

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

Capital¹ 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.

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 semi-

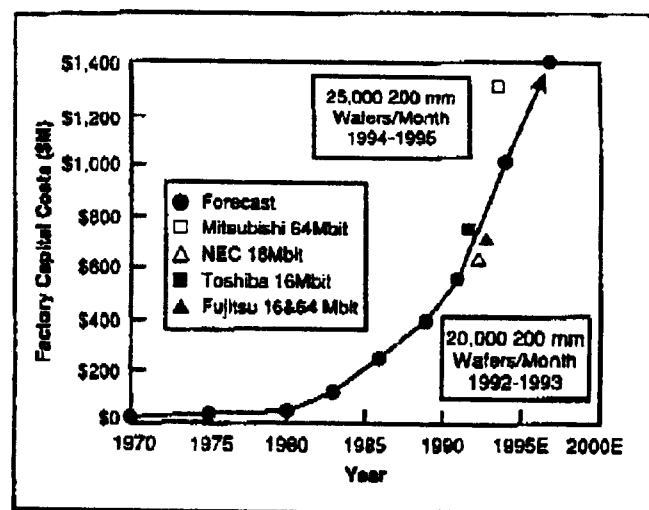


Figure 73.1 Capital cost for a new IC factory. (Source: *Texas Instruments Technical Journal*, 9(5), 8, 1992.)

conductor 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)

¹This 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.

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

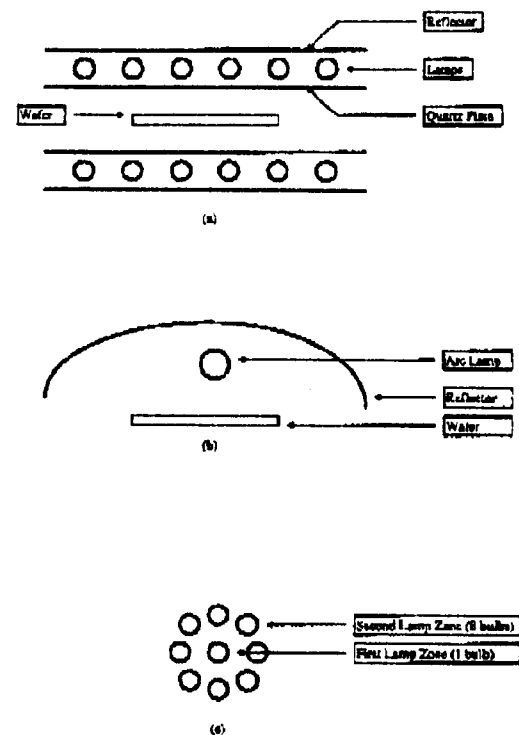


Figure 73.2 RTP lamp configurations: (a) bank of linear lamps, (b) single arc lamp, (c) two-zone lamp array.

only for a narrow range of conditions.

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-

73.1. INTRODUCTION

473

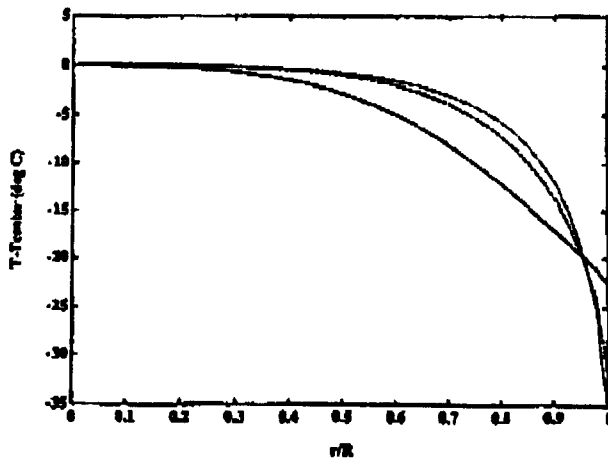


Figure 73.3 Nonuniformity in temperature induced by uniform energy flux impinging on the wafer top surface (center temperatures - solid line: 600°C; dashed line: 1000°C; dotted line: 1150°C.). R is the radius of the wafer, r is the radial distance from the center of the wafer.

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 uncer-

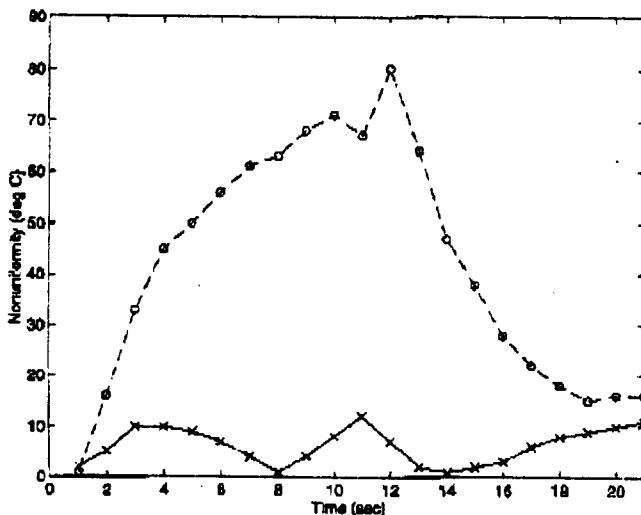


Figure 73.4 Temperature nonuniformity when the powers to the lamp were manually adjusted (“hand-tuning”). These nonuniformities correspond to a ramp and hold from nearly room temperature to 600°C at roughly 40°C/s. The upper curve (-o-) corresponds to scalar control (fixed power ratio to lamps). The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

tain 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.

