Optimization of Wafer Temperature Uniformity in Rapid Thermal Processing Systems

Stephen A. Norman*†

Information Systems Laboratory
Dept. of Electrical Engineering
Stanford University, Stanford CA 94305

Submitted to *IEEE Transactions on Electron Devices*
June 3, 1991

---

*PHONE: (415) 725-5698; FAX: (415) 723-8473;
E-MAIL: norman@isl.stanford.edu.

†This research 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 No. F49620-90-C-0014.

ABSTRACT

During rapid thermal processing (RTP) of semiconductor wafers it is essential that the wafer temperature closely tracks a pre-specified temperature trajectory and that the temperature profile across the wafer is nearly uniform whenever the wafer is at high temperatures. This paper considers axisymmetric RTP systems with multiple independently-controllable lamps. In such systems the relative lamp power settings can be adjusted so that the lamps provide a range of distributions of power over the wafer; this is useful because the ideal distribution of incident energy varies according to processing conditions and changes substantially during a single process cycle.

In this paper numerical techniques are given for finding lamp power settings to minimize temperature error across the wafer both during steady-state hold and during transients; these numerical techniques involve the formulation of the minimum temperature error problem as a linear programming problem. A heuristic scheme for near-optimal trajectory following is presented. Open-loop-plus-feedback control is proposed for RTP temperature control — open-loop control allows precise trajectory planning and feedback prevents drift due to disturbances and modeling errors. All of the above ideas are demonstrated in simulation on a fictional system qualitatively similar in its behavior to some real RTP systems.

1 Introduction

1.1 The RTP temperature control problem

Rapid thermal processing (RTP) is a category of techniques proposed for semiconductor fabrication processes such as annealing, oxidation, chemical vapor deposition and others [SMG+89, RP90]. In RTP, wafers are processed one at a time in a small, cold-walled reaction chamber. In tens of seconds or a few minutes, a wafer is heated from room temperature to as high as 1200°C, processed, and cooled. Wafers are heated by means of visible and/or infrared radiation supplied by one or more powerful lamps separated from the chamber by a transparent window.

Precise control of wafer temperature during RTP is essential. There are two main issues: uniformity and accurate trajectory following. Uniformity of temperature across the wafer at all times ensures uniform processing and prevents slip defects that result from excessive thermal stress. Accurate trajectory following is necessary to ensure process repeatability from one wafer to the next; in general it is desirable that the ramp-up and cool-down portions of temperature trajectories be short in order to minimize dopant redistribution during processing and to maximize system throughput. Numerous recent papers have discussed various aspects of the RTP temperature control problem. A by-no-means-exhaustive list is [HJB88, Lor88, KB89, VMDK89, RP90, YM90, GRS91, CAK+91, KCL91, MKY91, Kak91].

As many authors have noted, near-perfect wafer temperature uniformity requires that the radiative energy per unit area supplied by the lamps at all points on the wafer closely match a profile determined by the desired rate of wafer temperature change and by the distribution of radiative and convective heat loss across the wafer. It is well known that this profile varies in shape with changes in the processing environment. In particular the ideal lamp radiation profile for steady-state hold differs radically from that for fast ramp-up. Dynamic adjustment of the lamp radiation profile is therefore necessary to achieve temperature uniformity throughout an RTP process cycle. The need for such dynamic adjustment is pointed out in various papers [KB89, YM90, CAK+91, Kak91], but so far no practical methods have been presented for implementation of this dynamic adjustment. This paper will show how an RTP system with multiple, independently-controllable lamps can be used to ensure near-perfect wafer temperature uniformity and temperature trajectory following.

1.2 Scalar Control and Multivariable Control

Control engineers use the term *multivariable system* to describe a system with multiple controlled inputs and/or multiple sensed outputs; *multivariable control*, then, is control of such systems. In this paper the term *scalar system* will be used for a system that has only one controlled input and at most one sensed output; *scalar control* thus means control of scalar systems.

Controlling wafer temperature in an RTP system can be either a scalar control problem or a multivariable control problem, depending on the system. If the system has one big lamp and one sensor — typically a pyrometer focussed on a spot near the center of the wafer — then the control problem is scalar. On the other hand, if the system has multiple, independently controllable heat sources and/or multiple sensors, then then the control problem is multivariable. The sensor signals in a multivariable RTP temperature control scheme would typically be measurements of temperature at different points on the wafer.[1] Examples of RTP systems that are multivariable from the temperature control viewpoint are the rapid thermal multiprocessor (RTM) at Stanford [AWSM91] and a similar system at Texas Instruments [MKY91], both of which have concentric, circular, independently controllable arrays of halogen lamps; there are many other examples in existence.

It is possible to turn a multivariable RTP system into a scalar one by fixing the ratio between the powers supplied to the lamps. This is illustrated for a three-lamp system in Figure 1. In this diagram $P_j(t)$ is the power supplied to lamp j at time t and $P_j^{\max}$ is the maximum power that lamp j can radiate; the power supply inputs are assumed to be real numbers between 0 and 1. In the scalar case the power ratio is fixed by the constants g_1, g_2, and g_3, all between 0 and 1; only one controlled input, $u(t)$, is driven, either by the process engineer or by an automatic controller. In the multivariable case three such inputs, $u_1(t)$, $u_2(t)$, and $u_3(t)$, are independently driven; no g_j's are necessary.

This conversion of multi-lamp RTP systems into what are effectively single-lamp systems is a common practice for engineers trying to design RTP temperature control systems. Some criterion is used to select the coefficients g_1, g_2 and g_3 and then a scalar control system is designed in the hope that the performance penalty paid for surrendering flexibility will be small. One of the aims of this paper is to demonstrate that for RTP temperature control that this performance penalty is unacceptably large, that is, that the flexibility offered by multivariable control is necessary to obtain adequate temperature uniformity.

1.3 Outline of this paper

Section 2 introduces a generic thermal model for axisymmetric RTP systems; this model will be used throughout the rest of the paper.

Section 3 introduces a model RTP system called System S that will be used for simulation studies in Sections 4, 5 and 6.

[1]Sensor signals might also include measurements of other system variables pertaining to wafer temperature regulation — examples of such variables are lamp current, temperatures at various points in the RTP chamber, wafer surface conditions (*e.g.*, thickness of deposited layers), gas pressure, gas composition, etc. — if it turns out the system performance improvement obtained by use of these measurements warrants the added complexity and expense.

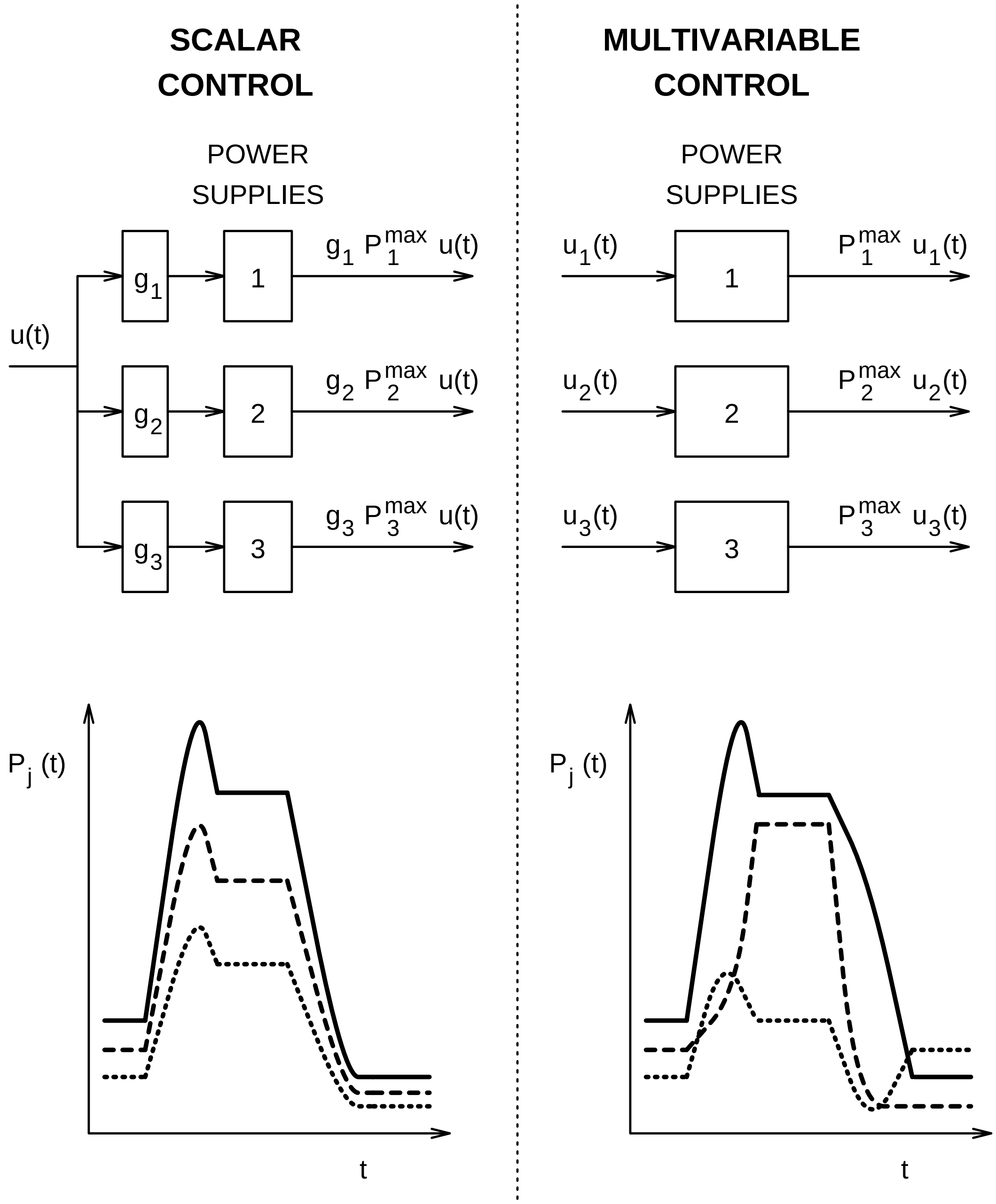


Figure 1: Scalar control and multivariable control for multi-lamp RTP systems.

Section 4 discusses methods for finding RTP lamp settings to optimize wafer temperature uniformity during steady-state hold.

Section 5 discusses two methods for designing lamp power trajectories to minimize error in following wafer temperature trajectories, one method being a simple heuristic and the other a more complicated but exactly optimal technique.

Section 6 introduces the idea of complementing open-loop control schemes based on ideas in Sections 4 and 5 with simple feedback control; application of feedback will greatly alleviate problems due to disturbances and modeling error.

Section 7 contains some discussion of the practical significance of the ideas presented in the paper. It also briefly describes some simple extensions of these ideas.

Section 8 makes some concluding remarks.

2 A Model for Axisymmetric RTP Systems

This section presents and discusses a generic model for axisymmetric RTP systems that is essentially an extension of a model presented by Lord [Lor88]. The primary reason for the appearance of this section is to clearly state assumptions and introduce notation that will be used in the ensuing sections. The model derived here will be an ordinary differential equation (ODE) approximation to the heat equation in cylindrical co-ordinates, which is a partial differential equation (PDE).

2.1 PDE wafer heat transfer model

Spatial distribution of temperature and other quantities in a silicon wafer is most naturally expressed in a cylindrical coordinate system. The wafer is assumed to be perfectly cylindrical with radius R and thickness Z. The z-axis of the coordinate system coincides with the central axis of the wafer; $z = 0$ on the bottom surface of the wafer and $z = Z$ on the top. Radial position is denoted by r and azimuth by θ. The dimensions R and Z are assumed to be constant, that is, thermal expansion of the wafer is neglected.

