to determine the appropriateness of the model was to be able to make the prediction error smaller than the quantization level using as small a number of ARX model parameters as possible. For our applications, this order was three A matrices and three B matrices with subsampling selected as four.

We are now going to show the value of the identified models by using them to study an important characteristic of the RTP system, its DC gain. For the ARX model with coefficients $\{A_i, B_i\}$, the identified DC gain was given by the formula

$$\text{DC gain} = \mathbf{D}_o = (I - \sum_{i=1}^{3} A_i)^{-1} \sum_{l=1}^{3} B_i$$

Substituting in the appropriate values for 700°C (with $I = 3$, $J = 3$),

$$\mathbf{D}_o = \begin{bmatrix} 2.07 & 4.41 & 4.50 \\ 1.11 & 4.78 & 4.91 \\ 0.73 & 5.08 & 5.51 \end{bmatrix}. \qquad (73.19)$$

Note that the magnitude of the first column of $\mathbf{D}_o$ is smaller than those of the second and third columns, which was due to the difference in the maximum power of each lamp: the first (center) lamp has 2 kW maximum power, the second (intermediate) lamp 12 kW maximum power, and the third (outer) lamp 24 kW maximum power. Also note the similarity of the second and third column of $\mathbf{D}_o$, which says that the second lamp will affect the wafer temperature in a manner similar to the third lamp. As a result, we have effectively two lamps rather than three (recall we have physically three lamps), which may cause difficulties in maintaining temperature uniformity in the steady state because of an inadequate number of degrees of freedom. This conclusion will be more clearly seen from an SVD (Singular Value Decomposition) analysis, about which more will be said later. To increase the independence of the control effects of the two outside lamps, a baffle was installed to redistribute the light energy from the lamps to the wafer. The same identification technique described earlier was used to identify the RTP system model, and the DC gain was computed from the identified model with the result

$$\mathbf{D}_n = \begin{bmatrix} 2.02 & 5.27 & 3.97 \\ 1.19 & 5.53 & 4.01 \\ 0.83 & 5.11 & 5.15 \end{bmatrix}. \qquad (73.20)$$

We can observe that the second column of the new DC gain matrix was no longer similar to the third column, as it was in Equation 73.19. As a result, the three lamps heated the wafer in different ways. The first (center) lamp heated mostly the center of the wafer, and the third (outer) lamp heated mostly the edge of the wafer. On the other hand, the second (intermediate) lamp heated the wafer overall, acting like a bulk heater. Of course, the second lamp heated the intermediate portion of the wafer more than the center and edge of the wafer, but the difference was not so significant.

Even if the idea of installing a baffle was partly motivated by the direct investigation of the DC gain matrix, it was in fact deduced from an SVD (Singular Value Decomposition) analysis of the DC gain matrix.

The SVD of $\mathbf{D}_o$ in Equation 73.19 is given by

$$\begin{aligned} u_1 &= [0.54, 0.57, 0.63], \; u_2 = [-0.80, 0.13, 0.58], \\ u_3 &= [0.25, -0.81, 0.53] \qquad (73.21) \\ v_1 &= [0.18, 0.68, 0.71], \; v_2 = [-0.98, 0.04, 0.21], \\ v_3 &= [0.12, -0.73, 0.67] \qquad (73.22) \\ \sigma_1 &= 12.15, \sigma_2 = 1.12, \sigma_3 = 0.11 \qquad (73.23) \end{aligned}$$

From this, we can conclude that $[1, 1, 1]$ (u_1) is a strong output direction. Of course, u_1 is $[0.54, 0.57, 0.62]$ and is not exactly equal to $[1, 1, 1]$. However, $[0.54, 0.57, 0.62]$ was close to $[1, 1, 1]$ in terms of direction in a 3-dimensional coordinate system and was denoted as the $[1, 1, 1]$ direction, here. Since $[1, 1, 1]$ was the strong output direction, we can affect the wafer temperature in the $[1, 1, 1]$ direction by a minimal input power change. This means that if we maintain the temperature uniformity at the reference temperature (700°C), we can maintain the uniformity near 700°C (say, at 710°C) with a small change in the input lamp power. The weak output direction (the vector u_3 — approximately $[1, -1, 1]$) says that it is difficult to increase the temperature of the center and outer portions of the wafer while cooling down the intermediate portion of the wafer, which was, more or less, expected. The gain of the weak direction (σ_3) was two orders of magnitude smaller than that of the strong direction (σ_3). This meant that there were effectively only two lamps in the RTP system in terms of controlling the temperature of the wafer, even if there were physically three lamps. This naturally led to the idea of redesigning the RTP chamber to get a better lamp illumination pattern. Installing a baffle (see [10] for more details)

DEC-11-1995 12:01 CRC PRESS-EDP DEPT. 407+994+3625 P.03

into the existing RTP system improved our situation as shown in the SVD analysis of the new DC gain $\mathbf{D}_n$ in Equation 73.20. The SVD of $\mathbf{D}_n$ was given by

$$u_1 = [0.56, 0.57, 0.60], \; u_2 = [-0.63, -0.17, 0.76],$$
$$u_3 = [0.53, -0.80, 0.26] \quad (73.24)$$
$$v_1 = [0.19, -0.72, 0.67], \; v_2 = [0.76, -0.3, -0.57],$$
$$v_3 = [0.63, 0.61, 0.48] \quad (73.25)$$
$$\sigma_1 = 12.14, \; \sigma_2 = 1.17, \; \sigma_3 = 0.52 \quad (73.26)$$

Compared to the SVD of the previous DC gain, the lowest singular value (σ_3) has been increased by a factor of 5, a significant improvement over the previous RTP system. In other words, only one-fifth of the power required to control the temperature in the weak direction, using the previous RTP system, was necessary for the same task with the new RTP system. As a result, we obtained three independent lamps by merely installing a baffle into the existing RTP system. Independence of the three lamps in the new RTP system was crucial in maintaining the temperature uniformity of the wafer.