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 procedure 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 $\pm 5^\circ\text{C}$ from 400°C to the processing temperature and better than $\pm 0.5^\circ\text{C}$ 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

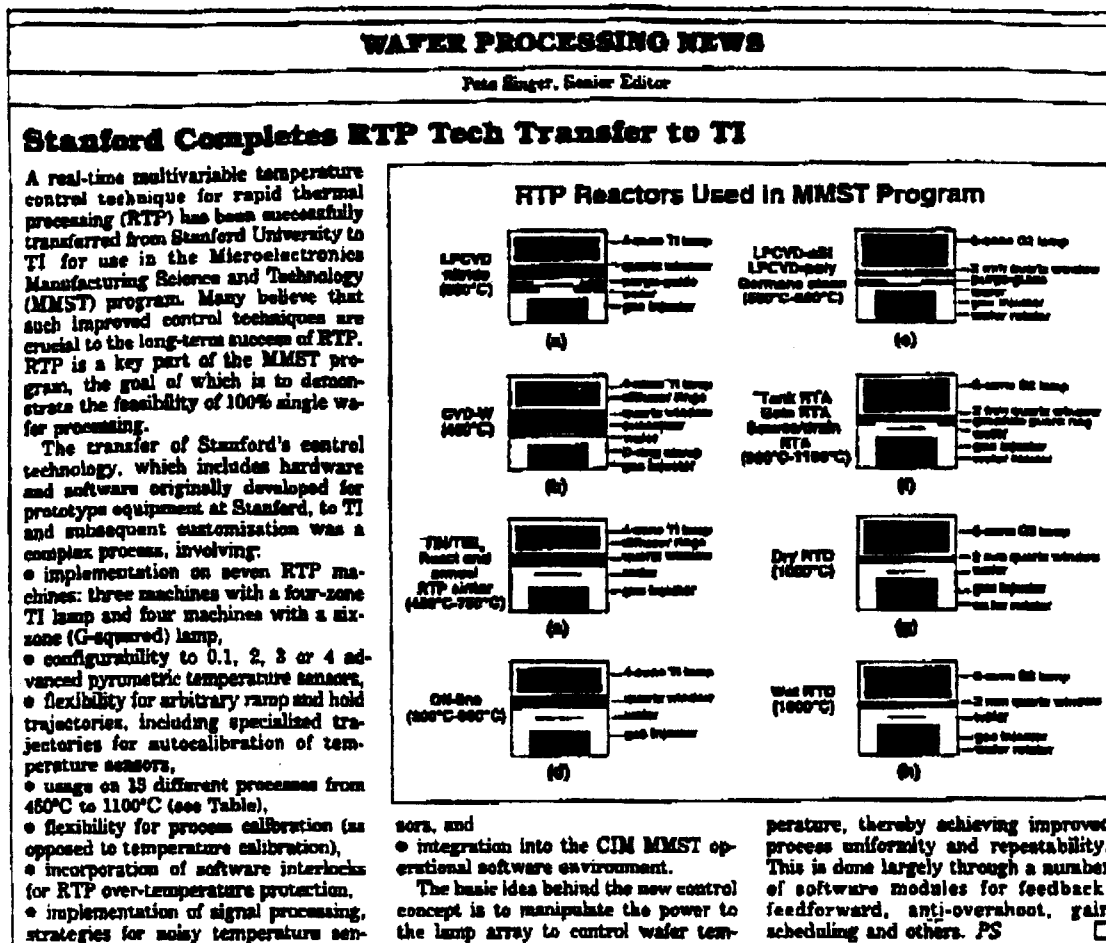


Figure 73.5 Description of technology transfer in *Semiconductor International*, 16(7), 58, 1993.

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

73.3. PERFORMANCE ANALYSIS

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

$$C\dot{T} = K^{rad}T^4 + K^{cond}T + K^{conv}(T - T_{gas}) + FP + q^{wall} + q^{dist} \quad (73.2)$$

where

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

where N denotes the number of wafer elements and M denotes the number of radiant heating zones; K^{rad} is a full matrix describing the radiation emission characteristics of the wafer, K^{cond} is a tridiagonal matrix describing the conductive heat transfer effects across the wafer, K^{conv} is a diagonal matrix describing the convective heat transfer effects from the wafer to the surrounding gas, 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 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,

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

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 $0 \leq P_j \leq P_j^{max}$. Using the finite difference model, the objective function of Equation 73.3 was approximated as

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

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 $\pm 1^\circ\text{C}$, much better than for the case of uniform energy flux.