Expressed in terms of this cylindrical coordinate system, the PDE of heat conduction in the wafer is

$$\frac{1}{r}\frac{\partial}{\partial r}\left(kr\frac{\partial T}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \theta}\left(k\frac{\partial T}{\partial \theta}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) = \rho C_P \frac{\partial T}{\partial t}, \qquad (1)$$

where k is thermal conductivity, ρ, mass density, and C_P, specific heat. The values of k and C_P are assumed to depend on temperature but not on time or position; if thermal expansion is neglected, the value of ρ is fixed. Boundary conditions are:

$$k\frac{\partial T}{\partial r} = q_{\mathrm{edge}}(\theta, z) \quad \text{at } r = R;$$

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

the quantities q_{edge}, q_{bottom}, and q_{top} are heat flow per unit area into the wafer on the wafer edge, bottom and top; at a particular point on the wafer, these quantities depend on the values of the following quantities at that point: wafer surface temperature, incident thermal radiation, and radiative and convective heat transfer coefficients.

If temperature is independent of θ and the dependence of temperature on z is neglected, (1) and its associated boundary conditions can be re-written as the PDE

$$
\frac{1}{r}\frac{\partial}{\partial r}\left(kr\frac{\partial T}{\partial r}\right) + \frac{q_{\text{bottom}}(r) + q_{\text{top}}(r)}{Z} = \rho C_P \frac{\partial T}{\partial t}, \tag{2}
$$

with boundary conditions

$$
\begin{aligned}
\frac{\partial T}{\partial r} &= 0 \quad \text{at } r=0, \quad \text{and} \\
k\frac{\partial T}{\partial r} &= q_{\text{edge}} \quad \text{at } r=R.
\end{aligned}
$$

The assumption that temperature is θ-independent is fair to make if there are no large-scale, non-axisymmetric patterns on the wafer surface, since considerable pains are taken to ensure effective axisymmetry in RTP system design. As shown in [VMDK89], some large-scale rectangular patterns can have serious non-axisymmetric effects on the wafer temperature profile; wafer temperature control in such cases is beyond the scope of this paper.

The assumption that temperature is z-independent will be used in this paper, but it should be pointed out that this assumption is not completely correct. It has been noted [Lor88, CAK$^+$91] that if the lamps heat the wafer from one side only the temperature difference between the illuminated side of the wafer and other side can be as much as 4°C.[2] It is generally assumed that this axial nonuniformity does not contribute significantly to thermal stress since the wafer is free to deform in response to axial temperature nonuniformity.

[2] If absorbed radiation and convective and radiative loss are nonuniform across the top and bottom surfaces this axial temperature difference may vary in size with radial position. So if temperature is perfectly uniform on the wafer backside, nonuniformity on the frontside could be on the order of 0.5°C. This nonuniformity is too small to matter in RTP applications, but it does mean that in Section 4 simulation results that show apparent temperature nonuniformities considerably smaller than 0.5°C should be interpreted with caution.

2.2 A finite difference approximation to the heat equation

The quantities q_{bottom}, q_{top} and q_{edge} in (2) include radiative and convective loss from the wafer and radiative absorption by the wafer of lamp power. Radiative loss is a nonlinear function of wafer temperature and lamp power can have an arbitrary distribution across the wafer, so solution of (2) is most easily accomplished numerically. The approach used here is to break the wafer into I concentric elements, one cylindrical and the others annular. Let r_i^{out} be the outer radius of element i:

$$0 < r_1^{\text{out}} < r_2^{\text{out}} < \ldots < r_I^{\text{out}} = R.$$

For convenience let r_i^{in} (element inner radius) and r_i^{cen} (element "center" radius) be defined thus:

$$r_i^{\text{in}} = \begin{cases} 0 & \text{if } i = 1 \\ r_{i-1}^{\text{out}} & \text{otherwise} \end{cases}$$

and

$$r_i^{\text{cen}} = \left(\frac{(r_i^{\text{in}})^2 + (r_i^{\text{out}})^2}{2} \right)^{\frac{1}{2}}.$$

Then for element i the mass m_i is given by

$$m_i = \rho\pi \left((r_i^{\text{out}})^2 - (r_i^{\text{in}})^2 \right) Z,$$

where ρ is the mass density, and the exposed area A_i is given by

$$A_i = \begin{cases} 2\pi \left((r_i^{\text{out}})^2 - (r_i^{\text{in}})^2 \right) & \text{for } i = 1, \ldots, I-1 \\ 2\pi \left((r_i^{\text{out}})^2 - (r_i^{\text{in}})^2 \right) + 2\pi RZ & \text{for } i = I \end{cases}.$$

The temperature at each element "center" position r_i^{cen} is denoted T_i; to determine radiative and convective loss from element i the temperature is approximated as constant throughout the element and equal to T_i.

The net heat flow q_i into each wafer element i is taken to be the sum of five terms:

$$q_i = q_i^{\text{em}} + q_i^{\text{ab}} + q_i^{\text{cond}} + q_i^{\text{conv}} + q_i^{\text{dist}}; \tag{3}$$

each term is explained in a subsection below. The time derivative of T_i is given by

$$\dot{T}_i = \frac{q_i}{m_i C_P(T_i)}; \tag{4}$$

note that the specific heat C_P is strongly dependent on temperature.

The model of heat conduction outlined in Section 2.5 approximates the spatial derivatives of temperature as finite differences, so that (3) and (4) together for $i = 1, \ldots, I$ approximate the PDE (2) and its boundary conditions as a system of first-order ODE's with time as the independent variable.

2.3 Radiation emitted by the wafer

Each wafer element i is taken to be a gray, opaque body, emitting heat diffusely at rate $-q_i^{\mathrm{em}}$ given by the Stefan-Boltzmann law:

$$q_i^{\mathrm{em}} = -\epsilon_i \sigma A_i T_i^4, \tag{5}$$

where σ is the Stefan-Boltzmann constant, 5.67×10^{-8} W m^{-2} K^{-4}, and ϵ_i is the spatially-and-spectrally-averaged emissivity of element i.

2.4 Radiation absorbed by the wafer

The main sources of radiative energy in an RTP system are the lamps and the wafer itself. Photons emitted by the sources and absorbed by the wafer may pass directly from source to wafer or may reflect off of various surfaces along the way; of course photons from the wafer must undergo at least one reflection in order to return to the wafer. The relation between power emitted by a source and power absorbed by a wafer element is assumed to be linear. Let P_j denote the power radiated by lamp zone j, and J the total number of lamp zones in the system. Then the radiation absorbed by element i can be written as

$$q_i^{\mathrm{ab}} = \sum_{j=1}^{J} L_{i,j} P_j + \sum_{i'=1}^{I} W_{i,i'}\, q_{i'}^{\mathrm{em}} + q_i^{\mathrm{wall}}. \tag{6}$$

The coefficients $L_{i,j}$ and $W_{i,i'}$ are the fraction of the power radiated by lamp zone j and the fraction of the power radiated by the wafer element i' that are absorbed by element i; q_i^{wall} is the power radiated by the chamber walls that is absorbed by element i. It is assumed that $0 \leq P_j \leq P_j^{\mathrm{max}}$, where P_j^{max} is the maximum power that can be radiated by lamp zone j.

The assumption of linearity in (6) is exactly correct if reflection and absorption properties of the surfaces involved are independent of light wavelength. It is often a reasonable approximation in the case of practical RTP systems.

The quantities $W_{i,i'}$ and $L_{i,j}$ depend on the lamp and chamber geometries, which can regarded as fixed, but also depend on the absorptivities and reflectivities of the wafer surfaces and chamber surfaces. The absorptivities and reflectivities of the wafer may change quite drastically during some RTP processes, so the $W_{i,i'}$'s and $L_{i,j}$'s may well vary significantly during processing. Simulation studies in this paper will assume however that the $W_{i,i'}$'s and $L_{i,j}$'s are fixed. Results in Section 6 indicate that if good temperature measurements are available then feedback control can be used to ensure that control schemes based on estimates of the $W_{i,i'}$'s and $L_{i,j}$'s are robust with respect to rather gross errors in the estimates.

2.5 Conduction

If temperature in the wafer is a function of radial position only then at radius r heat flux due to conduction in the radial direction is given by

$$q^{\text{cond}} = -kA\frac{\partial T}{\partial r} \tag{7}$$

where k is the thermal conductivity and $A = 2\pi r Z$ is the area across which conduction is occurring. Application of (7) to infinitesimal elements leads to the PDE (2). But (7) can also be used as described below to get an approximation for q_i^{cond}, the conductive heat flow into element i.

The thermal conductivity at the inner boundary of element i is approximated by the thermal conductivity of a linearly interpolated temperature:

$$k_{i,i-1} = k\left(T_i + \frac{(T_{i-1} - T_i)(r_i^{\text{cen}} - r_i^{\text{in}})}{r_i^{\text{cen}} - r_{i-1}^{\text{cen}}}\right)$$

and $k_{i,i+1}$, the thermal conductivity at the outer boundary, is approximated using the same technique with elements i and $i+1$. Then using finite difference approximations for $\frac{\partial T}{\partial r}$ at the element boundaries the equation for conduction is

$$q_i^{\text{cond}} = -\left(2k_{i,i-1}\,\frac{T_i - T_{i-1}}{r_i^{\text{cen}} - r_{i-1}^{\text{cen}}}\,\pi\, r_i^{\text{in}}\, Z \;+\; 2k_{i,i+1}\,\frac{T_i - T_{i+1}}{r_i^{\text{cen}} - r_{i+1}^{\text{cen}}}\,\pi\, r_i^{\text{out}}\, Z\right). \tag{8}$$

If element i is the innermost or outermost element, then conduction occurs at only one boundary and the appropriate term in (8) is dropped.

2.6 Convection

Convective heat transfer into the wafer to the surrounding gas is taken to occur according to

$$q_i^{\text{conv}} = -h_i A_i \cdot (T_i - T_{\text{gas}}), \tag{9}$$

where T_{gas} is the ambient gas temperature and h_i is the convective heat transfer coefficient, taken to be constant over the element surface. The coefficient h_i depends on the location in space of element i, and the composition, pressure, and flow rate of the gas in the chamber.

Work by Campbell and colleagues [CAK+91, KCL91] shows that convective heat transfer in RTP systems is a significantly more complex phenomenon than is suggested by (9), but for the purposes of this paper (9) will be a good enough model.

2.7 Disturbances

The fifth term on the right side of (3) is q_i^{dist}, a disturbance heat flow. There are two main components to q_i^{dist}: heat flow from sources not included in the model; and heat flow due to errors in model parameters or model structure.

An example of heat flow from unmodeled sources is heat transfer to the wafer from the window separating the chamber and the lamp array. If cooling for the window is too weak to keep the window near room temperature, the window may become hot enough to contribute noticeably to the wafer energy balance.

Heat flow due to errors in model parameters arises due to discrepancies between the actual values of such coefficients as ϵ and h and the values used in the model. Heat flow due to errors in model structure results from discrepancies between the model structure and physical reality such as those discussed in Sections 2.4 and 2.6.

In Sections 4 and 5 it will be assumed that q_i^{dist} is zero and then in Section 6 it will be shown that the application of feedback control based on temperature measurements can almost entirely cancel the effect of q_i^{dist} even when the effect is quite large.

