To appear in SPIE Proceedings of
Rapid Thermal and Integrated Processing, 1991.

Modeling and Control of Rapid Thermal Processing

C. Schaper, Y. Cho, P. Park, S. Norman, P. Gyugyi, G. Hoffmann, S. Balemi,
S. Boyd, G. Franklin, T. Kailath, and K. Saraswat

Department of Electrical Engineering
Stanford University, Stanford, CA 94305-4055

Abstract

A first-principles low-order model of rapid thermal processing of semiconductor wafers is derived. The nonlinear model describes the steady-state and transient thermal behavior of a wafer with approximate spatial temperature uniformity undergoing rapid heating and cooling in a multilamp RTP chamber. The model is verified experimentally for a range of operating temperatures from 400°C to 900°C and pressure of 1 torr in an inert N_2 environment. Advantages of the low-order model over detailed models include ease of identification and implementation for real-time predictive applications in signal processing and temperature control. This physics-based model is used in the design of an advanced real-time multivariable control strategy. The strategy employed a feedforward mechanism to predict temperature transients and a feedback mechanism to correct for errors in the prediction. The controller is applied to achieve a ramp from 20°C to 900°C at a rate of 45°C/second in a one atmosphere environment with less than 15°C nonuniformity during the ramp and less than 1°C average nonuniformity during the hold as measured by three thermocouples.

1 INTRODUCTION

Rapid thermal processing (RTP) systems are currently being developed for single-wafer manufacturing of integrated circuits. The process performs thermal related fabrication steps such as annealing, formation of thin dielectric films and chemical vapor deposition. The advantages associated with RTP have been well documented [1, 2].

In order to achieve slip-free and uniform processing, it is necessary to maintain near uniform temperature distribution over the wafer during both steady-state and transient (fast ramping of wafer temperature) situations. Furthermore, this uniform distribution must be achieved for a range of operating conditions including different pressures, gasses, and processing temperatures. In this manner, an RTP system is available for flexible manufacturing applications that can adapt and optimize to changes in processing specifications.

Recent innovations in the design of RTP systems have provided the ability to achieve temperature uniformity over a range of processing conditions [3]. The basic requirement is the ability to vary the spatial energy flux distribution radiating to the wafer as the necessities for wafer temperature uniformity change as a function of operating conditions. To achieve this requirement, one approach is the use of multiple concentric circular rings of lamps that can be manipulated independently, (for example, see [4]). An example of this design is the Stanford RTM (Rapid Thermal Multiprocessor), of which a cross-sectional schematic of the three-zone lamp heating system is shown in Figure 1. An automatic control strategy for the RTM is available to manipulate the power to each of the three lamp zones to achieve wafer temperature uniformity at steady-state and transient conditions. The control strategy employs a multi-point sensor reading to provide real-time measurement of the temperature distribution.

In this paper, an automatic multivariable controller is developed for multi-zone lamp, multi-point sensor RTP systems. It is applied to the RTM. In formulating this controller, the nonlinear dynamics of wafer temperature transients are first studied. This information is used to suggest the type of control structure that needs to be applied. Next, a model of the energy transport mechanisms in RTP systems is developed. This model is used to design the multivariable controller. Because the model is used for control applications, it need only be accurate enough to describe transient thermal behavior of a wafer with approximate spatial temperature uniformity. The model is also required to be flexible to cover a wide range of operating conditions and to retain a parsimonious structure for robust controller design. For our application, the physical laws of wafer heating suggested an appropriate model structure. By formulating the model for control applications, a nonlinear model was obtained with a low number of uncertain parameters that can be accurately determined from a small number of experiments.

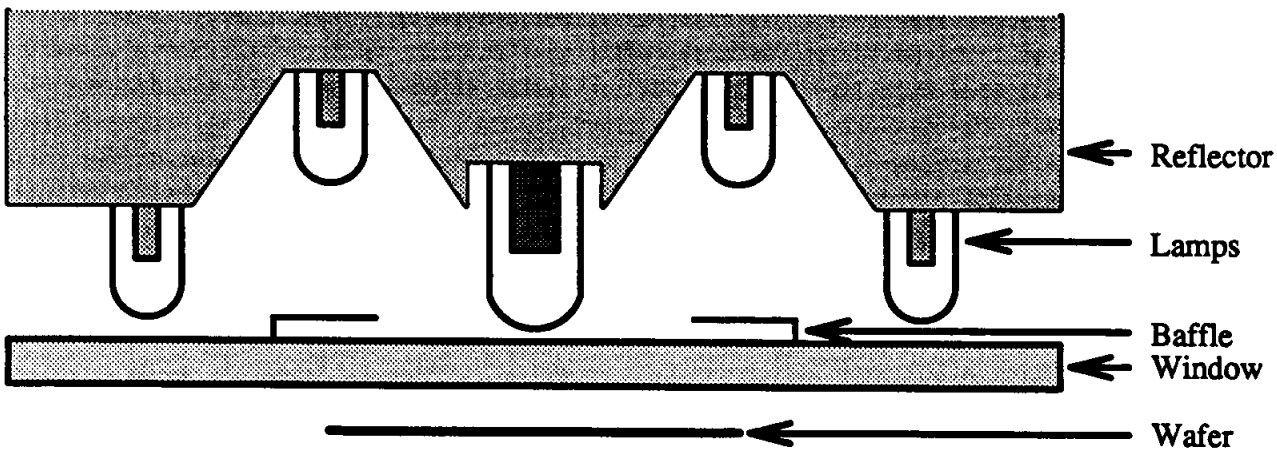


Figure 1: Cross-sectional diagram of the lamp heating arrangement for the Stanford Rapid Thermal Multiprocessor.

The paper is organized as follows. The fundamental model is derived and evaluated in Section 2. The model is then used in the design of a controller described and tested in Section 3 for fast ramps. Design issues of the RTM are also presented in Section 3.

2 MODELING

Modeling of the heat transfer characteristics of RTP systems is useful for several applications including:

- Lamp and chamber design — A model can assist in evaluating alternative designs before they are built. In this case, a detailed model is needed to determine the energy flux requirements to achieve wafer temperature uniformity at steady-state and transient situations under a range of operating conditions. This model may take the chamber and reflector geometry into account, in addition to the three mechanisms of wafer heating, radiation, convection and conduction energy transfer.

- Design of automatic temperature controllers — Models can be used to synthesize advanced process control systems. It is necessary that these models take the nonlinear behavior of wafer temperature heating into account. (This nonlinear behavior is described below.) A model structure that is more parsimonious than the one used for design can be used for control law synthesis and real-time prediction. The model needs to be accurate over a wide range of operating conditions (pressures,

temperatures, flow rates) in order to achieve precision control ($+/-5°C$) although the controller can be designed to be robust to a certain magnitude of modeling error.

- Temperature measurement — Models can be used for advanced signal processing of sensor readings for control and diagnostic applications. In this application, the model can be used to predict wafer temperature at some locations on the wafer where a sensor is not present. Furthermore, the signal processing method can be used to filter noise in the sensor signal and reduce biased temperature readings (for example, a pyrometer with an incorrect estimate of emissivity can be improved using model-based signal processing techniques [5].)