73.5 Control Design

The general strategy of feedback combined with feedforward control was investigated for RTP control. In this strategy, a feedforward value of the lamp power was computed (in response to a change in the temperature set point) according to a predetermined relationship. This feedforward value was then added to a feedback value and the resultant lamp power was sent to the system. The concept behind this approach was that the feedforward power brings the temperature close to the desired temperature; the feedback value compensates for modeling errors and disturbances.

The feedback value can be determined with a variety of design techniques, two of which are described below. Several approaches were investigated to determine the feedforward value. One approach was based on replaying the lamp powers of previous runs. Another approach was based on a model-based optimization.

The physics-based model was employed to develop a controller using a variation of the classical Internal Model Control (IMC) design procedure [13], [6]. The IMC approach consisted of factoring the linearized form of the nonlinear low-order model (see Equation 73.16) as,

$$\tilde{G}_p(z) = \tilde{G}_p^+(z)\tilde{G}_p^-(z) \quad (73.27)$$

where $\tilde{G}_p^+(z)$ contains the time delay terms, z^{-d}, all right half-plane zeros, zeros that are close to $(-1, 0)$ on the unit disk, and has unity gain. The IMC controller is then obtained by

$$G_c^*(z) = \tilde{G}_p^-(z)^{-1}F(z) \quad (73.28)$$

where $F(z)$ is a matrix of filters used to tune the closed-loop performance and robustness and to obtain a physically realizable controller. The inversion $\tilde{G}_p^-(z)^{-1}$ was relatively straightforward because the dynamics of the annular wafer elements of the linearized form of the nonlinear model were decoupled.

The tuning matrix, or IMC filter, $F(z)$ was selected to satisfy several requirements of RTP. The first requirement was related to repeatability in which zero offset between the actual and desired trajectory was to be guaranteed at steady-state condition despite modeling error. The second requirement was related to uniformity in which the closed-loop dynamics of the wafer temperature should exhibit similar behavior. The third requirement was related to robustness and implementation in which the controller should be as low-order as possible. Other requirements were ease of operator usage and flexibility. One simple selection of $F(z)$ that meets these requirements was given by the first-order filter

$$F(z) = f(z)\mathbf{I}, \quad (73.29)$$
$$f(z) = \frac{1-\alpha}{1-\alpha z^{-1}}, \quad (73.30)$$

where α is a tuning parameter, the speed of response. This provided us with a simple controller with parameters that could be interpreted from a physical standpoint.

In this approach to control design, the nonlinear dependency of K and τ_i on temperature can be parameterized explicitly in the controller. Hence, a continuous gain-scheduling procedure can be applied. It was noted that, as temperature increased, the process gain decreased. Since the controller involved the inverse of the process model, the controller gain increased as temperature was increased. Consequently, the gain-scheduling provided consistent response over the entire temperature range. Thus, control design at one temperature should also apply at other temperatures.

In addition to the IMC approach, a multivariable feedback control law was determined by an LQG design which incorporated integral control action to reduce run-to-run variations [12], [14]. The controller needed to be designed carefully, because, in a nearly singular system such as the experimental RTP, actuator saturation and integrator windup can cause problems. To solve this problem partially, integral control was applied in only the (strongly) controllable directions in temperature error space, helping to prevent the controller from trying to remove uncontrollable disturbances.

For the LQG design, the black-box model was used. It can be expressed in the time domain as follows:

$$\mathbf{y}_k^0 = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}\mathbf{u}_0^0 + \ldots + \mathbf{C}\mathbf{A}\mathbf{B}\mathbf{u}_{k-2}^0 + \mathbf{C}\mathbf{B}\mathbf{u}_{k-1}^0,$$

and the resulting equations ordered from $\mathbf{y}_1^0$ to $\mathbf{y}_N^0$:

$$\begin{bmatrix} \mathbf{y}_1^0 \\ \mathbf{y}_2^0 \\ \vdots \\ \mathbf{y}_N^0 \end{bmatrix} = \begin{bmatrix} \mathbf{CB} & \cdots & \mathbf{0} \\ \mathbf{CAB} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{CA}^{N-1}\mathbf{B} & \cdots & \mathbf{CB} \end{bmatrix} \begin{bmatrix} \mathbf{u}_0^0 \\ \mathbf{u}_1^0 \\ \vdots \\ \mathbf{u}_{N-1}^0 \end{bmatrix}.$$

These combined equations determine a linear relationship between the input and output of the system in the form $\mathbf{Y} = \mathbf{HU}$,

DEC-11-1995 12:02 CRC PRESS-EDP DEPT. 407+994+3625 P.04

where the notation Y and U is used to designate the $n_o N \times 1$ and $n_i N \times 1$ stacked vectors.

The identified model of the system was augmented with new states representing the integral of the error along the m easiest to control directions, defined as $\xi \stackrel{def}{=} [\ \xi_1\ \xi_2\ \dots\ \xi_m\ ]^T$. The new system model was, then,

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \xi_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{U}_{1:m}^T\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \xi_k \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{O} \end{bmatrix} \mathbf{u},$$

$$\mathbf{y}_k = \begin{bmatrix} \mathbf{C} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}_k \\ \xi_k \end{bmatrix},$$

where $\mathbf{U}_{1:m}$ is the first m columns of $\mathbf{U}$, the output matrix from the SVD of the open-loop transfer matrix $H(z)|_{z=1} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, and represented the (easily) controllable subspace. The output y then consisted of thermocouple readings and the integrator states (in practice the integrator states were computed in software from the measured temperature errors). Weights for the integrator states were chosen to provide a good transient response. A complete description of the control design can be found in [14], [15].