2.8 Model in vector-matrix form

Let the following vector notation be used for temperatures, heat flows, and lamp powers:

$$\mathbf{T} = \begin{bmatrix} T_1 \\ \vdots \\ T_I \end{bmatrix}, \quad \mathbf{T}^p = \begin{bmatrix} T_1^p \\ \vdots \\ T_I^p \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_I \end{bmatrix}, \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} P_1 \\ \vdots \\ P_J \end{bmatrix}.$$

Let the $I \times I$ matrices W, K^{rad}, K^{conv}, $K^{\text{cond}}(\mathbf{T})$, and $C(\mathbf{T})$ (recall that thermal conductivity and specific heat are temperature-dependent) be defined as follows:

$$\begin{aligned}
W &= [W_{i,i'}] \\
K^{\text{rad}} &= (-\mathcal{I} + W)\,\text{diag}\begin{bmatrix} \epsilon_1 \sigma A_1 \\ \vdots \\ \epsilon_I \sigma A_I \end{bmatrix} \\
K^{\text{conv}} &= \text{diag}\begin{bmatrix} -h_1 A_1 \\ \vdots \\ -h_I A_I \end{bmatrix} \\
K_{i,i}^{\text{cond}} &= -\left(\frac{2\pi r_i^{\text{in}} Z k_{i,i-1}}{r_i^{\text{cen}} - r_{i-1}^{\text{cen}}} + \frac{2\pi r_i^{\text{out}} Z k_{i,i+1}}{r_{i+1}^{\text{cen}} - r_i^{\text{cen}}} \right) \quad \text{if } 2 \le i \le I-1 \\
K_{1,1}^{\text{cond}} &= -\frac{2\pi r_1^{\text{out}} Z k_{1,2}}{r_2^{\text{cen}} - r_1^{\text{cen}}} \\
K_{I,I}^{\text{cond}} &= -\frac{2\pi r_I^{\text{in}} Z k_{I,I-1}}{r_I^{\text{cen}} - r_{I-1}^{\text{cen}}} \\
K_{i-1,i}^{\text{cond}} &= \frac{2\pi r_i^{\text{in}} Z k_{i,i-1}}{r_i^{\text{cen}} - r_{i-1}^{\text{cen}}} \quad \text{if } 2 \le i \le I
\end{aligned}$$

$$
\begin{aligned}
K^{\text{cond}}_{i+1,i} &= \frac{2\pi r_i^{\text{out}} Z k_{i,i+1}}{r_{i+1}^{\text{cen}} - r_i^{\text{cen}}} \quad \text{if } 1 \le i \le I-1 \\
K^{\text{cond}}_{i,i'} &= 0 \quad \text{if } |i-i'| > 1 \\
C &= \operatorname{diag} \begin{bmatrix} m_1 C_P(T_1) \\ \vdots \\ m_I C_P(T_I) \end{bmatrix},
\end{aligned}
$$

where $\mathcal{I}$ is the identity matrix and $\operatorname{diag} \mathbf{x}$ is the diagonal matrix with each i,i-entry equal to the ith entry of the vector $\mathbf{x}$. Let the $I \times J$ matrix L be given by

$$L = \left[L_{i,j} \right].$$

Then equation (3) can be written as

$$\mathbf{q} = K^{\text{rad}}\mathbf{T}^4 + K^{\text{cond}}(\mathbf{T})\mathbf{T} + K^{\text{conv}} \left(\mathbf{T} - \begin{bmatrix} T_{\text{gas}} \\ \vdots \\ T_{\text{gas}} \end{bmatrix} \right) + L\mathbf{P} + \mathbf{q}^{\text{wall}} + \mathbf{q}^{\text{dist}}, \quad (10)$$

and equation (4) as

$$\dot{\mathbf{T}} = C^{-1}\mathbf{q}. \quad (11)$$

2.9 Numerical determination of temperature profiles

Suppose for a given $\mathbf{P}$ and $\mathbf{q}^{\text{dist}}$, both constant in time, we wish to know the corresponding steady-state temperature profile $\mathbf{T}^{\text{ss}}$. A nonlinear equation relating $\mathbf{P}$, $\mathbf{q}^{\text{dist}}$ and $\mathbf{T}^{\text{ss}}$ is given by (10) with $\mathbf{q} = 0$ because in steady-state $\dot{\mathbf{T}}$ is zero. Starting with an initial guess $\mathbf{T}^{\text{ss}}_0$ we can use Newton's method to converge iteratively to the solution:

$$
\begin{aligned}
\mathbf{q}_m &= K^{\text{rad}}(\mathbf{T}^{\text{ss}}_m)^4 + K^{\text{cond}}(\mathbf{T}^{\text{ss}}_m)\mathbf{T}^{\text{ss}}_m + K^{\text{conv}} \left(\mathbf{T}^{\text{ss}}_m - \begin{bmatrix} T_{\text{gas}} \\ \vdots \\ T_{\text{gas}} \end{bmatrix} \right) \\
&\quad + L\mathbf{P} + \mathbf{q}^{\text{wall}} + \mathbf{q}^{\text{dist}} \\
\mathbf{T}^{\text{ss}}_{m+1} &= \mathbf{T}^{\text{ss}}_m - \left\{ 4K^{\text{rad}} \operatorname{diag}\left((\mathbf{T}^{\text{ss}}_m)^3 \right) + K^{\text{cond}}(\mathbf{T}^{\text{ss}}_m) + K^{\text{conv}} \right\}^{-1} \mathbf{q}_m.
\end{aligned}
$$

(This is not *exactly* Newton's method because a term involving the derivative of $K^{\text{cond}}(\mathbf{T})$ with respect to $\mathbf{T}$ has been neglected.) Experience has shown that the iteration always converges and that adequate precision is obtained in at most five or six iterations even if the error in the initial guess is hundreds of degrees.

Transient temperature profiles can be obtained using an ODE solver. Simulations in this paper used a Runge-Kutta method implementation called *ode45*, part of the numerical analysis package PRO-MATLAB [Mat90].

3 A Simulated RTP System

Most of the numerical results in this paper have come from computer simulations with a model of the form introduced in Section 2. Parameters for this model were obtained for the idealized RTP system shown in Figure 2. For convenience this system will be referred to as System S.

While some of the assumptions made in the formulation of a model for System S are unrealistic, the behavior of the model qualitatively reflects the behavior of an actual RTP system well enough that the model can be used to make some important observations about RTP temperature control. Knowledge gained from fictional systems like System S has already proved useful in temperature control system design and analysis for real RTP systems.

The radiative heat transfer model for System S was adapted from a model presented by Lord [Lor88]. Similar models have been been used by Young and McDonald [YM90] and Campbell, Knutson, *et. al.* [CAK$^+$91, KCL91]; in [Lor88] and [CAK$^+$91] it is shown that simulation results agree well with experimental data. Simulation study results that appear in [NSB91] were obtained using system almost identical to System S except with four lamps and a six-inch wafer. The multiple-lamp radiative heat transfer model of [KCL91] is very similar to that used for System S and in [NSB91]; the radiative heat transfer model in [YM90] is somewhat related in that it includes multiple controllable radiative sources in the chamber.

In System S, the wafer is suspended on three quartz pins, the effects of which on heat transfer are neglected. The wafer is assumed to be a perfectly cylindrical piece of pure silicon. Data for specific heat, C_P, and thermal conductivity, k, is taken from [KS88] and [Wyb88]; linear interpolation is used to determine values of C_P and k at temperatures in between the temperature breakpoints of the data. The wafer diameter ($2R$) is 100.8 mm (four inches) and its thickness (Z) is 0.675 mm. The density of Si is taken to be 2330 kg m^{-3}. The top, bottom, and edge surfaces of the wafer are assumed to be gray and diffusely reflecting with a total emissivity of 0.6 and the wafer is assumed to be opaque.

The wafer is modeled as a collection of twenty concentric annular elements. Outer element radii — r_i^{out} for i equal to 1 through 20 — are:

$$\tfrac{1}{16}R,\ \tfrac{2}{16}R,\ \ldots,\ \tfrac{14}{16}R,\ \tfrac{29}{32}R,\ \tfrac{60}{64}R,\ \tfrac{61}{64}R,\ \tfrac{62}{64}R,\ \tfrac{63}{64}R,\ R.$$

The selection of a somewhat nonuniform grid is based on the heuristic reasoning that since the highest temperature gradient is expected near the wafer edge the grid should be finer in that region.

The interior of the chamber is also axisymmetric and is coaxial with the wafer. The inner diameter of the chamber is 260 mm, the distance from chamber floor to chamber ceiling 150 mm, and the distance from wafer bottom to chamber floor 100 mm.

The inner wall of the chamber is modeled as a gray, diffuse, opaque surface with an emissivity of 0.5, with the exception of special zones in the chamber

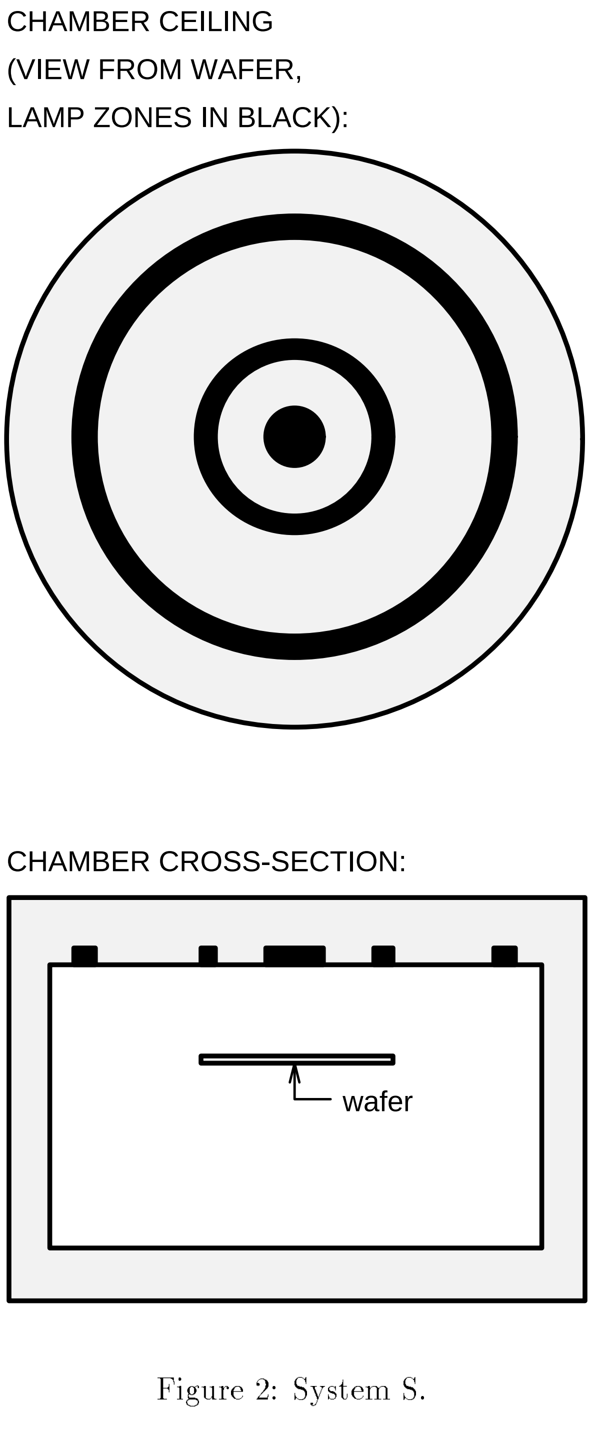


Figure 2: System S.

ceiling. These special zones are three lamps, the center one circular and the other two annular, all coaxial with the wafer, modeled as diffuse, black surfaces. It is assumed that the power radiated by each lamp can be directly commanded by the RTP system's controller. The lamps are numbered 1 to 3 starting at the center lamp and moving outward. Lamp maximum powers are

$$P_1^{\max} = 2\,\mathrm{kW}, \quad P_2^{\max} = 10\,\mathrm{kW}, \quad \text{and } P_3^{\max} = 35\,\mathrm{kW}.$$

The remainder of the chamber inner surface is taken to have a temperature of 300 K at all times.