In this section, we focus on developing a model of wafer heating that can be used for control. There are two possible approaches. In the first approach, the RTP system is viewed as a black-box. The physical laws governing the energy transfer in an RTP system are not employed by this approach. Instead, the model is developed by fitting experimental data to an assumed correlation that relates the input signals to the output signals . A black-box model of the RTP system would then relate the signals sent to the lamps to the sensors that record temperature. These black-box models are usually linear in structure. Because of the substantial (input/output) nonlinearities of wafer heating due to radiative heat transfer, a series of linear models is needed. A control strategy can then use this series of linear models to adapt to the nonlinearities.

In this section, the second approach to modeling for control is pursued. In this approach, the physical laws that describe wafer heating are employed. For an RTP system, a physics-based model of the system behavior can be derived and is shown to be nonlinear in structure. With this approach, the underlying nonlinear structure of wafer heating is captured by a single global model. Experimental data can then be fit to the various unknown parameters of this physics-based structure. This single fundamental nonlinear model is used instead of a multitude of linearized black-box models.

2.1 Low-Order Model Development

The development of a physical model of an RTP system is complex. In Figure 2, the relationship between the systems inputs and the system outputs is presented. This relationship essentially consists of three blocks. In the first block, a voltage signal (0 — 5 Volts) is sent to the lamps. In the Stanford RTM, three concentric rings of lamps can be manipulated independently. In this case, three voltage signals are sent to the system. This voltage signal is then converted into radiative power by heating a tungsten filament lamp. There are modeling dynamics associated with the transfer of electrical energy to radiative energy. In the second block, the radiative power from the lamps is absorbed by the semiconductor wafer. In addition, convective, conductive and additional radiative heat transfer mechanisms that take place in the wafer. The energy flux to the wafer is influenced by the chamber, process gasses, pressure, and lamps. In the third block, a sensor measures the temperature of the wafer. This measurement may be based on a radiance signal to a pyrometer or an electrical signal from a thermocouple.

A novel model is now derived for control applications that describes the second block, or the relationship of lamp power to wafer temperature. Previous approaches to modeling wafer heating can be classified as detailed or high-order approaches [6, 7, 8, 9, 10]. In these approaches, a series of partial differential equations is written to describe wafer heating. The system of equations can then be expressed as a series of finite difference approximations and solved using numerical methods. These detailed modeling approaches employ numerous parameters (such as high-dimensional view-factor matrices), which

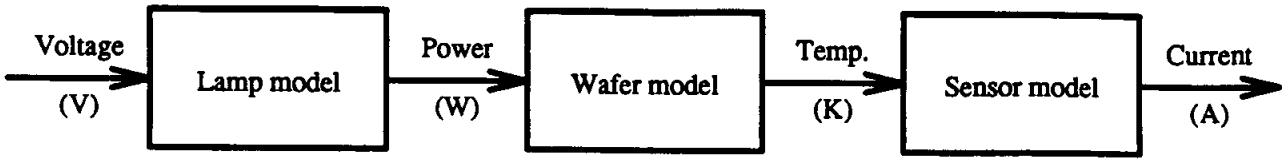


Figure 2: Block diagram of physical models that are derived to describe a rapid thermal processing system.

for an actual system with complex geometries are often not known accurately. Furthermore, the computational complexity of the detailed models generally preclude usage for prediction purposes in real-time applications. Consequently, while the detailed models are useful for qualitative statements about lamp and chamber design, their use for predictive applications in signal processing and process control design is limited.

We begin the development by starting with a complete model of wafer heating, that is one that considers radiative, conductive, and convective energy transfer mechanisms. In [11], it is shown that a detailed model of RTP can be given by

$$\dot{T} = -\mathbf{A}_{rad}T^4 - \mathbf{A}_{conv}(T - T_a) - \mathbf{A}_{cond}T + \mathbf{B}P \tag{1}$$

where T_a is the ambient temperature expressed as an $N \times 1$ vector. The heating effects due to the presence of a thick window are not taken into account because the heating mode is two orders of magnitude slower than the wafer and hence is not important for the predictive applications of interest as will be discussed later and shown by experimental data.

Radiative heat transfer is described by the full matrix $\mathbf{A}_{rad}$ given by

$$\mathbf{A}_{rad} = \mathbf{A}_l\mathbf{D} \tag{2}$$

where

$$\mathbf{A}_l = \mathbf{diag}\left[\frac{\epsilon(T_1)\sigma A_{s,1}}{m_1 C_p(T_1)}, \frac{\epsilon(T_2)\sigma A_{s,2}}{m_2 C_p(T_2)}, \ldots, \frac{\epsilon(T_N)\sigma A_{s,N}}{m_N C_p(T_N)}\right] \tag{3}$$

The mass, heat capacity and surface area of wafer element i are denoted by m_i, $C_p(T_i)$, and $A_{s,i}$, respectively. The total emissivity is denoted by $\epsilon(T_i)$, the Stefan-Boltzmann constant is denoted by σ, and the reflection matrix is denoted by the full matrix $\mathbf{D}$ (see [11] for a derivation of $\mathbf{D}$). We note that the temperature dependence of the emissivity and heat capacity are taken into account.

The convective heat transfer is described by the diagonal matrix:

$$\mathbf{A}_{conv} = \mathbf{diag}\left[\frac{h_1(T_1)A_{s,1}}{m_1 C_p(T_1)}, \frac{h_2(T_2)A_{s,2}}{m_2 C_p(T_2)}, \ldots, \frac{h_N(T_N)A_{s,N}}{m_N C_p(T_N)}\right] \tag{4}$$

where h_i denotes the convective heat transfer coefficient.

The term accounting for heat conduction, $\mathbf{A}_{cond}$, is represented by a tridiagonal matrix where

$$\begin{aligned}
\mathbf{A}_{cond}(i,i+1) &= \frac{k_c(T_i)A_{c,i}}{m_i C_p(T_i)(r_{c,i+1} - r_{c,i})}, && 1 \le i < N \\
\mathbf{A}_{cond}(i,i) &= \frac{k_c(T_{i-1})A_{c,i-1}}{m_i C_p(T_i)(r_{c,i} - r_{c,i-1})} - \frac{k_c(T_i)A_{c,i}}{m_i C_p(T_i)(r_{c,i+1} - r_{c,i})}, && 1 \le i \le N \\
\mathbf{A}_{cond}(i+1,i) &= \frac{k_c(T_{i-1})A_{c,i-1}}{m_i C_p(T_i)(r_{c,i} - r_{c,i-1})}, && 1 < i < N
\end{aligned}$$

where k_c is the thermal conduction coefficient and $A_{c,i}$ is the cross sectional area evaluated at the mean radius, r_c, of element i.