The goal of our feedforward control was to design, in advance, a reference input trajectory which will cause the system to follow a predetermined reference output trajectory, assuming no noise or modeling error. The approach built upon the analysis done in [16] and expressed the open-loop trajectory generation problem as a convex optimization with linear constraints and a quadratic objective function [14], [15].

The input/output relationship of the system can be described by a linear matrix equation. This relationship was used to convert convex constraints on the output trajectory into convex constraints on the input trajectory.

The RTP system imposed a number of linear constraints. These were linear constraints imposed by the RTP hardware, specifically, the actuators had both saturation effects and a maximum rate of increase.

Saturation constraints were modeled as follows: $\mathbf{P}_k^{lm}$ is defined as the $n_i \times 1$ vector of steady-state powers at the point of linearization of the above model at time k. With an outer feedback control loop running, to insure that the feedback controller has some room to work (for example ±10% leeway), the total powers should be constrained to the operating range of $10 \leq \mathbf{P}^{total} \leq 90$, which translates into a constraint on U of $(10 - \mathbf{P}^{lm}) \leq \mathbf{U} \leq (90 - \mathbf{P}^{lm})$.

Maximum rates of increase (or slew rate limits) for our actuators were included also. These were due to the dynamics of the halogen bulbs in our lamp. We included this constraint as $\mathbf{u}_{k+1}^0 - \mathbf{u}_k^0 \leq 5$, which can be expressed in a matrix equation of the form $\mathbf{SU} \leq 5$, where S has 1 on the main diagonal and -1 on the off-diagonal.

The quality of our optimized trajectory can be measured in two ways: minimized tracking error (following a reference trajectory) and minimized spatial nonuniformity across the wafer. Because the tracking error placed an upper bound on the nonuniformity error, we concentrate on it here. We define the desired trajectory $\mathbf{Y}^{ref}$ as relative to the same linearized starting point used for system identification. The tracking error E can be defined as $\mathbf{E} = \mathbf{Y} - \mathbf{Y}^{ref}$, where E again denotes the stacked error vector.

We define our objective function to be a quadratic constraint on E as $F(x) = \mathbf{E}^T\mathbf{E}$, and expand

$$F(x) = \mathbf{U}^T\mathbf{H}^T\mathbf{H}\mathbf{U} - 2(\mathbf{Y}^{ref})^T\mathbf{H}\mathbf{U} + (\mathbf{Y}^{ref})^T\mathbf{Y}^{ref}. \quad (73.31)$$

Software programs exist, such as the FORTRAN program LSSOL [17], which can take the convex constraints and produce a unique solution, if one exists.

After achieving successful results in simulation, the control system was implemented in a real-time computing environment linked to the actual RTP equipment. The computing environment included a VxWorks real-time operating system, SUN IPC workstation, VME I/O boards and a Motorola 68030 processor.

73.6 Proof-of-Concept Testing

The Stanford RTP system was used for multiprocessing applications in which sequential thermal process steps were performed within the same reactor. A schematic of the RTM is shown in Figure 73.10. A concentric three-zone 38-kW illuminator, constructed and donated by Texas Instruments, was used for wafer heating. The center zone consisted of a 2-kW bulb, the intermediate zone consisted of 12 1-kW bulbs and the outer zone consisted of 24 1-kW bulbs. A picture of the three-zone lamp is presented in Figure 73.11. The reflector was water and air cooled. An annular gold-plated stainless steel opaque ring was placed on the quartz window to provide improved compartmentalization between the intermediate and outer zones. This improvement was achieved by reducing the radiative energy from the outer zone impinging on the interior location of the wafer and from the intermediate zone impinging on the edge of the wafer. The RTM was used for 4-inch wafer processing. The wafer was manually loaded onto three supporting quartz pins of low thermal mass. The wafer was placed in the center of the chamber which was approximately 15 inches in diameter and 6 inches in height. Gas was injected via two jets. For the control experiments presented below, temperature was measured with a thermocouple instrumented wafer. Three thermocouples were bonded to the wafer along a common diameter at radial positions of 0 inch (center), 1 inch, and 1 7/8 inches. The experiments were conducted in a N_2 environment at 1 atmosphere pressure.

The control algorithms were evaluated for control of temperature uniformity in achieving a ramp from room temperature to 900°C at a ramp rate of 45°C/s followed by a hold for 5 minutes at 1 atm pressure and 1000 sccm (cc/min gas at standard conditions) N_2 [4]. This trajectory typified low-temperature thermal oxidation or annealing operations. The ramp rate was selected to correspond to the performance limit (in terms of satisfying uniformity requirements) of the equipment. The control system utilized simultaneous IMC feedback and feedforward control. Gain scheduling was employed to compensate for the nonlinearities induced by radiative heating.