The chamber ceiling is divided into ten concentric zones (three of which are the lamps), the chamber floor into nine concentric zones, the chamber side wall into eight vertically stacked tubular zones, and the wafer surface into forty-one zones (twenty element tops, twenty element bottoms, and the edge). Radiative heat transfer in the system can be computed using angle factors [SC78, Lor88]. Angle factors were computed for every pair of zones in the system, taking the total or partial shading caused by the wafer into account if necessary. From these angle factors the vector $\mathbf{q}^{\mathrm{wall}}$ and the matrices F and W can be computed.

Figure 3 shows graphically what each column of the matrix F looks like for System S. Each bar graph j shows how much power from lamp j is absorbed by each wafer element when $P_j = P_j^{\max}$; the powers absorbed by the elements is normalized with respect to the element masses. Two facts about the columns of F are worth emphasizing:

- Each lamp supplies significant power to each wafer element. These lamps are *not* designed to illuminate only part of the wafer.
- Lamp 3 supplies much more power/mass to element 20 than to element 19, because lamp 3 supplies considerable power to the wafer edge. This will allow compensation for radiative and convective loss from the wafer edge.

The convective heat transfer model for System S is intended to produce effects similar in size and shape to those produced by convection in actual RTP systems. Two different chamber gas conditions are modeled: vacuum and slowly flowing N_2 at 1 atm. For the N_2 case T_{gas} is taken to be 300 K and the convective loss coefficient for element i is taken to be

$$h_i = \left(7.1 + 4.3\left(\frac{r_i^{\mathrm{cen}}}{R}\right)^4\right)\frac{\mathrm{W}}{\mathrm{m^2K}}; \tag{12}$$

the formula and the numbers are taken from [Lor88]. Evidence that (9) and (12) together form an adequate model by the criterion of effects of similar size and shape is shown in Figure 4. The solid curve shows the steady-state wafer temperature profile for System S in the N_2 case with the lamps adjusted for optimal uniformity (see Section 4 for details on this adjustment) with a

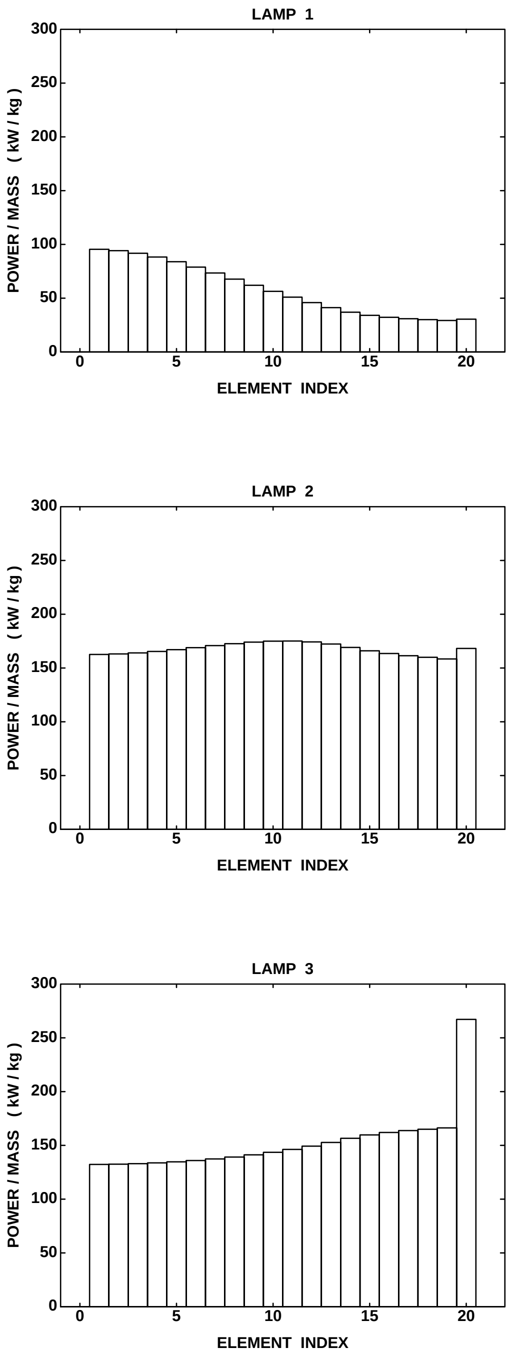


Figure 3: Illumination patterns for System S.

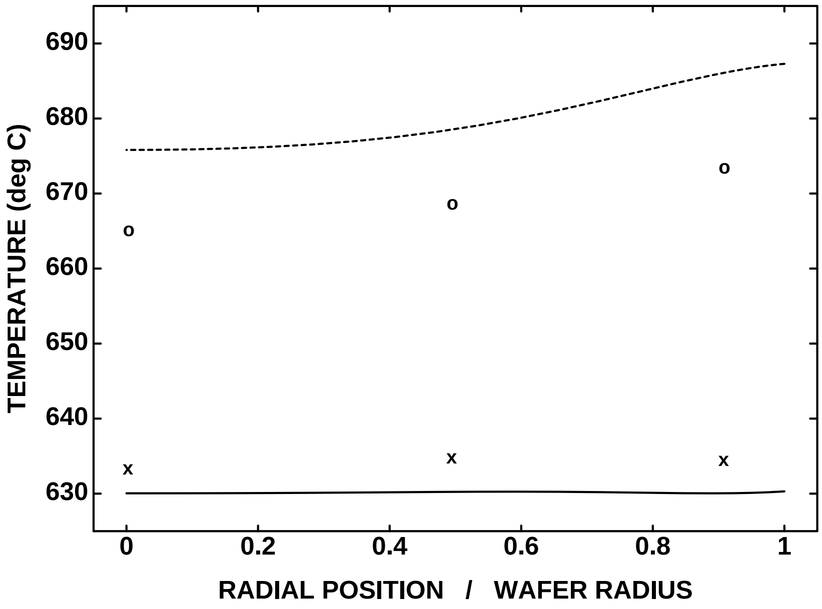


Figure 4: Effects of convective heat loss in System S and the Stanford RTM.

setpoint of 630°C and the dashed curve shows the steady-state profile with the same lamp settings when the chamber is evacuated. The O's are steady-state thermocouple readings for certain lamp settings in the Stanford RTM system with a 1 atm N_2 ambient and the X's are steady-state thermocouple readings for the same lamp settings with a 1 torr N_2 ambient. The wafer size in both the fictional system and the real system is four inches.

4 Optimization of Steady-State Uniformity

Consider the problem of choosing lamp power setpoints that result in the steady-state temperature at each point on the wafer being close to some specified temperature T^{set}. The most natural goal is to find lamp settings that minimize the worst-case difference between wafer temperature and T^{set}, that is minimize the quantity

$$\max_{0 \leq r \leq R} \left| T^{\text{ss}}(r, \mathbf{P}) - T^{\text{set}} \right|, \tag{13}$$

where $T^{\text{ss}}(r, \mathbf{P})$ is the steady-state temperature at radius r with the constant lamp power vector $\mathbf{P}$.

4.1 Minimax error approximation and linear programming

Before the presentation of a technique for minimizing (13), a brief digression to introduce some notation and to state some standard, simple optimization results is necessary. A reference for this material is [Chv83, Ch. 14].

Consider the system of linear equations represented by

$$A\mathbf{x} = \mathbf{b}, \tag{14}$$

where A is an $m \times n$ matrix, $\mathbf{x}$ is an $n \times 1$ column vector, $\mathbf{b}$ is an $m \times 1$ column vector, and $m > n$. The system is overdetermined: In general, for no value of $\mathbf{x}$ will (14) be exactly satisfied. It is therefore of interest to find an approximate solution, that is, a value of $\mathbf{x}$ for which the vector $A\mathbf{x} - \mathbf{b}$ is in some sense as small as possible.

The most familiar approximate solution to (14) is the *least-squares* solution, which is given by the solution[3] $\hat{\mathbf{x}}_{\mathrm{LS}}$ of

$$A^T A\, \hat{\mathbf{x}}_{\mathrm{LS}} = A^T \mathbf{b}.$$

The least-squares solution is the value of $\mathbf{x}$ for which $\|A\mathbf{x} - \mathbf{b}\|_2$, the two-norm of the approximation error $A\mathbf{x} - \mathbf{b}$, is minimal. The *two-norm* (or *Euclidean norm*), $\|\mathbf{v}\|_2$, of a vector $\mathbf{v}$, is the square root of $\sum_i v_i^2$, where the v_i's are the entries of $\mathbf{v}$.

An alternate approximate solution to (14) is what will be called in this paper the *minimax-error* solution. The minimax-error solution $\hat{\mathbf{x}}_{\mathrm{MME}}$ is the value of $\mathbf{x}$ for which $\|A\mathbf{x} - \mathbf{b}\|_\infty$ is minimal. The *infinity-norm* (or *max-norm*), $\|\mathbf{v}\|_\infty$, of a vector $\mathbf{v}$ is $\max_i |v_i|$, the largest absolute value of the entries of $\mathbf{v}$. In many cases, including several soon to be discussed in this paper, the minimax-error solution more exactly suits an engineer's need than does the least-squares solution.

Unlike the least-squares problem, the minimax-error problem does not have a simple, "closed-form" solution. The minimax-error problem is equivalent to a linear programming problem (hereafter abbreviated LP) [Chv83, Dan63, Lue84]:

$$\begin{aligned} &\text{Minimize} && \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ e \end{bmatrix} \\ &\text{subject to} && \begin{bmatrix} A & -\mathbf{1}_I \\ -A & -\mathbf{1}_I \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ e \end{bmatrix} \le \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix}, \end{aligned} \tag{15}$$

where $\mathbf{1}_I$ is a column vector with all I entries equal to one, and where for vectors $\mathbf{v}$ and $\mathbf{w}$, $\mathbf{v} \le \mathbf{w}$ means every entry of $\mathbf{v}$ is less than the corresponding entry of $\mathbf{w}$. (Note: Some LP formulations stipulate that all of the variables be nonnegative; *no* such constraint is present here.) If $\begin{bmatrix} \mathbf{x}^* \\ e^* \end{bmatrix}$ is a solution[4] to the LP, then

$$\hat{\mathbf{x}}_{\mathrm{MME}} = \mathbf{x}^* \quad \text{and} \quad \|A\hat{\mathbf{x}}_{\mathrm{MME}} - \mathbf{b}\|_\infty = e^*.$$

[3] There are multiple solutions if the rank of A is strictly less than n.

[4] Here too there there may be multiple solutions.