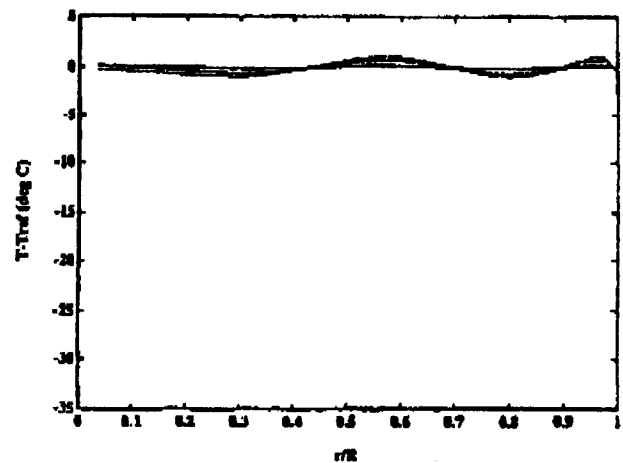


Figure 73.6 Optimal temperature profiles using a multizone RTP system (center temperatures - solid line: 600°C; dashed line: 1000°C; dotted line: 1150°C).

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

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

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

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.

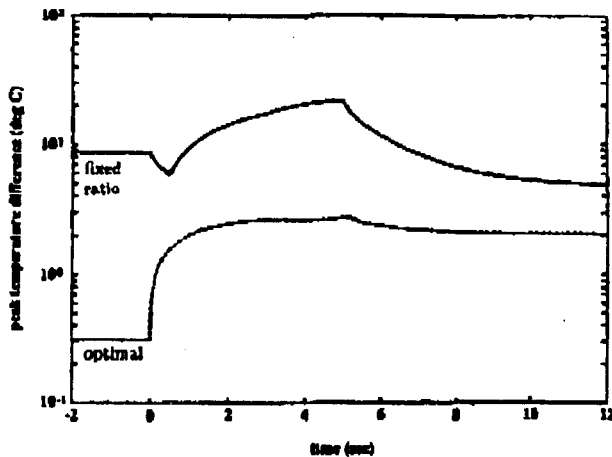


Figure 73.7 Peak temperature nonuniformity during ramp.

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

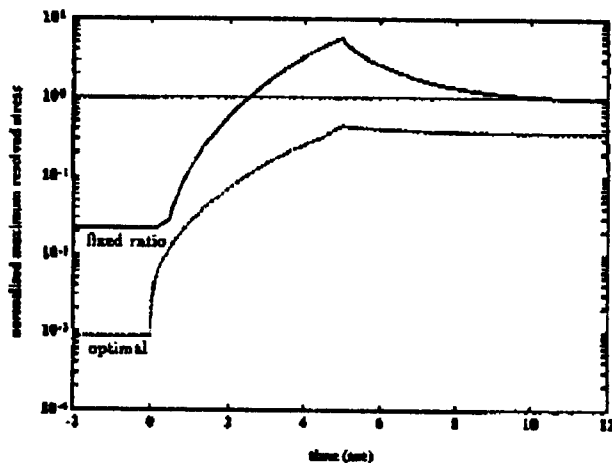


Figure 73.8 Normalized maximum resolved stress during ramp.

(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 multiple 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 marshall 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 *i*th annular element can be ex-

73.4 MODELS FOR CONTROL

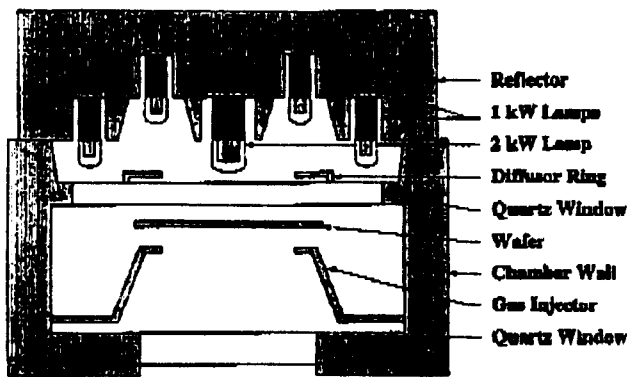


Figure 73.10 Schematic of the rapid thermal processor.

This will come out better in the final version =>

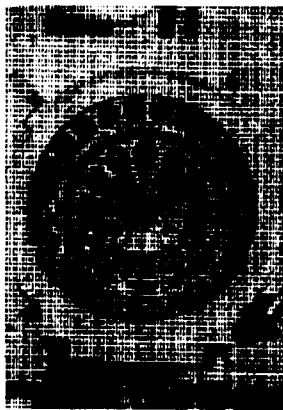


Figure 73.11 Picture of the Stanford three-zone RTM lamp.

pressed as [3] and [6]

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

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 i^{th} annular element from the j^{th} lamp zone, and P_j is the power from the j^{th} 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,

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

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

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

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

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) ,

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

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

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

where the gain and time-constant are given by

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

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

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