The wafer temperature for the desired trajectory over the first 100 seconds is plotted in Figure 73.14 for the center, middle, and

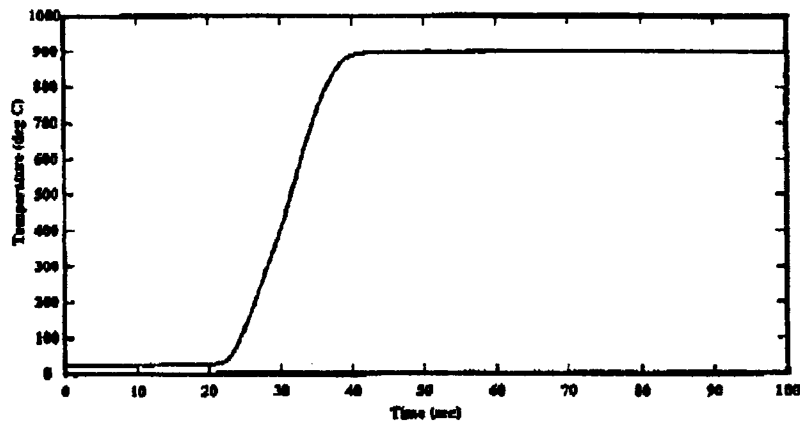


Figure 73.14 Temperature trajectory of the three sensors over the first 100 seconds and the 5 minute process hold.

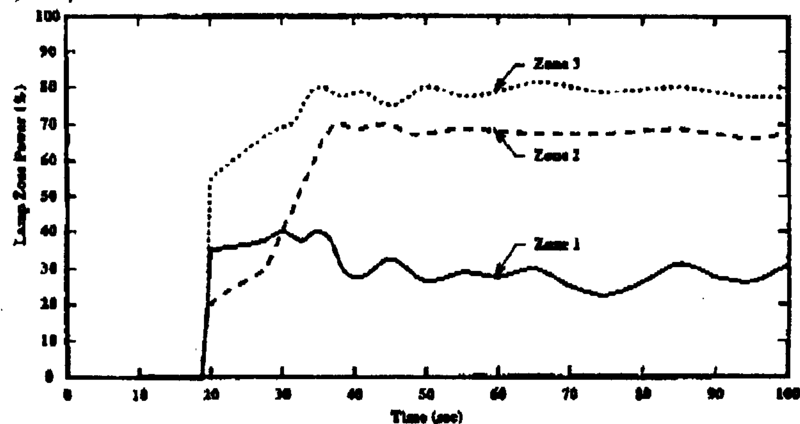


Figure 73.15 Powers of the three zones used to control temperature over the first 100 seconds.

edge locations where thermocouples are bonded to the wafer along a common diameter at radial positions of 0 inch (center), 1 inch, and 1 7/8 inches.

The ramp rate gradually increased to the specified 45°C/s and then decreased as the desired process hold temperature was approached. The corresponding lamp powers, that were manipulated by commands from the control system to achieve the desired temperature trajectory, are shown in Figure 73.15.

The time delay of the system can be seen by comparing the starting times of the lamp powers to the temperature response. Approximately, a two second delay existed in the beginning of the response. Of this delay, approximately 1.5 seconds was caused by a power surge interlock on the lamp power supplies which only functions when the lamp power is below 15% of the total power. The remaining delay was caused by the sensor and filament heating dynamics. In the power profile plot, the rate limiting of the lamp powers is seen. This rate-limiting strategy was employed as a safety precaution to prevent a large inrush current to the lamps. However, these interlocks prevented higher values of ramp rates from being achieved.

The nonuniformity of the controlled temperature trajectory was then analyzed. From the measurements of the entire five minute run (i.e., the first 100 seconds shown in Figure 73.14 along with an additional 400 second hold at 900°C not shown in the figure), the nonuniformity was computed by the peak-to-peak temperature error of the temperature measurements of the three thermocouples. The result is plotted in Figure 73.16. The maximum temperature nonuniformity of approximately 15°C occurred during the ramp around a mean temperature of 350°C. This nonuniformity occurred at a low temperature and does not effect processing or damage the wafer via slip. As the ramp progressed from this point, the nonuniformity decreased. The significant sensor noise can be seen.

The capability of the controller to hold the wafer temperature at a desired process temperature despite the presence of dynamic heating from extraneous sources was then examined. As seen in Figure 73.14, the control system held the wafer temperature at the desired value of 900°C. Although the sensors were quite noisy and had resolution of 0.5°C, the wafer temperature averaged over the entire hold portion for the three sensors corresponded to 900.9°C, 900.7°C, and 900.8°C, respectively. This result was desired because the uniformity of the process parameters, such as film thickness and resistivity, generally depend on the integrated or averaged temperature over time. The capability of the control system to hold the wafer temperature at the desired value, albeit slightly higher, is demonstrated by plotting the dynamics of the quartz window and chamber base of the RTM in Figure 73.17.

The slow heating of these components of the RTM corresponded to slow disturbances to the wafer temperature. Because of the reduced gain of the controller to compensate for time de-

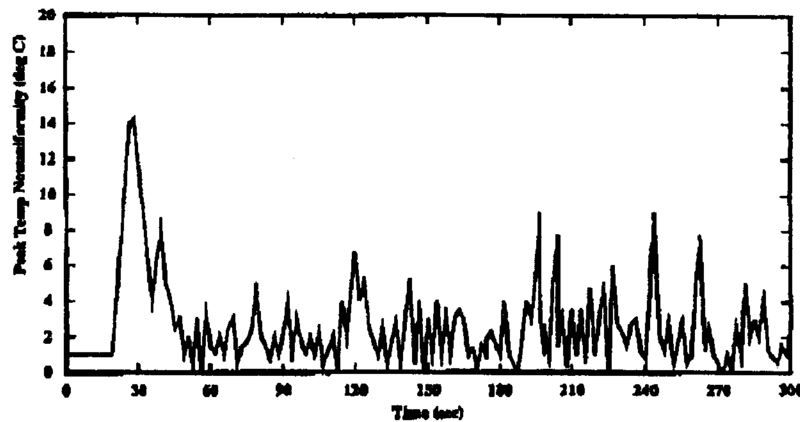


Figure 73.16 Temperature nonuniformity for the 5 minute run as measured by the three temperature sensors.