A relative of the minimax-error problem is what will be called here the *constrained minimax-error* problem:

$$\text{Minimize } \|A\mathbf{x}-\mathbf{b}\|_\infty \text{ subject to } \mathbf{l} \le \mathbf{x} \le \mathbf{u}; \tag{16}$$

here there are lower and upper bounds placed on each of the entries of $\mathbf{x}$. Solving this problem is really no harder to solve than is the unconstrained minimax-error problem. The constrained minimax-error problem is equivalent to an LP with the same variables and objective function as (15) but with a slightly expanded constraint set. Here too if $\begin{bmatrix} \mathbf{x}^* \\ e^* \end{bmatrix}$ is a solution to the LP, then

$$\hat{\mathbf{x}}_{\text{CMME}} = \mathbf{x}^* \quad \text{and} \quad \|A\hat{\mathbf{x}}_{\text{CMME}} - \mathbf{b}\|_\infty = e^*.$$

Iterative algorithms are required to solve LP's. One such algorithm is the famous Simplex Algorithm, a variant of which, the computer code LSSOL [GHM+86], was used to solve the LP's generated later in this paper. For constrained minimax-error problems where A is 20×3 — all of the examples solved in this paper are this size except for the exactly optimal trajectory design problem in Section 5.4 — the LP is solved in a few seconds by a Sun workstation. To solve the constrained minimax-error problem of Section 5.4 takes tens of seconds to a tens of minutes, depending on the size of the problem.

4.2 Numerical optimization via model

Suppose an RTP system is well described by the model developed in Section 2 and all the model parameters are known. With very little error we can approximate (13) as

$$\max_i \left| T_i^{\text{ss}}(\mathbf{P}) - T^{\text{set}} \right| = \left\| \mathbf{T}^{\text{ss}}(\mathbf{P}) - \mathbf{T}^{\text{set}} \right\|_\infty, \tag{17}$$

where $T_i^{\text{ss}}(\mathbf{P})$ is the steady-state temperature of element i with constant lamp power vector $\mathbf{P}$ and $\mathbf{T}^{\text{set}}$ is a vector with all entries equal to T^{set}.

Because $T_i^{\text{ss}}(\mathbf{P})$ is a nonlinear function of $\mathbf{P}$, the problem of minimizing (17) is not a minimax-error problem as defined in Section 4.1. However, the problem can be solved in two steps, each of which involves solving a minimax-error problem.

Step 1. Let

$$\mathbf{q}^{\text{perfect}} = -K^{\text{rad}}(\mathbf{T}^{\text{set}})^4 - K^{\text{conv}} \left(\mathbf{T}^{\text{set}} - \begin{bmatrix} T_{\text{gas}} \\ \vdots \\ T_{\text{gas}} \end{bmatrix} \right) - \mathbf{q}^{\text{wall}};$$

comparison with equation (10) indicates that $\mathbf{q}^{\text{perfect}}$ is the heat flux vector that would have have to be supplied by the lamps if the steady-state temperature profile were to be exactly $\mathbf{T}^{\text{set}}$. Find the solution $\hat{\mathbf{P}}$ to the constrained minimax-error problem

$$\text{Minimize } \|L\mathbf{P} - \mathbf{q}^{\text{perfect}}\|_\infty \text{ subject to } \mathbf{0} \le \mathbf{P} \le \mathbf{P}^{\max},$$

and determine $\mathbf{T}^{ss}(\hat{\mathbf{P}})$, which will be very close to $\mathbf{T}^{set}$ — if not, either the lamp array provides an inadequate range of flux patterns or T^{set} is higher than the maximum wafer temperature obtainable with the RTP system.

Step 2. Since $\mathbf{T}^{ss}(\hat{\mathbf{P}})$ is close to $\mathbf{T}^{set}$ the following first-order approximation will be quite accurate:

$$\mathbf{T}^{ss}(\hat{\mathbf{P}} + \mathbf{\Delta P}) \approx \mathbf{T}^{ss}(\hat{\mathbf{P}}) + D_{\mathbf{P}}\mathbf{T}^{ss}(\hat{\mathbf{P}})\mathbf{\Delta P}, \tag{18}$$

where $D_{\mathbf{P}}\mathbf{T}^{ss}(\hat{\mathbf{P}})$ is the matrix of partial derivatives

$$\left.\begin{bmatrix} \frac{\partial T_1^{ss}}{\partial P_1} & \cdots & \frac{\partial T_1^{ss}}{\partial P_J} \\ \vdots & & \vdots \\ \frac{\partial T_I^{ss}}{\partial P_1} & \cdots & \frac{\partial T_I^{ss}}{\partial P_J} \end{bmatrix}\right|_{\mathbf{P}=\hat{\mathbf{P}}}.$$

Setting $\mathbf{q} = \mathbf{0}$ in equation (10), differentiating with respect to $\mathbf{P}$, assuming that $D_{\mathbf{T}}\,K^{\text{cond}}$ is small, and re-arranging, one finds that

$$D_{\mathbf{P}}\mathbf{T}^{ss} \approx -\left(4K^{\text{rad}}\,\text{diag}\left((\mathbf{T}^{ss})^3\right) + K^{\text{cond}}(\mathbf{T}^{ss}) + K^{\text{conv}}\right)^{-1} L. \tag{19}$$

So starting from $\hat{\mathbf{P}}$ and $\mathbf{T}^{ss}(\hat{\mathbf{P}})$ the problem of minimizing (17) can be reformulated as

$$\begin{array}{c} \text{Minimize} \left\| D_{\mathbf{P}}\mathbf{T}^{ss}(\hat{\mathbf{P}})\,\mathbf{\Delta P} - \left(\mathbf{T}^{set} - \mathbf{T}^{ss}(\hat{\mathbf{P}})\right)\right\|_\infty \\ \text{subject to } -\hat{\mathbf{P}} \leq \mathbf{\Delta P} \leq \mathbf{P}^{\max} - \hat{\mathbf{P}}, \end{array} \tag{20}$$

which when the approximation (19) is applied is a constrained minimax-error problem. If $\mathbf{\Delta P}^*$ is the solution to (20) then $\hat{\mathbf{P}} + \mathbf{\Delta P}^*$ minimizes (17).

4.3 Numerical optimization without a model

Objections that might be raised to the uniformity optimization technique described in Section 4.2 are:

- It requires a detailed thermal model of the RTP system. Such a model may not be available or would require considerable effort to develop.
- Even if a model were available, inaccuracies in the model would seriously degrade the performance of the optimization technique.

If a real RTP system with multiple temperature sensors is available, these two problems can be completely circumvented by applying the modifications described below.

In this subsection $\mathbf{T}^{ss}$ will denote a vector of sensed or estimated temperatures at various radial positions on the wafer rather than a vector of element temperatures. Ideally, this vector would have enough entries to capture the entire wafer temperature profile. It could be obtained with a scanning pyrometer or by interpolation of data from several sensors measuring temperature at fixed positions.[5] With this redefinition of $\mathbf{T}^{\mathrm{ss}}$ the two-step procedure of the preceding subsection can be modified as follows.

Modifications to Step 1. The goal in this step is simply to find a $\hat{\mathbf{P}}$ such that $\mathbf{T}^{\mathrm{ss}}(\hat{\mathbf{P}})$ is close to $\mathbf{T}^{\mathrm{set}}$. There many ways to do this with an actual system. One is to manually adjust the lamp settings until the temperature profile is near $\mathbf{T}^{\mathrm{set}}$, and another is to apply some kind of closed-loop control that achieves the same effect. (In fact, in many practical applications the resulting temperature profile would be so good that Step 2 would be a waste of time.)

Modifications to Step 2. The matrix $D_{\mathbf{P}}\mathbf{T}^{\mathrm{ss}}(\hat{\mathbf{P}})$ can be determined experimentally, by measuring $\mathbf{T}^{\mathrm{ss}}(\mathbf{P}^{(1)})$ through $\mathbf{T}^{\mathrm{ss}}(\mathbf{P}^{(J)})$, where each $\mathbf{P}^{(j)}$ is $\hat{\mathbf{P}}$, slightly perturbed as follows:

$$\mathbf{P}^{(1)} = \hat{\mathbf{P}} + \begin{bmatrix} \Delta P_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{P}^{(2)} = \hat{\mathbf{P}} + \begin{bmatrix} 0 \\ \Delta P_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{etc.,}$$

and approximating each column j of $D_{\mathbf{P}}\mathbf{T}^{\mathrm{ss}}(\hat{\mathbf{P}})$ as

$$\frac{\partial}{\partial P_j}\mathbf{T}^{\mathrm{ss}} \approx \frac{1}{\Delta P_j}\left(\mathbf{T}^{\mathrm{ss}}(\mathbf{P}^{(j)}) - \mathbf{T}^{\mathrm{ss}}(\hat{\mathbf{P}})\right).$$

A further objection is: What if $\mathbf{q}^{\mathrm{dist}}$ is unpredictable and significant? If this is true, then the experimentally-determined optimal temperature profiles may not be repeatable, or may drift slowly away from $\mathbf{T}^{\mathrm{set}}$. This difficulty can be largely overcome by use of feedback control, as described in Section 6.

4.4 Optimal steady-state uniformity for System S

Figure 5 shows results from the application to System S of the method of Section 4.2. The plots show the lamp power settings that minimize the worst-case difference between wafer temperature and T^{set} for values of T^{set} between 575°C and 1125°C. The upper plot is for the case of a 1 atm N_2 ambient and the lower plot is for the case of a vacuum ambient. The lamp settings are shown as fractions of maximum power and plotted using a log scale so that at any abscissa the spacings between the curves indicate the ratios between lamp settings at that abscissa.

In the vacuum case, these ratios remain constant. This is because heat loss from the wafer is radiative only — the magnitude of heat loss is proportional

[5] The suggestion that temperature controllers in RTP systems *actually used in IC fabs* must measure temperature virtually continuously across the wafer should *not* be inferred from these comments. On the other hand, having such measurements available during RTP system *development* would be useful.