The radiation of energy from the lamps to the wafer is given through the full matrix,

$$\mathbf{B} = \mathbf{B}_l \mathbf{F} \tag{5}$$

where

$$\mathbf{B}_l = \mathbf{diag}\left[\frac{\epsilon(T_1)}{m_1 C_p(T_1)}, \frac{\epsilon(T_2)}{m_2 C_p(T_2)}, \ldots, \frac{\epsilon(T_N)}{m_N C_p(T_N)}\right] \tag{6}$$

The high-order model, (1), consists of a large number of parameters that need to be estimated, including the reflection matrix $\mathbf{D}$ in $\mathbf{A}_{rad}$, the view factor matrix $\mathbf{F}$ in $\mathbf{B}$, and the diagonal convection matrix $\mathbf{A}_{conv}$. These parameters are difficult to determine accurately *a priori* based on theoretical or ray-tracing considerations. The number of unknown parameters is at least $N \times N$ from $\mathbf{D}$, $N \times M$ from $\mathbf{F}$ and $N \times 1$ from $\mathbf{A}_{conv}$ which is equal to $N(N + M + 1)$. This number is large when one considers that for a four inch wafer with, say, ten annular zones and three lamps the number of unknowns is 140.

We now propose to reduce this high dimensional system representation, (1), to a lower order model that can be used for a computationally efficient prediction of the wafer dynamics for process control or signal processing applications. The basic assumption used to develop the low-order model is that the spatial distribution of temperature is reasonably uniform. This assumption is really a necessity for RTP to reach fruition and is the basic reason for large research efforts in lamp design and multipoint temperature measurement. Because the temperature will be roughly uniform in RTP applications, the thermal gradient across the wafer will be small enough that the contribution of energy into an annular zone from conduction will be much smaller than that from radiation and convection. Consequently, $\mathbf{A}_{cond}$ will be dropped from (1).

The term $\mathbf{A}_{rad}T^4$ in (1) accounting for radiative heat transfer for annular zone i can be written as

$$\beta_i \sum_{j=1}^{N} \mathbf{D}_{i,j} T_j^4 = \beta_i \sum_{j=1}^{N} \mathbf{D}_{i,j}(T_i^4 + 4\Delta_{i,j}T_i^3 + 6\Delta_{i,j}^2 T_i^2 + 4\Delta_{i,j}^3 T_i + \Delta_{i,j}^4) \tag{7}$$

where

$$\begin{aligned} \Delta_{i,j} &= T_j - T_i \\ \beta_i &= \frac{\epsilon(T_i)\sigma A_{s,i}}{m_i C_p(T_i)} \end{aligned}$$

Since $\Delta_{i,j}$ denotes the temperature difference from one annular zone to the next, it will be a small number relative to T_i, by our assumption of spatial temperature uniformity. In addition, we consider that the diagonal elements of $\mathbf{D}$ consist of two parts, emission and reflection components, while the off-diagonal elements only consist of reflection components. Consequently, the matrix is diagonally dominant. Based on these considerations, the radiation matrix can be approximated by the following diagonal matrix

$$\tilde{\mathbf{A}}_{rad} = \mathbf{diag}\left[\beta_1 \sum_{j=1}^{N} \mathbf{D}_{1,j}, \ldots, \beta_N \sum_{j=1}^{N} \mathbf{D}_{N,j}\right] \tag{8}$$

The error associated with this approximation is given for a particular wafer element i as

$$\delta_i = \frac{\sum_{j=1}^{N} \mathbf{D}_{i,j}(4\Delta_{i,j}T_i^3 + 6\Delta_{i,j}^2 T_i^2 + 4\Delta_{i,j}^3 T_i + \Delta_{i,j}^4)}{\sum_{j=1}^{N} \mathbf{D}_{i,j}(T_i^4 + 4\Delta_{i,j}T_i^3 + 6\Delta_{i,j}^2 T_i^2 + 4\Delta_{i,j}^3 T_i + \Delta_{i,j}^4)} \tag{9}$$

(It turns out that this error is generally not very large, for the an unlikely situation where $T_i = 1000$ K and $\Delta_{i,j} = 5$ K for all $j(j \neq i)$, the error is only about 2 percent.)

The resulting state description of the process is expressed as

$$\dot{T} = -\tilde{\mathbf{A}}_{rad}T^4 - \mathbf{A}_{conv}(T - T_a) + \mathbf{B}P \tag{10}$$

The reduction of dimensionality is substantial. The number of unknown parameters is now $N \times 1$ from $\tilde{\mathbf{A}}_{rad}$, $N \times 1$ from $\mathbf{A}_{conv}$, and $N \times M$ from $\mathbf{B}$. We may also note that by neglecting conductive effects and radiative transfer from one wafer element to another via reflections, the low-order model, (10), is made noninteracting with regard to one wafer element influencing another. Consequently the model can be reduced further by considering only those wafer elements of concern (*e.g.* where sensors are located). We denote the reduced number of elements by $\tilde{N}$. The number of unknown parameters is thus $\tilde{N}(2 + M)$. For a 3 sensor 3 lamp system this equals 15, compared to 140 for the high-order model. Furthermore, the estimation of these parameters can be accomplished through a small number of experiments. In addition, the computational time to solve such a system of equations is low and can be accomplished in real-time for predictive purposes.

2.2 Model Linearization

Linearization of the low-order model will further aid in analyzing the process dynamics and evaluating the assumptions used in the development of the low-order model. We consider a Taylor series expansion about a reference value of power and temperature, denoted by P_r and T_r respectively,

$$T_d = T - T_r \tag{11}$$

$$P_d = P - P_r \tag{12}$$

Taylor series expansion of (10) and truncation after the first terms yields

$$\dot{T}_d = -\mathbf{A}_{lin}(T_r)T_d + \mathbf{B}_{lin}(T_r)P_d \tag{13}$$

where

$$\mathbf{A}_{lin}(T_r) = \left.\frac{\partial \tilde{\mathbf{A}}_{rad}T^4}{\partial T}\right|_{T_r} + \left.\frac{\partial \mathbf{A}_{conv}(T - T_a)}{\partial T}\right|_{T_r} + \left.\frac{\partial \mathbf{B}\mathbf{P}_r}{\partial T}\right|_{T_r,P_r} \tag{14}$$

The solution to the above differential equation is given by

$$T_d(t) = e^{-\mathbf{A}_{lin}(T_r)t}T_d(0) + \int_0^t e^{-\mathbf{A}_{lin}(T_r)(t-\phi)}\mathbf{B}_{lin}(T_r)P_d(\phi)d\phi \tag{15}$$

where $T_d(0)$ is the initial condition.