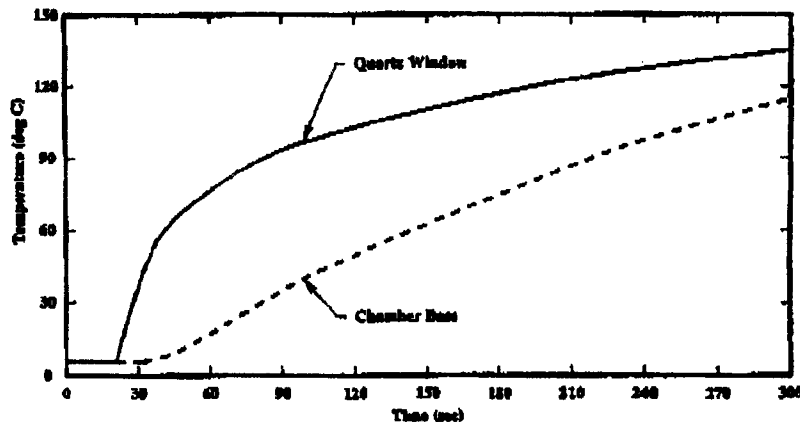


Figure 73.17 Temperatures of the quartz window and chamber base over the 5 minute run.

lays, these disturbances impacted the closed-loop response by raising the temperature to a value slightly higher than the set point. However, without the feedback temperature control system, the wafer temperature would have drifted to a value more than 50°C higher than the set point as opposed to less than 1°C in the measured wafer temperature.

73.7 Technology Transfer to Industry

After demonstrating the prototype RTP equipment at Stanford, the multivariable control strategy (including hardware and software) was transferred to Texas Instruments for application in the MMST program. This transfer involved integration on eight RTP reactors: seven on-line and one off-line. These RTP systems were eventually to be used in a 1000 wafer demonstration of two full-flow sub-half-micron CMOS process technologies in the MMST program at TI [18], [19],[20],[21] and [22].

Although there were similarities between the RTP equipment at TI and the three-zone RTP system at Stanford, there were also substantial differences. Two types of illuminators were used for MMST, a four-zone system constructed at Texas Instruments for MMST applications and a six-zone system manufactured by G^2 Semiconductor Corp. Both systems utilized concentric zone heating; the TI system employed a circular arrangement of lamps, and the G^2 system used a hexagonal arrangement of lamps. A thick (roughly 15 mm) quartz window was used in the TI system to separate the lamps from the reaction chamber, and a thin (3 mm) quartz window was used in the G^2 system. Wafer rotation was not employed with the TI system but was used with the G^2 system. The rotation rate was approximately 20 rpm. Moreover, six-inch wafer processing took place using up to four pyrometers for feedback. Most reactors employed different configurations purge ring assemblies, guard rings, and susceptors (see Figure 73.5).

The on-line RTP reactors configured with the IMC controller were used for thirteen different thermal processes: LPCVD Nitride, LPCVD Tungsten, Silicide react, Silicide anneal, sinter, LPCVD polysilicon, LPCVD amorphous silicon, germane clean, dry RTO, wet RTO, source/drain anneal, gate anneal, and tank anneal. These processes ranged from 450° to 1100°C, from 1 to 650 torr pressure, and from 30 seconds to 5 minutes of processing time (see Figure 73.18).

There were several challenges in customizing the temperature control system for operation in an all-RTP factory environment [13]. These challenges included substantial differences among the eight reactors and thirteen processes, operation in a prototyping development environment, ill-conditioned processing equipment, calibrated pyrometers required for temperature sensing, equipment reliability tied to control power trajectories,

DEC-07-1995 16:29 CRC PRESS-EDP DEPT. 407+994+3625 P.12/12

RTP Process	Carrier Gas	T_{ph} [°C]	T_{pr} [°C]	t_{ph} [s]	t_{pr} [s]	P [Torr]
sinter	N_2	450	450	0	180	650
LPCVD-W	Ar/H_2	425-475	425-475	0	60-180	30
LPCVD-amor Si	Ar	450-500	500-560	5-15	60-180	15
LPCVD-poly	Ar	450-550	650	5-15	120-240	15
silicide react	N_2	450-500	650	5-15	180	1
silicide anneal	Ar	400-550	750	5-15	60	1
LPCVD-SiO_2	O_2	450-650	750	5-15	30-180	1-5
germanate clean	H_2	480-550	650-750	5-15	120	15
LPCVD-Si_3N_4	NH_3	550-650	850	5-15	60-180	1-5
gate RTA	Ar	750-800	900	5-15	30	650
dry RTO	O_2	750-800	1000	5-15	120-180	650
wet RTO	O_2	750-800	1000	5-15	120-180	650
source/drain RTA	Ar	750-800	1000-1050	5-15	15-30	650
tank RTA	NH_3	750-800	1100	5-15	300	650

Figure 73.18 List of processes controlled during the MMST program. The preheat temperature (T_{ph}) and time (t_{ph}) and the process temperature (T_{pr}) and time (t_{pr}) are given. The carrier gases and operating pressures are also presented.

multiple lamp-zone/sensor configurations, detection of equipment failures, and numerous operational and communication modes.