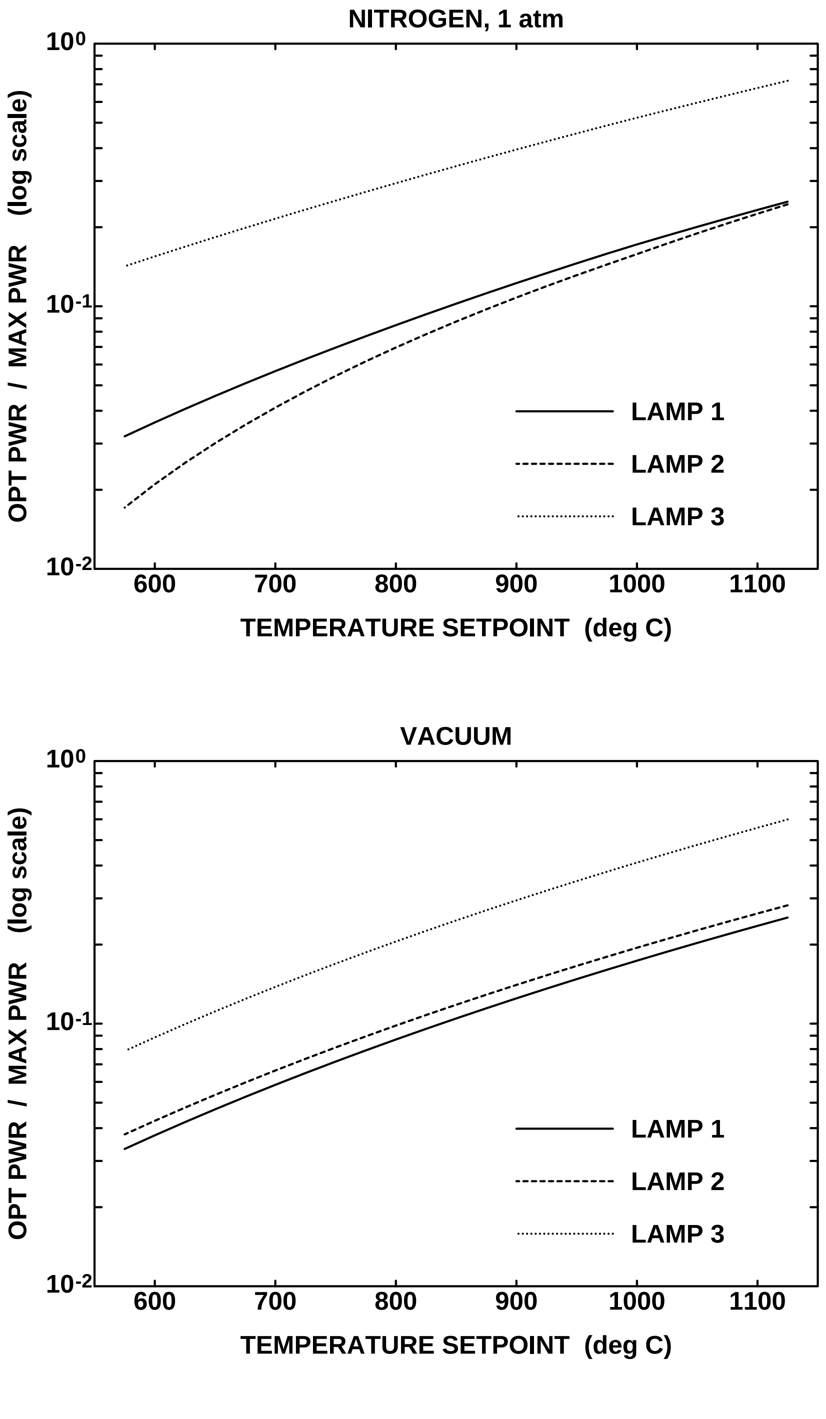


Figure 5: Optimal steady-state lamp settings as a function of T^{set}.

to the fourth power of T^{set} but the shape of the spatial distribution of heat loss is independent of T^{set}. In the case of a gas ambient, however, the relative importances of radiative loss and convective loss change with T^{set}. The spatial distribution of heat loss from convection has a different shape than that from radiation. Thus the vertical spacings between the curves in the upper plot of Figure 5 change substantially with T^{set}. It is also noteworthy that the ratios between optimal gas ambient lamp settings for every value of T^{set} are quite different than the ratios between the optimal vacuum settings.

When T^{set} ranges between 575°C and 1125°C, and optimal lamp settings are used, the worst-case value of the max-norm of the temperature error vector (*i.e.,* the value of (17)) is 0.16°C for the N_2 ambient and 0.34°C for vacuum; therefore the largest difference in steady-state temperature between any two points on the wafer at any one T^{set} value is no more than 0.32°C for N_2 and 0.68°C for vacuum. These numbers are so small that it is fair to conclude that for System S temperature uniformity in steady-state is essentially perfect in both vacuum and 1 atm N_2 ambient over the range of setpoint temperatures considered.[6]

Figure 6 shows a typical optimal temperature profile. It happens to be for a T^{set} value of 850°C and the 1 atm N_2 ambient. The max-norm of the temperature error vector is 0.12°C.[7] It can be seen from the plot that the error is this large at four radii on the wafer. This can be explained as follows: If the maximum error occurred at three or fewer locations, there would be enough degrees of freedom to reduce the error at all of these locations. Thus as long as none of the lamp settings are at their lower or upper bounds the maximum error must occur at four or more locations.

In Figure 7 optimal steady-state temperature profiles are plotted for the N_2 ambient case and T^{set} values of 600°C and 1100°C, along with temperature profiles obtained by re-scaling the optimal lamp setting for 850°C in the N_2 ambient. The latter temperature profiles are what would be obtained with scalar PID control based on the optimal setting for 850°C, if the sensor of the scalar control system were measuring wafer center temperature. It should be clear that a significant performance penalty is paid for using scalar control. A similar conclusion can be reached from inspection of Figure 8, in which the optimal 850°C temperature profile for vacuum is plotted along with a temperature profile obtained by re-scaling the optimal N_2 ambient lamp setting for 850°C.

[6] On the other hand, these numbers are so small that it is *not* fair to use them as accurate estimates of nonuniformity but rather as guarantees that actual nonuniformity is less than 1°C. See Footnote 2.

[7] As noted in Footnote 6 with reference to Footnote 2, the temperature error vector is so small that it likely does not represent the actual error vector very well, but this is irrelevant to the qualitative point being made here.

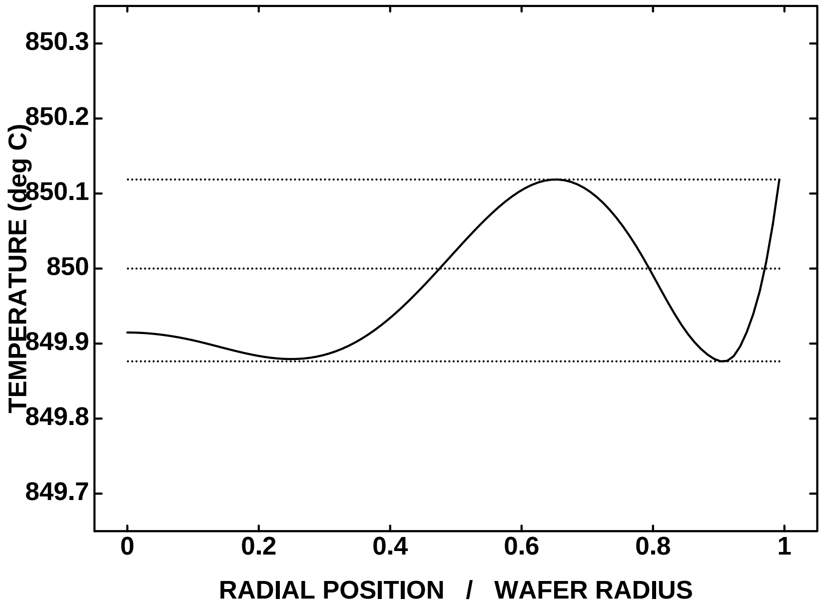


Figure 6: Optimal steady-state temperature profile in 1 atm N_2 ambient for $T^{\mathrm{set}} = 850°\mathrm{C}$.

4.5 Discussion

The idea of using different relative lamp power settings for steady-state hold at different temperature setpoints also appears in a paper by Gyurcsik, Riley and Sorrell [GRS91]. The method used in [GRS91] for selecting lamp power settings involves the following:

- ray tracing to estimate the wafer-surface flux patterns created by each lamp;
- using a least-squares fit of simulation data to approximate the temperature at a given radius as a quadratic function of incident flux density at that point;
- and using least-squares to match not the desired temperature profile but the desired illumination pattern determined using the aforementioned approximation of the relation between incident flux density and temperature.

None of these approximations are unreasonable, but their cumulative effect may mean that the steady-state temperature profile resulting from the lamp setting selected this way is significantly worse than the optimal profile. If so, it is impossible to tell how far from optimal the temperature profile is.

In contrast, the method presented here in Section 4.3 directly minimizes wafer temperature nonuniformity. By using temperature measurement data the problem of estimating lamp flux density patterns can be avoided entirely.

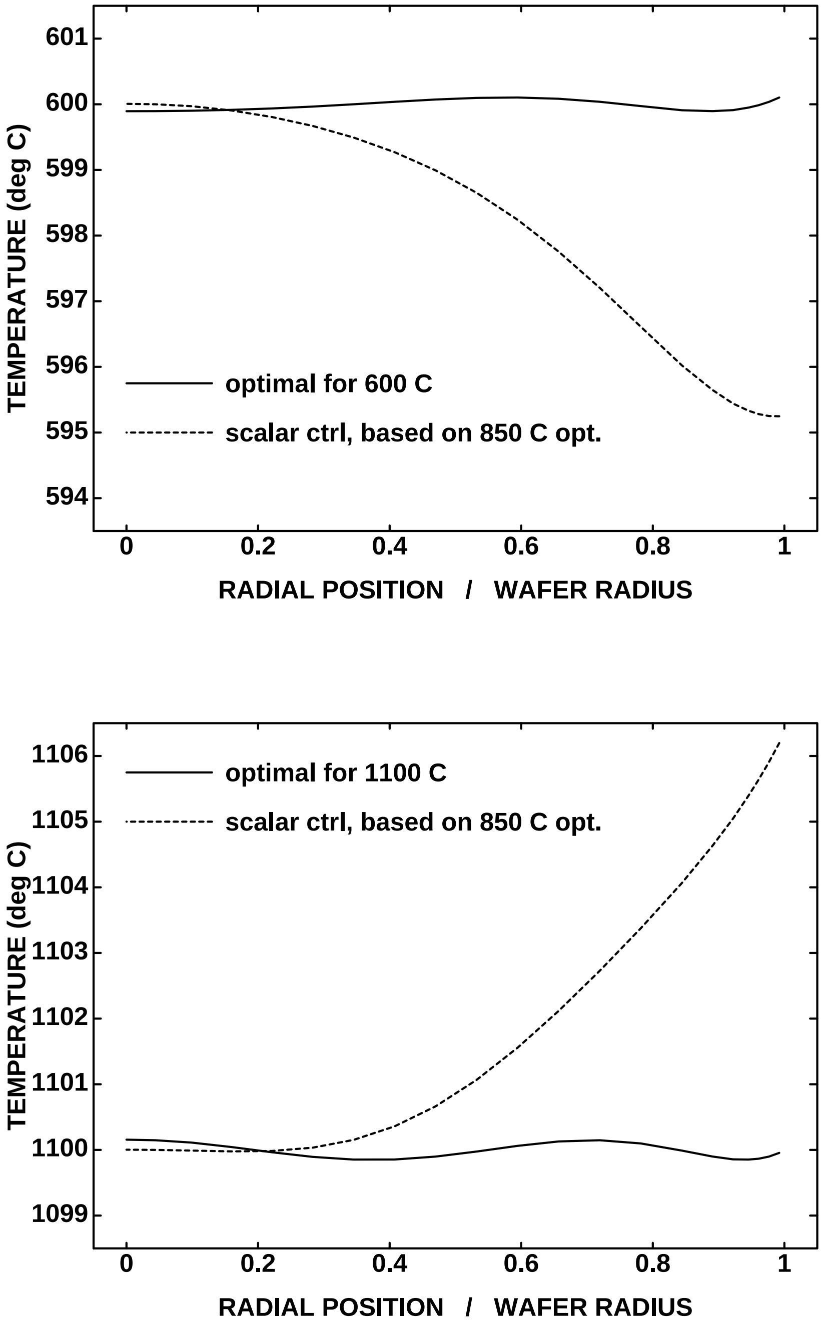


Figure 7: Scalar *vs.* multivariable control for setpoints of 600°C and 1100°C.