If we consider the coefficients to be weakly dependent on temperature, the matrices in the linearized model can then be approximated as

$$\mathbf{A}_{lin}(T_r) = \mathbf{diag}\left[\frac{4\epsilon(T_{r,1})\sigma A_1 T_{r,1}^3 \sum_{j=1}^N \mathbf{D}_{1,j} + h_1(T_{r,1})A_1}{m_1 C_p(T_{r,1})}, \right.$$
$$\left. \cdots, \frac{4\epsilon(T_{r,N})\sigma A_{\tilde{N}} T_{r,\tilde{N}}^3 \sum_{j=1}^N \mathbf{D}_{\tilde{N},j} + h_{\tilde{N}}(T_{r,\tilde{N}})A_{\tilde{N}}}{m_{\tilde{N}} C_p(T_{r,\tilde{N}})}\right] \tag{16}$$

$$\mathbf{B}_{lin}(T_r) = \mathbf{diag}\left[\frac{\epsilon(T_{r,1})}{m_1 C_p(T_{r,1})}, \cdots, \frac{\epsilon(T_{r,\tilde{N}})}{m_{\tilde{N}} C_p(T_{r,\tilde{N}})}\right]\mathbf{F} \tag{17}$$

The eigenvalues characterize the dynamics of the system response. For this model, they are real and are given by the diagonal elements of $\mathbf{A}_{lin}(T_r)$,

$$\lambda_i = \frac{4\epsilon(T_{r,i})A_{s,i}T_{r,i}^3 \sum_{j=1}^{N} \mathbf{D}_{i,j} + h_i(T_{r,i})A_{s,i}}{m_i C_p(T_{r,i})}, \qquad i = 1, \cdots, \tilde{N} \tag{18}$$

For zones that do not encompass the wafer edge, the eigenvalue can also be expressed as

$$\lambda_i = \frac{8\epsilon(T_{r,i})T_{r,i}^3 \sum_{j=1}^{N} \mathbf{D}_{i,j} + 2h_i(T_{r,i})}{\rho_w d C_p(T_{r,i})}, \qquad i = 1, \cdots, \tilde{N} \tag{19}$$

where d is the thickness of the wafer. For the case where an annular zone of interest is at the edge (indicated by $\tilde{N}$), the eigenvalue at the edge is given by

$$\lambda_{edge} = \frac{8\epsilon(T_{r,\tilde{N}})T_{r,\tilde{N}}(1 + A_{s,e}/2A_{\tilde{N}}) \sum_{j=1}^{N} \mathbf{D}_{i,j} + 2h_{\tilde{N}}(1 + A_{s,e}/2A_{\tilde{N}})}{\rho_w d C_p(T_{r,\tilde{N}})} \tag{20}$$

where $A_{s,e}$ is the surface area of the edge.

In process control, the inverse of the eigenvalue is referred to as the time constant. For first order systems, the response time (*i.e.* time to settle to 98% of the final value) is approximately equal to four time constants. The large dependence on temperature is seen in (19). Also, note that the eigenvalue is independent of outside sources such as quartz window and the lamp. In addition, note that the eigenvalue λ_{edge} of the edge zone will be larger than that of adjacent zones because of the term $A_{s,e}$. The result is that the edge responds to system changes much faster than the rest of the wafer. Consequently, in uniform flux systems, the wafer edge will be hotter than the rest of the wafer during heating, and cooler than the rest of the wafer during cool-down.

The gain matrix describing the static effects from P_d to T_d is given by

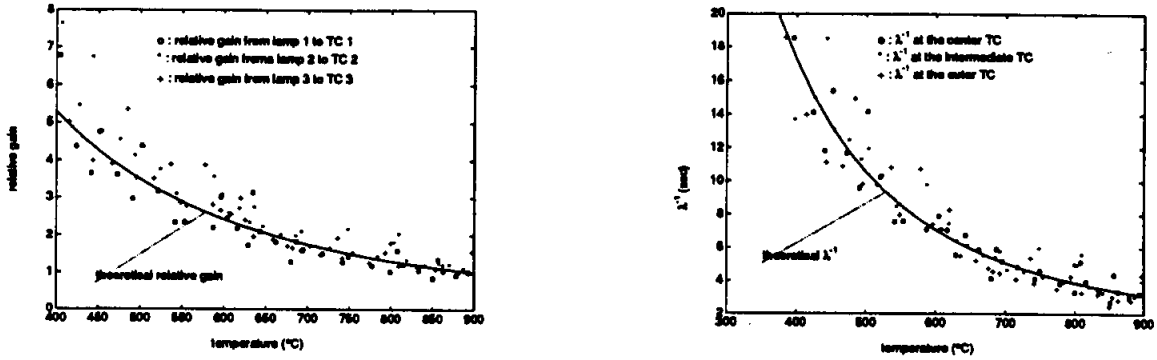
$$\mathbf{K} = \mathbf{diag}\left[\frac{\epsilon(T_{r,1})}{4\epsilon(T_{r,1})\sigma T_{r,1}^3 \sum_{j=1}^{N} \mathbf{D}_{1,j} + h_1}, \ \cdots, \ \frac{\epsilon(T_{r,\tilde{N}})}{4\epsilon(T_{r,\tilde{N}})\sigma T_{r,\tilde{N}}^3 \sum_{j=1}^{N} \mathbf{D}_{\tilde{N},j} + h_{\tilde{N}}} \right] \mathbf{F/A} \tag{21}$$

where the coefficients are evaluated at the reference temperature T_r and $\mathbf{F/A}$ denotes $\mathbf{F}$ with the elements of row i divided by A_i. As is the case for the eigenvalues, the gain is highly dependent on temperature. We will demonstrate these nonlinear effects in the next subsection by validating the low-order nonlinear model.

2.3 Low-Order Model Validation

The low-order model is validated by first studying its ability to characterize the temperature dependence of two useful measures of the transient behavior of RTP. These measures are the *gain* and *time constant* of the system. The gain (see (21)) describes the amount that the wafer temperature will change in response to a given change in lamp power after transients have subsided. The gain is an intrinsic measure of RTP and serves as a characteristic of the steady-state behavior of RTP. In (21), the linearized form of the low-order model predicts that the gain is proportional to $1/T^3$, where T denotes absolute wafer temperature. (The actual numerical value of the gain is a function of view factors and wafer thermal parameters as derived above.) Experiments were conducted at 1 torr pressure to evaluate this prediction by making independent step changes to each of the three lamps of the Stanford RTM

and then observing the change in temperature of three thermocouples bonded at 0, 1, and 1.75 inch radii of a four-inch wafer. These locations will be denoted as the center, intermediate, and outer thermocouple positions. The gain is the ratio of change in wafer temperature (*i.e.* from one steady-state condition to another) to change in lamp power. Because we have three lamp zones (inputs) and three thermocouple measurements (outputs), there will actually be nine gains which can be arranged in a 3 $\times$ 3 matrix. The three diagonal terms of this matrix will be comprised of the gains associated with, 1) center lamp power to center thermocouple measurement, 2) intermediate lamp power to intermediate thermocouple measurement, 3) outer lamp power to outer thermocouple measurement. In Figure 3 (a), the comparison between the $1/T^3$ prediction and experimental data for these diagonal terms is presented. The gains are normalized to their values at 900°C so that various parameters (view factors) involved in the prediction of the numerical value of the gain will cancel and the $1/T^3$ trend can be studied. Close agreement between the $1/T^3$ prediction of the low-order model and the data is seen. The strong nonlinear effect of temperature on the gain is also observed. This effect must be considered when designing wafer temperature control systems.



(a) (b)

Figure 3: Model validation at 1 torr, 100 sccm N_2 by examination of: (**a**) the gain (relative to its value at 900 °C) from a lamp to a thermocouple as a function of temperature and, (**b**) the time constant predicted by theory and fit to experimental data.