Nonetheless, it was possible to develop a single computer control code with the flexibility of achieving all of the desired objectives. This was accomplished by developing a controller in a modular framework based on a standardized model of the process and equipment. The control structure remained the same while the model-based parameters of the controller differed from process to process and reactor to reactor. It was possible to read these parameters from a data file while holding the controller code and logic constant. Consequently, it was only necessary to maintain and modify a single computer control code for the entire RTP factory.

We present results here for an LPCVD-Nitride process that employed a TI illuminator and for an LPCVD-Poly process that employed a G^2 illuminator. Additional temperature and process control results are presented in [13] and [6].

The desired temperature trajectory for the LPCVD-Nitride process involved a ramp to 850°C and then a hold at 850°C for roughly 180 seconds in a SiH_4/NH_3 deposition environment. Temperature was measured using four radially distributed 3.3 μm InAs pyrometers. The center and edge pyrometers were actively used for real-time feedback control and the inner two pyrometers were used to monitor the temperature. The reasons for this analysis were: (1) repeatable results were possible using only two pyrometers, (2) an analysis of the benefits of using pyrometers for feedback could be assessed, and (3) fewer pyrometers were maintained during the marathon demonstration. In Figure 73.19, the center temperature measurement is shown for a 24-wafer lot process. The offsets in the plot during the ramps are merely due to differences in the starting points of the ramps.

During the hold at 850°C, the reactive gases were injected, and the deposition took place. The standard deviation (computed over the 24 runs) of the temperature measurements during the deposition time was analyzed. In Figure 73.20, the standard deviation of the four sensor measurements are shown. The controlled sensors improved repeatability over the monitored sensor locations by a factor of seven. A three-sigma interpretation shows roughly that the controlled sensors held temperature to within $\pm$0.3°C and the monitored sensors were repeatable at $\pm$2.0°C.

We analyzed the power trajectories to the lamp zones to evaluate the repeatability of the equipment. In Figure 73.21, the power to the center zone is presented for the 24 runs. The intermediate two zones were biased off the center and edge zones, respectively. From these results, it was clear that the lamp power decreased substantially during a nitride deposition run because the chamber and window heat more slowly than the wafer; because the chamber and window provide energy to the wafer, the necessary energy from the lamps to achieve a specified wafer temperature was less as the chamber and window heat up. In addition, we noted the chamber and window heating effect from run-to-run by observing the lowered lamp energy requirements as the lot processing progresses. These observations can be used in developing fault detection algorithms.

To study the capability of temperature control on the process parameter, we compared the thickness of the LPCVD poly process determined at the center for each wafer of a 24-wafer lot where multizone feedback temperature control was used and no real-time feedback temperature control (i.e., open-loop operation) was used. For the open-loop case, a predetermined lamp power trajectory was replayed for each wafer of the 24-wafer lot. The comparison is shown in Figure 73.22. It is clear that the multizone feedback control is much better than open-loop control. In some sense, this comparison is a worst case analysis since the lamp powers themselves for both cases had no control, not usual in industry. In our experiments, variations in line voltage proceeded unfiltered through the reactor causing unprovoked fluctuations in the lamp power and inducing strong temperature effects. However, the feedback temperature control system can compensate somewhat for these fluctuations. For the open-loop case, these fluctuations pass on directly and result in unacceptable repeatability.

73.8 Conclusions

A systems approach has been used for a study in semiconductor manufacturing. This methodology has included developing models, analyzing alternative equipment designs from a control perspective, establishing model identification techniques to develop a model for control design, developing a real-time control system and embedding it within a control processor, proof-of-concept testing with a prototype system, and then transferring the control technology to industry. This application has shown the role that control methodologies can play in semiconductor device manufacturing.

TOTAL P.12

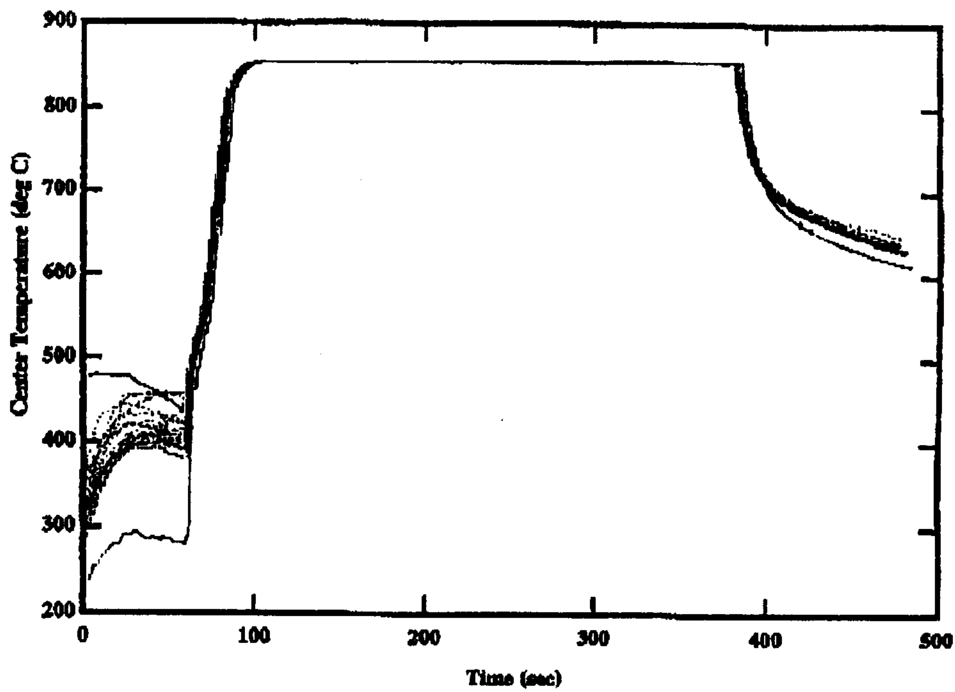


Figure 73.19 The center temperature of the LPCVD nitride process for a 24-wafer lot run.