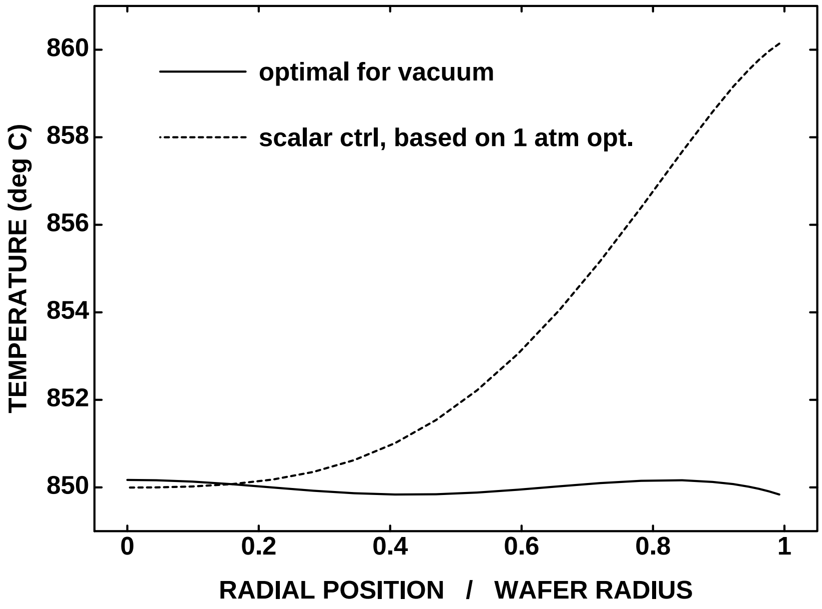


Figure 8: Steady-state temperature profiles in vacuum with $T^{\mathrm{set}} = 850°\mathrm{C}$.

An accurate answer can be obtained for the question: What is the best possible temperature uniformity achievable for temperature setpoint T^{set}?

This issue of accurate determination of limits of performance will be discussed again in Section 7.

5 Optimization of Error During Transients

This section discusses the problem of adjusting lamp power settings over time to keep both nonuniformity and trajectory error small. A natural goal is to minimize

$$\max_{0 \le r \le R,\ t_o \le t \le t_f} \left| T(r,t) - T^{\mathrm{ref}}(t) \right|, \tag{21}$$

the largest temperature error from the specified trajectory T^{ref} at any point on the wafer at any time between an initial time t_o and a final time t_f. As in Section 4 we can make an approximation: The quantity (21) is not much bigger than

$$\max_{t_o \le t \le t_f} \left\| \mathbf{T}(t) - \mathbf{T}^{\mathrm{ref}}(t) \right\|_\infty, \tag{22}$$

where $\mathbf{T}^{\mathrm{ref}}(t)$ is a vector with all its entries equal to $T^{\mathrm{ref}}(t)$.

5.1 Discretization in time

In an RTP system under automatic temperature control it is almost certain that the lamps will be controlled by a digital computer [FPW90, ÅW90]. The lamp power signals will most likely be zero-order-hold signals, that is, signals that are constant between sample instants. In this section Δt will denote the sampling period; accordingly, $1 / \Delta t$ is the sample rate. It will be assumed that the length of the interval $[t_o, t_f]$ is exactly some integer N times the sample rate:

$$t_f = t_o + N\Delta t.$$

The nth lamp power vector will be denoted $\mathbf{P}_n$:

$$\text{For } t_o + n\Delta t \leq t < t_o + (n+1)\Delta t, \ \ \mathbf{P}(t) = \mathbf{P}_n.$$

The sequence of temperature profiles and reference temperatures at sample times will be denoted $\mathbf{T}_n$ and T_n^{ref}:

$$\mathbf{T}_n = \mathbf{T}(t_o + n\Delta t); \quad \text{and } T_n^{\text{ref}} = T^{\text{ref}}(t_o + n\Delta t);$$

$\mathbf{T}_n^{\text{ref}}$ is an $I \times 1$ vector with all entries equal to T_n^{ref}.

5.2 A simple heuristic scheme

One way to approximately minimize (22) is to choose, for each t between t_o and t_f, the lamp vector $\mathbf{P}(t)$ to be the sum of two vectors:

1. The optimal lamp vector for steady-state hold at $T^{\text{ref}}(t)$, determined as in Section 4. Call this vector $\mathbf{P}^{\text{ss}}(T^{\text{ref}}(t))$.
2. The product

$$\mathbf{P}^{\text{dyn}}\left(T^{\text{ref}}(t), \dot{T}^{\text{ref}}(t)\right) = \dot{T}^{\text{ref}}(t) C_P \left(T^{\text{ref}}(t)\right) \mathbf{P}^{\text{unif}},$$

   where $\mathbf{P}^{\text{unif}}$ is chosen such that for each wafer element i,

$$\frac{1}{m_i} \left[\begin{array}{ccc} L_{i,1} & \cdots & L_{i,J} \end{array} \right] \mathbf{P}^{\text{unif}} \approx 1 \frac{\text{W}}{\text{kg}}.$$

The vector $\mathbf{P}^{\text{ss}}$ balances the loss of heat from the wafer by radiation and convection, while the vector $\mathbf{P}^{\text{dyn}}$ adds enough additional heat to obtain a rate of temperature increase of close to $\dot{T}^{\text{ref}}$ at all points on the wafer.

A block diagram for a discrete-time implementation of this scheme is shown in Figure 9. In the diagram, the box z represents a forward shift of one time-step. The non-causality of this scheme is not a problem because in RTP reference trajectories will be known in advance. In this scheme we choose

$$\mathbf{P}_n = \mathbf{P}_n^{\text{ss}} + \mathbf{P}_n^{\text{dyn}}, \tag{23}$$

where

$$\mathbf{P}_n^{\text{ss}} = \mathbf{P}^{\text{ss}}\left(\frac{T_{n+1}^{\text{ref}} + T_n^{\text{ref}}}{2}\right)$$

and

$$\mathbf{P}_n^{\rm dyn} = \left(\frac{T_{n+1}^{\rm ref} - T_n^{\rm ref}}{\Delta t}\right) C_P \left(\frac{T_{n+1}^{\rm ref} + T_n^{\rm ref}}{2}\right) \mathbf{P}^{\rm unif}.$$

The quantity $\mathbf{P}_n^{\rm ss}$ is evaluated by interpolation in a pre-computed table of values of $\mathbf{P}^{\rm ss}(T)$ at a grid of temperatures T spaced, say, 25°C or 50°C apart over the working range of the RTP system.

How can be $\mathbf{P}^{\rm unif}$ determined? Simulation studies with System S show that during wafer ramp-up and cool-down using the scheme outlined above, the temperature error profile settles to a near-constant vector. The max-norm of this error vector can be approximately minimized by picking $\mathbf{P}^{\rm unif}$ to minimize

$$\left\| -\left(D_{\mathbf{T}}\mathbf{q}(\mathbf{T}^{\rm mid})\right)^{-1} \left(C(\mathbf{T}^{\rm mid})\right)^{-1} \left(L\mathbf{P} - 1\frac{\rm W}{\rm kg}\begin{bmatrix} m_1 \\ \vdots \\ m_I \end{bmatrix}\right)\right\|_\infty, \qquad (24)$$

where, from (10),

$$D_{\mathbf{T}}\mathbf{q}(\mathbf{T}) = 4K^{\rm rad}\,\mathrm{diag}\left((\mathbf{T})^3\right) + K^{\rm cond}(\mathbf{T}) + K^{\rm conv}$$

and $T^{\rm mid}$ is some temperature in the middle of the working range of the RTP system. It may be necessary to use constraints on the entries of $\mathbf{P}^{\rm unif}$ to ensure that $(P_j^{\rm dyn})_n$ is between zero and $P_j^{\rm max}$ for every j and n.

It is possible to determine a good value for $\mathbf{P}^{\rm unif}$ using experimental data from a real system rather than from a detailed system model. There are many ways to do this; one way is as follows. The idea is similar to the idea used in Section 4.3; here again $\mathbf{T}$ will refer to a vector of measured or estimated temperatures. This is the procedure: Perform J experiments, starting each time from steady-state hold at some lamp setting $\mathbf{P}^{\rm start}$. In experiment j change the lamp setting immediately to $\mathbf{P}^{\rm start} + \mathbf{P}^{(j)}$, where $\mathbf{P}^{(j)}$ is as given in Section 4.3 but preferably with large values for the $\Delta P^{(j)}$'s, and estimate $\dot{\mathbf{T}}$ at the instant in time immediately after the power switch; call this estimate $\dot{\mathbf{T}}^{(j)}$. If the measurements are noisy some signal processing will have to be used to to get reliable derivative estimates. Let C_P be the specific heat of Si at the steady-state temperature corresponding to $\mathbf{P}^{\rm start}$. Choose as $\mathbf{P}^{\rm unif}$ the solution to this minimax error problem:

$$\text{Minimize} \quad \left\| \begin{bmatrix} \dfrac{\dot{\mathbf{T}}^{(1)}}{\Delta P^{(1)}} & \cdots & \dfrac{\dot{\mathbf{T}}^{(J)}}{\Delta P^{(J)}} \end{bmatrix} \mathbf{P} - \frac{1\,\mathrm{W\,kg^{-1}}}{C_P} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\|_\infty .$$

5.3 Heuristic scheme applied to System S

A representative RTP temperature trajectory — called hereafter, for convenience, Trajectory T — is shown in Figure 10. The initial condition is assumed to be steady-state hold with a setpoint of 600°C. The trajectory begins with

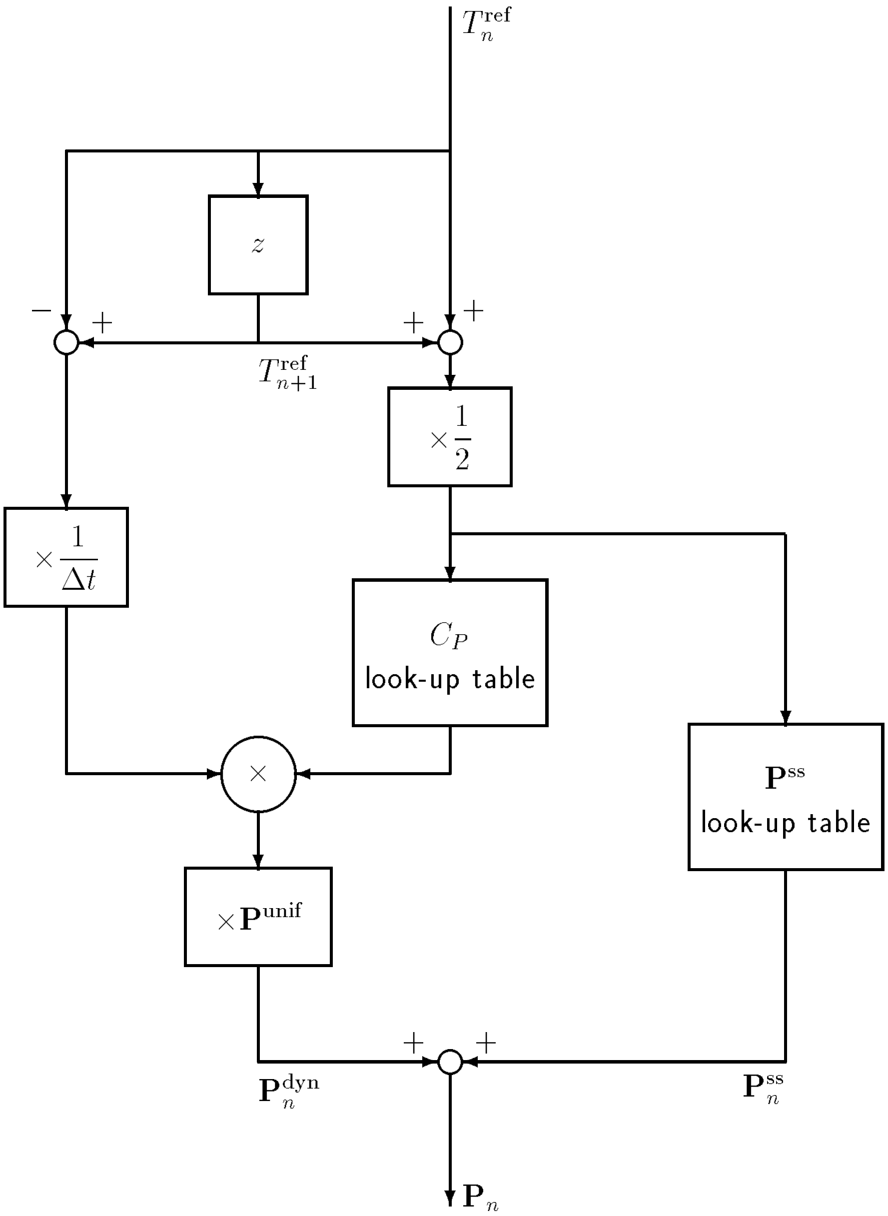


Figure 9: Discrete-time implementation of heuristic lamp power scheme.