The time constant (see (19)) is a useful measure to characterize the dynamic response of wafer temperature to changes in lamp power. The low-order model predicts that the time constant of the wafer temperature in vacuum is approximately equal to $(\rho d C_p)/(8\epsilon\sigma T^3)$. In Figure 3 (**b**), the time constant of the wafer (denoted by λ^{-1}) predicted by the low-order model is compared against an estimated quantity using experimental data at 1 torr pressure. The estimated quantity is determined by using a signal processing method, ESPRIT, to extract the most significant time constants from the data and then the dominant time constant was selected from them [11]. The data is in close agreement with the theoretical prediction of the low-order model over the range from 400°C to 900°C. The temperature dependence of time constant is apparent from this figure. The independence of the time constant with respect to wafer position can also be seen and is predicted by the low-order model at low pressures when convective heat transfer effects are negligible. We might also note that the time constant (and gain) of the wafer is independent of whether the wafer is being heated or cooled. The reason is simply that the gain and time constant of the wafer are independent of external factors such as the lamps, chamber, or window,

which act as forcing or input functions that actually drive wafer temperature deviations.

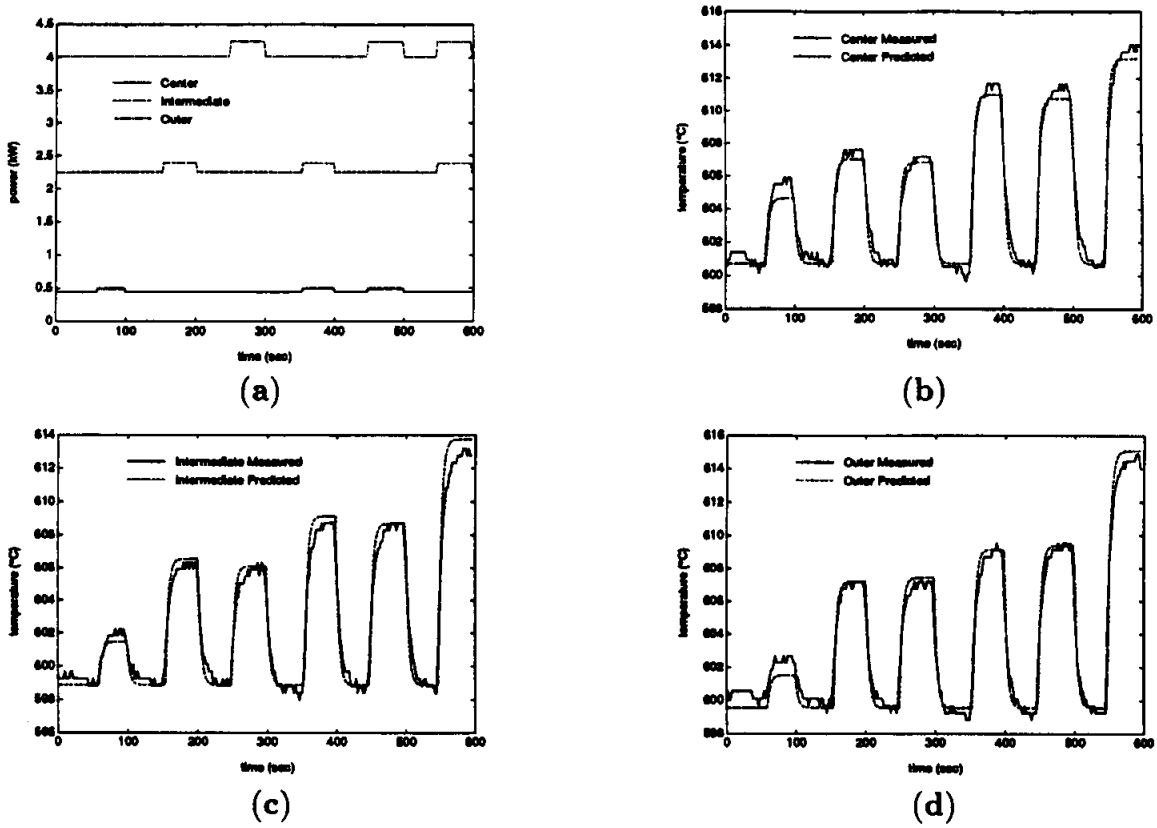


Figure 4: Model validation with (**a**) lamp power inputs and comparison between low-order model prediction and measurement about 600°C and 1 torr for (**b**) center, (**c**) intermediate, and (**d**) outer thermocouples.

To complete the validation of the low-order model, we investigate its input-output predictive capabilities. This is accomplished by comparing the response of the system to that predicted by the low-order model. The data employed in this study is not used in estimating the various low-order model coefficients. This verification procedure is known as cross-validation [12]. In Figure 4 (**a**), the pulsed sequence of input powers to the three lamp rings at 1 torr pressure are shown as a function of time. The corresponding responses of the center, intermediate, and outer thermocouples are shown in Figure 4 (**b**) — (**d**), respectively. The temperatures are close to 600°C. The bit resolution of the thermocouple sensors is evident by the roughly 0.5°C jumps in measurement signal. The predictions using the low-order model are also shown. The prediction is in agreement with the system response. Some deviation is to be expected since the low-order model does not include the time-constant of the heating element. For tungsten-halogen lamps, this time constant is roughly 200 milliseconds. The prediction is not effected by chamber and window heating because the system started in a thermally steady-state condition. Addition of these slow mode heating effects to the wafer transients could be included as an external source radiating energy to the wafer.

Because radiative heat transfer is proportional to T^4, the dynamics and steady-state characteristics significantly change with temperature. The ability of the model to capture the nonlinear effects in a predictive capacity is further presented in [11]. The experiments consisted of using the same low-order model (*i.e.* updating the coefficients with temperature according to the physical predictions derived

above) to predict the response of the wafer at other temperatures. The low-order model was shown to closely predict the steady-state and transient behavior of RTP under a variety of conditions.

These predictive results, combined with the structural validation plots of Figures 3 (a) and (b) demonstrate that the proposed nonlinear low-order model can be used to well approximate RTP over a range of operating conditions. Consequently, it is then only necessary to estimate the parameters of the low-order model about a few temperatures (only one temperature is actually necessary at 1 torr pressure) to predict the system behavior at other temperatures.

At higher operating pressures, convective heat transfer will play a role and needs to be considered in the modeling. In this case, the time constant of the wafer is given by $(\rho d C_p)/(8\epsilon\sigma T^3 + 2h)$ where h is the convective heat transfer coefficient. This coefficient accounts for variations in gas pressure, gas composition and flow rate. It needs to be estimated from data for each set of processing conditions since accurate theoretical predictions of this quantity are not available. Details of the necessary estimation procedure and model validating experiments at 1 atmosphere pressure are contained in [11].

3 CONTROL

The objective of an automatic temperature control system for RTP is to achieve spatial temperature uniformity while tracking prespecified trajectories. Because of the precise temperature control needed for RTP, successful control systems have to be able to handle several process complexities including limitations of achievable actuation, nonlinear heat transfer characteristics due to radiation, thermal memory of the system, apparent time-delays in the actuators and sensors and the multivariable coupling of lamps and sensors. In order to meet these requirements, sufficient control authority must exist. Control authority, for this case, refers to the ability to generate a wide range of energy flux distributions to the wafer. This range is necessary because of the wide variety of operating conditions subjected to the wafer with the constraint that the objective of wafer temperature uniformity be achieved. In this section, design issues are first discussed. A multivariable control strategy is then described and experimental evaluation is presented.