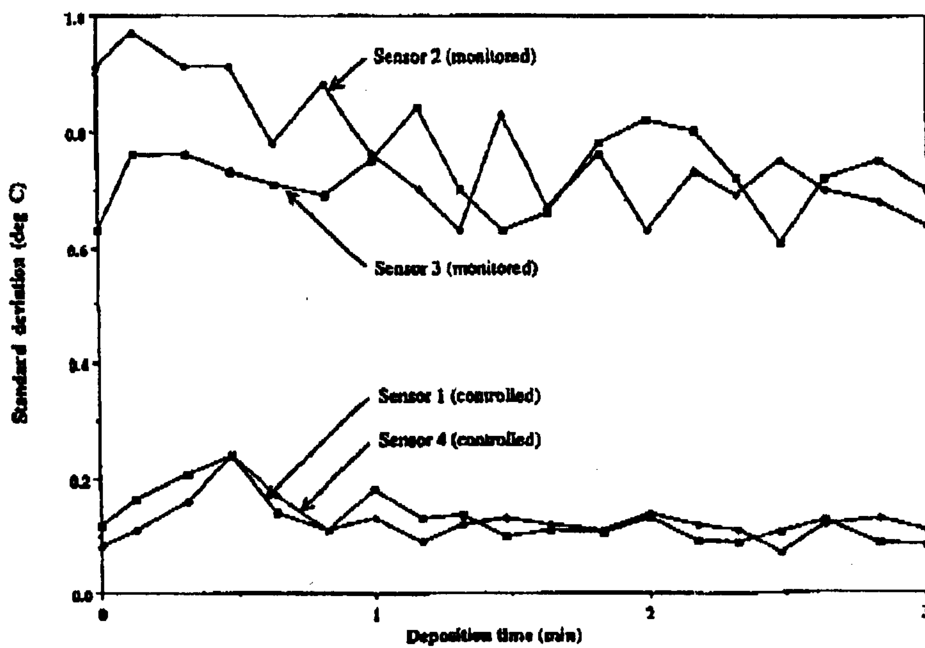


Figure 73.20 Standard deviation of temperature measurements during the three minute deposition step of the LPCVD nitride process for the 24-wafer lot run.

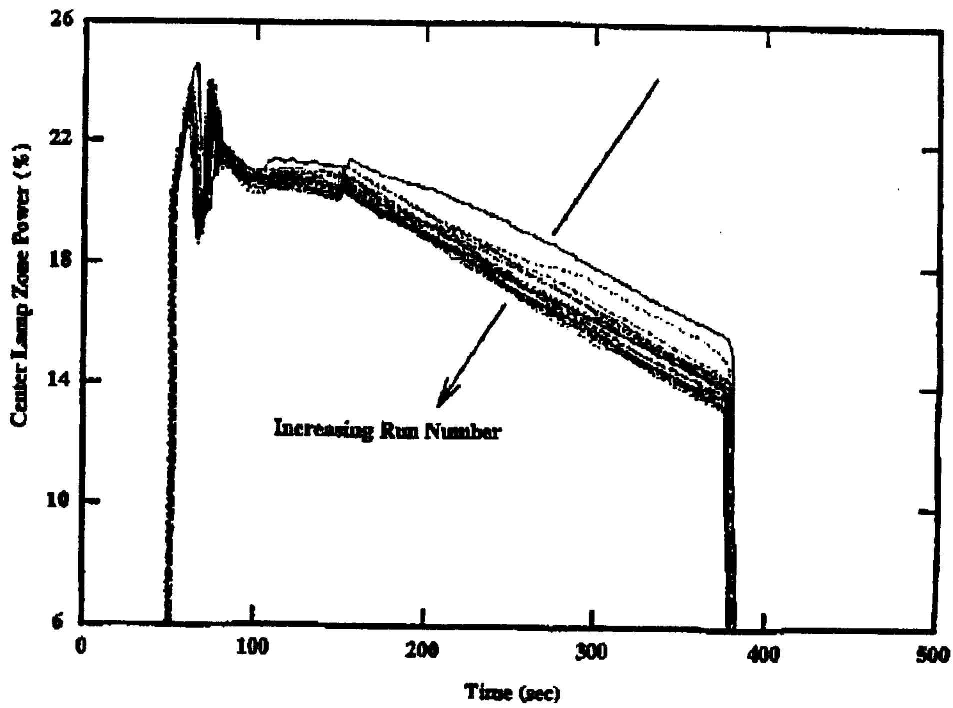


Figure 73.21 Power to the center zone for the 24-wafer lot run of the LPCVD nitride process.