3.1 Wafer Temperature Controllability

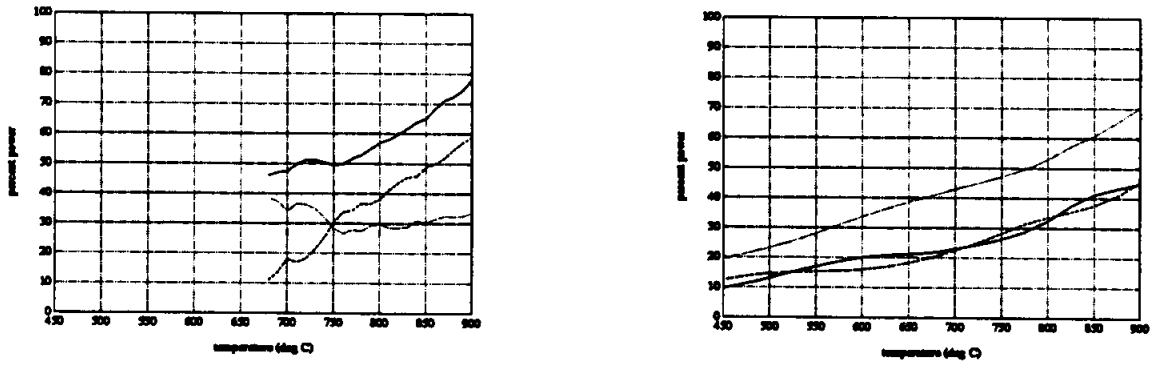
Original rapid thermal processing systems used a single power supply for wafer heating. This power supply was connected to either a single arc lamp or to a bank of linear lamps. It was found that slip-free processing was difficult to obtain because of severe temperature nonuniformities at high temperatures. It was shown in [3, 4, 10, 13] that a multivariable lamp actuation (or power supply) system could achieve slip-free processing over a range of operating conditions. In a multivariable system, the energy flux to the semiconductor wafer could be varied dynamically to account for the inherent nonlinearities of radiative heating.

A cross-sectional schematic of the multiple-actuation Stanford RTM is shown in Figure 1. The outer ring of lamps consists of twenty-four one-kilowatt bulbs. The intermediate ring of lamps consists of twelve one-kilowatt bulbs. The central lamp is a two-kilowatt bulb. Three power supplies are available to control the power to each lamp ring. The reflector head is water-cooled and the surface is gold-plated. The window is made of quartz. The wafer is four inches in diameter and supported by pins.

An annular gold-plated stainless steel baffle was added to improve the controllability (or the ability to generate a wider variety of energy flux profiles to the wafer). Prior to the addition of the baffle, studies indicated that the intermediate and outer lamp rings were largely coupled. The energy differential between the intermediate and outer zones had little influence on the wafer temperature profile. For the

case where no baffle is present, Figure 5 (a) presents the percentage power to each of the three zones that achieves equal temperature readings on three thermocouples bonded along a common wafer radius (at radii positions 0, 1, and 1.75 inches). As can be seen, it was impossible to achieve equal temperature readings at all three locations at temperatures below 650°C since the intermediate zone would saturate to zero. This saturation in effect was a loss in controllability and degrees of freedom. The three lamp heating system essentially operated as a two lamp heating system.

However, the baffle provided a twofold effect of attenuating direct radiation from 1) the intermediate lamps to the outer portion of the wafer and 2) the outer lamps to the intermediate portion of the wafer. Consequently, the baffle allowed the outer lamps to provide a higher degree of wafer edge heating while minimizing the contribution of energy to the interior locations of the wafer. Because of the increased surface area at the wafer edge, less energy flux is required than the rest of the wafer during heating. In Figure 5 (b), the improvement in the range of achievable uniformity is demonstrated for the case with the baffle present. The range of achievable uniformity is increased which can be seen by comparison to Figure 5 (a). Identification studies also indicated an improvement to the weakest direction of control (as measured by the minimum singular value of the gain matrix) by a factor of five.



(a) (b)

Figure 5: Powers of the three-zone lamps to achieve equal temperature readings at the three sensor locations for the cases when (a) no baffle was present and, (b) the baffle was added.

3.2 Multivariable Control

A multivariable control system is developed to automatically manipulate the power to each of the N (in our case, three) rings of lamps to control N (three) temperature sensors on the wafer. It may be advantageous to use more sensors than available lamps. However, for a specific simulated RTP system and operating conditions, Norman [10] has shown that little is to be gained over the $N \times N$ configuration. For this case, it was also shown that the placement of the sensors is an important consideration.

Because the desired temperature trajectory is known *a priori*, the appropriate control strategy to use for this problem combines feedback and feedforward mechanisms and employs gain-scheduling to handle the nonlinearities. A block diagram of the strategy is shown in Figure 6. The feedforward mechanism is used as a model-based prediction to get the wafer temperature close to the desired trajectory. The feedback mechanism is used to compensate for modeling errors and disturbances. Gain-scheduling is employed to adjust the feedforward and feedback control parameters to compensate for the nonlinearities.

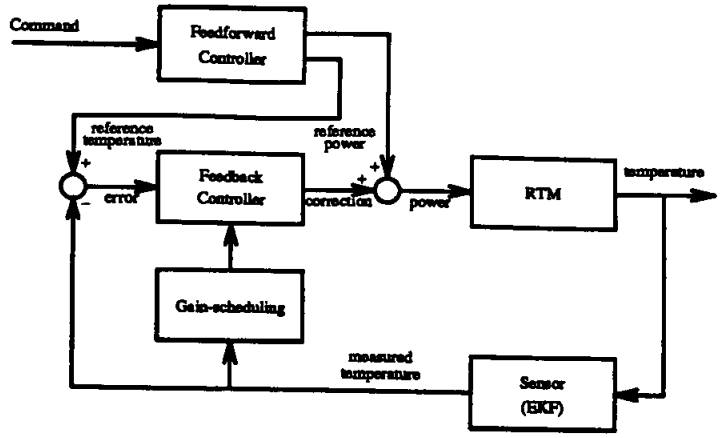


Figure 6: Block diagram of the control strategy used to control a multi-zone lamp, multi-point sensor RTP system. The strategy employs a feedforward loop for model-based prediction and a feedback loop to compensate for errors in the prediction. Gain-scheduling is employed to compensate for the nonlinear characteristics of wafer heating induced by radiation energy transfer phenomenon. The strategy includes an extended Kalman filter (EKF) to filter measurement noise.