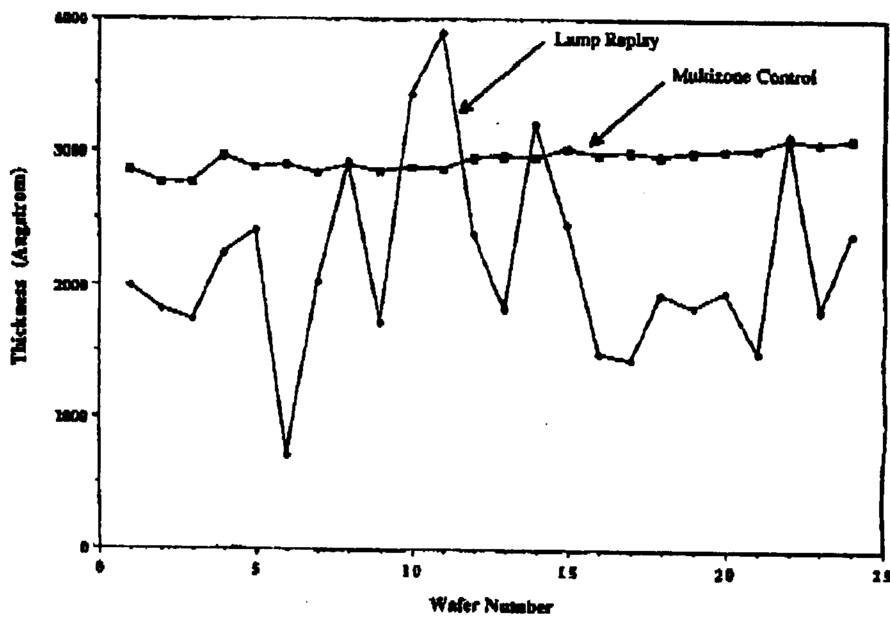


Figure 73.22 Comparison of multizone feedback control and open-loop lamp replay for LPCVD polysilicon process.

References

[1] Lord, H., Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven, *IEEE Trans. Semicond. Manufact.*, 1, 141–150, 1988.

[2] Norman, S.A., *Wafer Temperature Control in Rapid Thermal Processing*, Ph.D. Thesis, Stanford University, 1992.

[3] Cho, Y., Schaper, C. and Kailath, T., Low order modeling and dynamic characterization of rapid thermal processing, *Appl. Phys. A: Solids and Surfaces*, A:54(4), 317–326, 1992.

[4] Norman, S.A., Schaper, C.D. and Boyd, S.P., Improvement of temperature uniformity in rapid thermal processing systems using multivariable control. In *Mater. Res. Soc. Proc.: Rapid Thermal and Integrated Processing*. Materials Research Society, 1991.

[5] Saraswat, K. and Apte, P., Rapid thermal processing uniformity using multivariable control of a circularly symmetric three-zone lamp, *IEEE Trans. on Semicond. Manufact.*, 5, 1992.

[6] Saraswat, K., Schaper, C., Moslehi, M. and Kailath, T., Modeling, identification, and control of rapid thermal processing, *J. Electrochem. Soc.*, 141(11), 3200–3209, 1994.

[7] Cho, Y., *Fast Subspace Based System Identification: Theory and Practice*, Ph.D. Thesis, Stanford University, CA, 1993.

[8] Cho, Y. and Kailath, T., Model identification in rapid thermal processing systems, *IEEE Trans. Semicond. Manufact.*, 6(3), 233–245, 1993.

[9] Roy, R., Paulraj, A. and Kailath, T., ESPRIT - a subspace rotation approach to estimation of parameters of cisoids in noise, *IEEE Trans. ASSP*, 34(5), 1340–1342, 1986.

[10] Schaper, C., Cho, Y., Park, P., Norman, S., Gyugyi, P., Hoffmann, G., Balemi, S., Boyd, S., Franklin, G., Kailath, T., and Sarawat, K., Dynamics and control of a rapid thermal multiprocessor. In *SPIE Conference on Rapid Thermal and Integrated Processing*, September 1991.

[11] Cho, Y.M., Xu, G., and Kailath, T., Fast recursive identification of state-space models via exploitation of displacement structure, *Automatica*, 30(1), 45–59, 1994.

[12] Gyugyi, P., Cho, Y., Franklin, G., and Kailath, T., Control of rapid thermal processing: A system theoretic approach. In *IFAC World Congress*, 1993.

[13] Saraswat, K., Schaper, C., Moslehi, M., and Kailath, T., Control of MMST RTP: Uniformity, repeatability, and integration for flexible manufacturing, *IEEE Trans. on Semicond. Manifact.*, 7(2), 202–219, 1994.

[14] Gyugyi, P., *Application of Model-Based Control to Rapid Thermal Processing Systems*. Ph.D. Thesis, Stanford University, 1993.

[15] Gyugyi, P.J., Cho, Y.M., Franklin, G., and Kailath, T., Convex optimization of wafer temperature trajectories for rapid thermal processing. In *The 2nd IEEE Conf. Control Appl.*, Vancouver, 1993.

[16] Norman, S.A., Optimization of transient temperature uniformity in RTP systems, *IEEE Trans. Electron Dev.*, January 1992.

[17] Gill, P.E., Hammarling, S.J., Murray, W., Saunders, M.A., and Wright, M.H., User's guide for LSSOL (Version 1.0): A FORTRAN package for constrained least-squares and convex quadratic programming, Tech. Rep. SOL 86-1, Operations Research Dept., Stanford University, Stanford, CA, 1986.

[18] Chatterjee, P. and Larrabee, G., Manufacturing for the gigabit age, *IEEE Trans. on VLSI Technology*, 1, 1993.

[19] Bowling, A., Davis, C., Moslehi, M., and Luttmer, J., Microeletronics manufacturing science and technology: Equipment and sensor technologies, *TI Technical J.*, 9, 1992.

[20] Davis, C., Moslehi, M., and Bowling, A., Microeletronics manufacturing science and technology: Single-wafer thermal processing and wafer cleaning, *TI Technical J.*, 9, 1992.

[21] Moslehi, M. et al., Single-wafer processing tools for agile semiconductor production, *Solid State Technol.*, 37(1), 35–45, 1994.

[22] Saraswat, K. et al., Rapid thermal multiprocessing for a programmable factory for adaptable manufacturing of ic's, *IEEE Trans. on Semicond. Manufact.*, 7(2), 159–175, 1994.