A detailed derivation of the feedforward, feedback, and gain-scheduling control strategy that is used for the following results is presented in [14]. Presently, a summary of the controller algorithm is given. The feedforward action consists of two terms. The first term describes the lamp power needed to hold a uniform wafer temperature at steady-state. This term employs experimental results in relating steady-state temperatures to the corresponding steady-state powers. The second term provides the incremental lamp power needed to set the wafer temperature along a desired trajectory. This term employs the nonlinear physics-based model derived above. The feedback mechanism is based on a linearized version of the low-order model combined with models of the lamps and sensors. The feedback controller is designed using a model matching criterion. That is, the controller is designed so that disturbances are eliminated in a specified duration of time. The duration is specified to balance between speed of response and sensitivity to noise. A key feature of the feedback controller is the use of multiple integrators to guarantee zero-error between wafer and desired temperature despite modeling errors. The controller parameters are a function of the physics-based model parameters and temperature dependent functions. Consequently, a gain-scheduling (or temperature-scheduling) of the controller parameters is used to adapt for the nonlinear radiation effects and to achieve better control.

3.3 Experimental Results

The multivariable, gain-scheduled, feedforward/feedback control strategy was applied to the Stanford Rapid Thermal Multiprocessor (RTM). The schematic of the lamp heating system is shown in Figure 1. Three thermocouples were bonded to a four-inch diameter wafer at radii locations of 0, 1, and 1.75 inches.

The objective of the control experiment was the following: *Starting from room temperature, ramp to 900 °C at a rate of 45 °C/second and then hold at 900 °C for five minutes. Conduct the experiment at one atmosphere pressure with 100 sccm N_2 flow rate.* This specification corresponds to a thermal oxidation step where oxygen is used instead of nitrogen. Oxygen was not used because it would damage

the thermocouple wafer. Thermocouples were used since accurate noninvasive sensors were not yet available. A 45°C/second ramp was selected because earlier studies indicated that it was the maximum achievable ramping rate with temperature uniformity for the present control authority of the RTM. The differential energy flux provided by three lamp zones was not sufficient to achieve a higher ramping rate and still retain temperature uniformity.

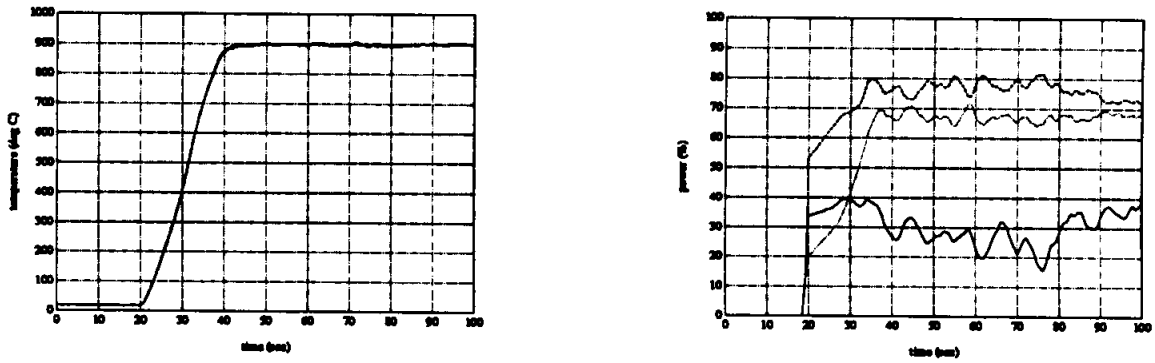
The following procedure was used to achieve the control objective:

- The gain-scheduled, integral action, feedback controller was used to obtain the relation of steady-state uniform temperature (T_s) to steady-state powers (P_s) over the temperature range from 20°C to 950°C at 50°C increments. It was important to generate this relation by:

  1. First, the feedback controller was used alone to exponentially rise from room temperature to the desired temperature. The feedback controller was designed so that the time to settle at the desired temperature was equal to the total temperature change divided by the desired ramp rate.
  2. Next, the wafer was returned returned to room temperature.

  This two step-procedure was then repeated until the range of temperatures at the specified increment was covered. This approach was important since the feedforward powers necessary to achieve a desired temperature were representative of the conditions under which the wafer was to be processed. That is, the system (*i.e.* quartz window, walls) was in a transient slow-heating condition and was not allowed to come to thermal equilibrium. The difference in necessary power to hold wafer temperature uniform was substantially dependent on the system temperatures.

- The relation of T_s to P_s and the physics-based model was then used to obtain the feedforward trajectory. The feedback controller was wrapped around this feedforward trajectory with a specified closed-loop time constant specified as 1.5 seconds.

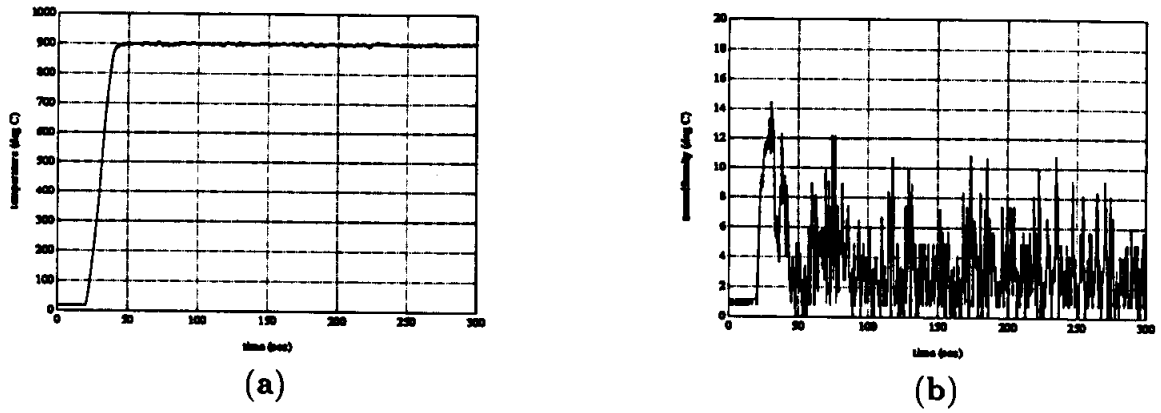


(a) (b)

Figure 7: Controlled temperature trajectory over the first 100 seconds of the desired ramp and hold trajectory: (a) Temperature response of the three thermocouples and, (b) Percent powers to the three-zone lamp heating system to achieve the desired temperature response.

The controlled ramp for the first 100 seconds is examined in Figure 7 (**a**) where the three thermocouple measurements are shown for the ramp and hold. The corresponding powers are shown in Figure 7 (**b**). The sensitivity of the controller to the sensor noise can be seen in this result. This sensitivity can easily be reduced by adding a filter to the sensor. It is interesting to note the time delay of the system by comparing the powers to the temperature response as the ramp started. Approximately a two second total delay existed in the beginning of the response. Of this delay, 1.5 seconds is due to a power surge protection scheme on the lamps which subsides after the percentage power to the lamps surpasses fifteen. A 0.5 second delay remains and is due to the thermocouples and lamps. This delay is very substantial when one considers a desired ramp rate of 45°C/second. This time delay prevented a smaller value of the closed-loop dynamics without employing special compensation procedures such as Smith or GAP predictors. In addition, from the power plot of Figure 7 (**b**), the rate limiting of the lamps is also seen. These rate-limits are set as a safety protection to prevent a large in-rush current.

The nonuniformity of the controller throughout the entire ramp and hold trajectory is studied. In Figure 8 (**a**), the temperature corresponding to the three thermocouple sensors is plotted for the five minute trajectory. In Figure 8 (**b**) the maximum temperature difference as measured by subtracting the minimum of the three thermocouple readings from the maximum is plotted as a function of time. The maximum nonuniformity during the ramp is approximately 15°C. The mean temperature corresponding to this error is approximately 350°C. This mean temperature is low and the nonuniformity will not effect processing nor damage the wafer. The substantial sensor noise of the thermocouples can also be seen when the mean temperature is 900°C.

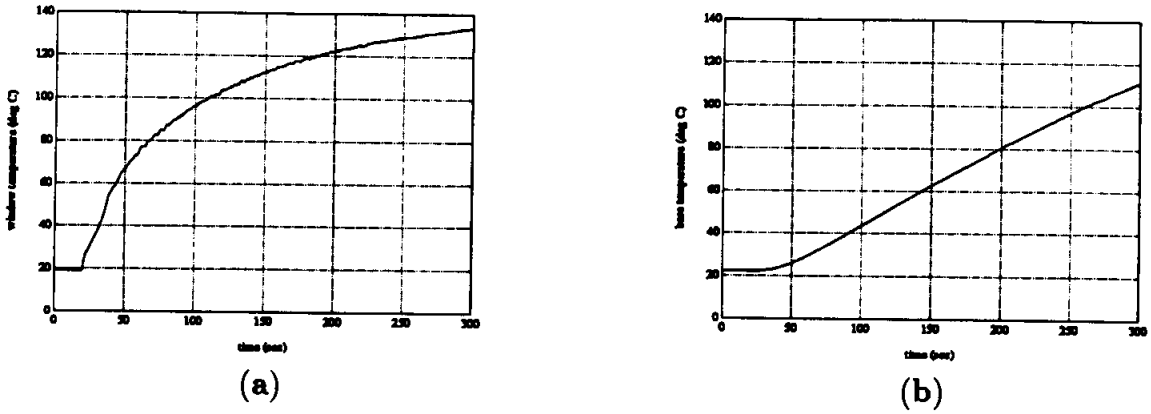


(**a**) (**b**)

Figure 8: Automatic control for the five minute ramp and hold trajectory: (**a**) the temperature as measured by three thermocouples and, (**b**) the temperature nonuniformity across the wafer as measured by the difference between the maximum and minimum readings of three thermocouples.

The capability of the controller to hold the wafer temperature at a desired processing temperature despite slow heating modes is examined. As seen in Figure 8 (**a**), the multivariable controller was able to hold the temperature about the desired value of 900°C. Although the sensors were quite noisy in the atmospheric environment and had resolution limited to approximately 0.5°C, the average temperature during the hold portion of the ramp for the three sensors was, 900.9°C, 900.7°C and 900.8°C. This result is important since the average temperature is nearly uniform and will result in uniform processing. The reason for the temperatures being higher that the targeted 900°C is explained by examining the slow

heating modes of the chamber walls and quartz window. The temperature of the quartz window and chamber base of the RTM are plotted in Figure 9 (a) and (b), respectively. This slow heating of the components of the RTM act as slow disturbances to the wafer. Because the gain of the controller was reduced to compensate for time delay, a slightly higher average temperature than the desired 900°C was achieved. However, without a feedback control with integral action, the wafer temperature would have drifted substantially to greater than 950°C.



(a) (b)

Figure 9: Temperature of RTM system components: (a) the quartz window as measured by a thermocouple mounted to the edge of the window on the nonprocessing side and, (b) the base as measured by a thermocouple mounted on the nonprocessing side of the chamber.

In summary, the controller met the objective and performed satisfactory in the presence of numerous challenges including substantial time delays, saturating actuators, sensor noise, system nonlinearities, and slow disturbances. Extensions to the control strategy include an adaptation feature to fine-tune the nonlinear physics-based model using data from a cycle. The controllers are then fine-tuned from the adapted model.

4 CONCLUSIONS

A physics-based nonlinear model of wafer heating has been derived and validated experimentally for a range of temperatures from 400°C to 900°C and pressures of 1 torr in an inert N_2 environment. Results for one atmosphere pressure are contained in [11]. The advantage of the low-order model is ease of identification and application for real-time prediction purposes in signal processing and wafer temperature control. The nonlinear effects of temperature have been demonstrated in validating the model. A controller for a multi-zone lamp, multi-point sensor has been described and demonstrated. A controlled ramp was achieved from 20°C to 900°C at a rate of 45°C/second with less than 15°C nonuniformity during the ramp and less than 1°C average nonuniformity during the hold.

ACKNOWLEDGEMENTS

We thank Len Booth for help with the experimental work. This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under contract F49620-90-C-0014.

REFERENCES

[1] M. Moslehi, Single-wafer optical processing of semiconductors: Thin insulator growth for integrated electronic device applications, *Appl. Phys. A*, **46**:255–273, 1988.

[2] K. Saraswat, Center for Research on Manufacturing Science and Technology for VLSI, Annual report, Center for Integrated Systems, Stanford, CA, 1989.

[3] P. Apte and K. Saraswat, Rapid thermal processing uniformity using multivariable control of a circularly symmetric three zone lamp, submitted, 1991.

[4] S. Norman, C. Schaper, and S. Boyd, 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, April 1991.

[5] Y. Cho, C. Schaper, and T. Kailath, *In-Situ* temperature estimation in rapid thermal processing systems using extended Kalman Filtering, In *Mater. Res. Soc. Proc.: Rapid Thermal and Integrated Processing.* Materials Research Society, April 1991.

[6] H. Lord, Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven, *IEEE Trans. Semicond. Manufact.*, **1**(3):105–114, August 1988.

[7] R. Kakoschke, E. Bußmann, and H. Föll, Modelling of wafer heating during rapid thermal processing, *Appl. Phys. A*, **50**(2):141–150, February 1990.

[8] S. Campbell, K. Ahn, K. Knutson, B. Liu, and J. Leighton, Steady-state thermal uniformity and gas flow patterns in a rapid thermal processing chamber, *IEEE Trans. Semicond. Manufact.*, **4**(1):14–20, February 1991.

[9] R. Gyurcsik, T. Riley, and F. Sorrell, A model for rapid thermal processing: Achieving uniformity through lamp control, *IEEE Trans. Semicond. Manufact.*, **4**(1):9–13, February 1991.

[10] S. Norman, RTP control system analysis using convex optimization, Technical report, Stanford Univ. Information Systems Lab., 1992.

[11] C. Schaper, Y. Cho, and T. Kailath, Low-order modeling and dynamic characterization of rapid thermal processing, *to appear in Applied Physics A*, 1992.

[12] L. Ljung, *System Identification : Theory for the User*, Prentice Hall,Inc., Englewood Cliffs, New Jersey, 1987.

[13] S. Norman, Optimization of transient temperature uniformity in RTP systems, *IEEE Trans. Electron Devices*, **39**:205–207, 1992.

[14] C. Schaper, P. Park, and T. Kailath, Control of a multi-zone lamp, multi-point sensor RTP system, Annual report, Stanford University, Stanford, CA, 1992.