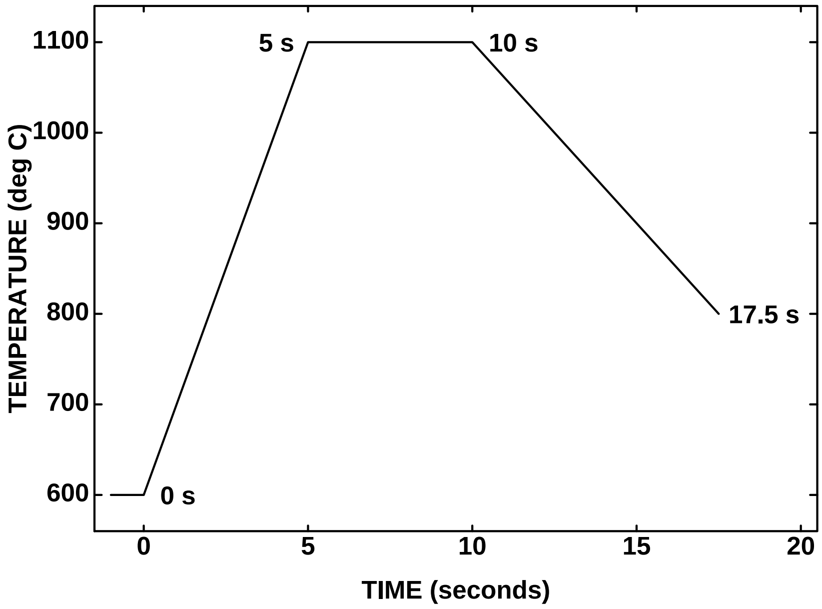


Figure 10: Trajectory T.

a ramp-up at 100 °C/s to 1100°C, continues with steady-state hold at 1100°C for 5 s, and finishes with a controlled cool-down at 40 °C/s to 800°C. It is assumed for this fictional process that once the wafer gets below 800°C the lamps can be simply turned off, that is, that the temperature nonuniformity resulting from free cool-down below 800°C will have no ill effects.

Figure 11 shows a lamp setting trajectory corresponding to Trajectory T computed using the scheme (23) for System S with the 1 atm N_2 ambient. The sample rate used was 10 Hz. Three different values of $\mathbf{P}^{\mathrm{unif}}$ were used: one for ramp-up, one for cool-down from 1100°C to 900°C, and one for cool-down from 900°C to 800°C. All three were obtained by minimization of (24); the latter two were chosen with constraints that ensured nonnegativity of the entries of $\mathbf{P}^{\mathrm{dyn}}$ at all times. Note that independent control of lamp powers is essential for obtaining this lamp power trajectory.

Figure 12 shows

$$\max_i \left(T_i(t) - T^{\mathrm{ref}}(t)\right) \quad \text{and} \quad \min_i \left(T_i(t) - T^{\mathrm{ref}}(t)\right)$$

when this lamp power trajectory is applied. The largest deviation from the specified trajectory at any time and at any point on the wafer is 4.2°C and the largest difference in temperature between any two points on the wafer at any one time is 5.7°C; both of these extremes are reached at the very end of the trajectory. Table 1 summarizes the largest errors and nonuniformities over parts of the trajectory. Approximate thermal stress computations[8] following the approach of [BCD84] show that the ratio of thermal stress in the wafer to

[8]The author thanks Dr. C. Schaper for software that performed these computations.

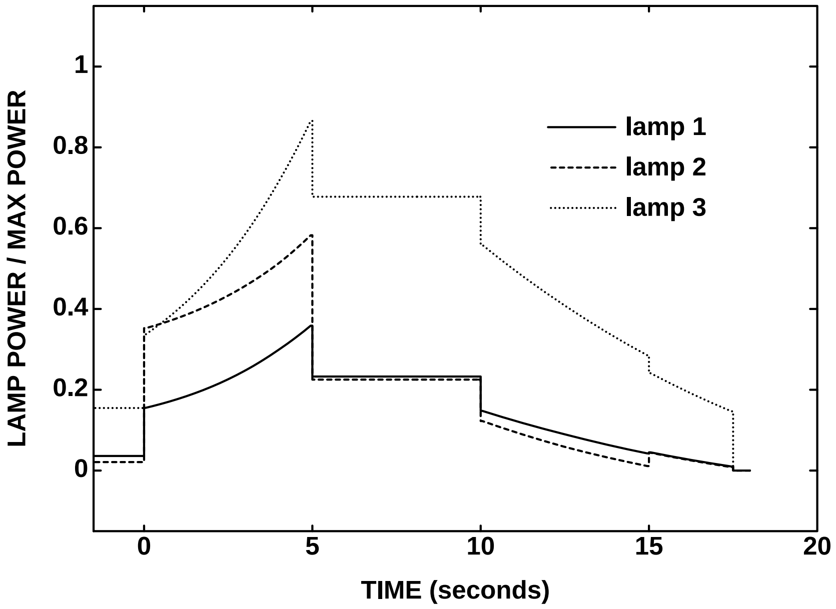


Figure 11: Lamp powers for Trajectory T using heuristic scheme.

yield stress has a peak value of 0.45, occurring just at the end of ramp-up, so the appearance of slip defects in the wafer is unlikely, although a wider safety margin is probably desirable. It is up to a process engineer to decide whether the temperature uniformity is adequate for uniformity of processing.

These results demonstrate that real-time independent variation of lamp powers allows much better uniformity during transients than does control with a fixed flux pattern. Numerous reports — for example, [Lor88, HJB88, KB89, YM90, Gyu91] — note that with a flux pattern tuned for a near-uniform steady-state temperature profile the wafer edge will be typically be tens of degrees hotter than the wafer center during fast ramp-up. The problem is that during ramp-up the total power applied is perhaps two or three times the total power applied for steady-state hold, so the compensation for edge losses is undesirably multiplied by two or three as well, resulting in serious over-compensation; the problem is compounded because edge losses are small at the low temperatures at which ramp-ups begin. Alternatively, if the flux pattern is adjusted for near-uniformity during ramp-up then the edge will be tens of degrees colder than the center in steady-state [KCL91].

5.4 Exact optimization using model

A valid criticism of the scheme (23) is that since it is merely heuristic, how does one know how well it works compared to the best scheme possible? This question becomes extremely important if the performance using the heuristic scheme is marginal or unacceptable. It would be extremely useful to know the exact solution to the problem of minimizing the quantity (21).

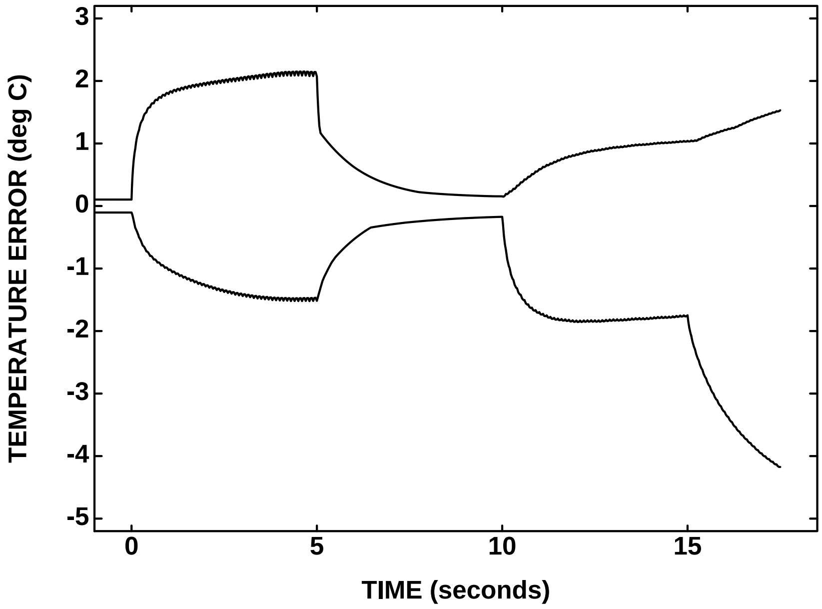


Figure 12: Errors using heuristic lamp power scheme.

Table 1: Errors and nonuniformities using heuristic lamp power scheme.

part	$\max_{i,t} \left(T_i(t) - T^{\text{ref}}(t)\right)$ °C	$\min_{i,t} \left(T_i(t) - T^{\text{ref}}(t)\right)$ °C	$\max_{i,j,t} \left(T_i(t) - T_j(t)\right)$ °C
1	2.2	-1.6	3.7
2	2.1	-1.6	3.7
3	1.1	-1.9	2.8
4	1.6	-4.2	5.7

part 1: 600°C up to 1100°C at 100 °C/s
part 2: hold at 1100°C
part 3: 1100°C down to 900°C at
-40 °C/s
part 4: 900°C down to 800°C at -40 °C/s

In this subsection a technique is presented for solving the following problem, which is a discretization in time of the problem of minimizing (22), that is, a discretization in time and space of the problem of minimizing (21).

$$\begin{array}{c}\text{Given } \mathbf{T}_0, \text{ find } \mathbf{P}_0, \mathbf{P}_1, \ldots \mathbf{P}_{N-1} \text{ to minimize} \\ \max\limits_{n=1,2,\ldots,N} \left\| \mathbf{T}_n - \mathbf{T}_n^{\text{ref}} \right\|_\infty \\ \text{subject to, for } n = 0, 1, \ldots, N-1, \\ \mathbf{0} \leq \mathbf{P}_n \leq \mathbf{P}^{\max}, \end{array} \tag{25}$$

where $\mathbf{T}_n$ is the temperature profile at time $t_o + n\Delta t$ resulting from the initial condition $\mathbf{T}_0$ and application of the lamp power sequence $\mathbf{P}_0$ through $\mathbf{P}_{n-1}$.

Define $\mathbf{T}_n^{\text{zero}}(t)$ to be the value at time t of the solution to (11) with initial condition $\mathbf{T}(t_o + n\Delta t) = \mathbf{T}_n^{\text{ref}}$ and with $\mathbf{P}(t)$ and $\mathbf{q}^{\text{dist}}(t)$ equal to zero for $t_o + n\Delta t \leq t \leq t_o + (n+1)\Delta t$. If $\mathbf{T}(t_o + n\Delta t)$ is close to $\mathbf{T}_n^{\text{ref}}$ then (11) can be linearized as follows:

$$\frac{d}{dt}\left(\mathbf{T} - \mathbf{T}_n^{\text{zero}}\right) \approx F_n \cdot \left(\mathbf{T} - \mathbf{T}_n^{\text{zero}}\right) + L\mathbf{P}_n, \tag{26}$$

where

$$F_n = D_{\mathbf{T}}\mathbf{q}\left(\mathbf{T}_n^{\text{ref}}\right)\left(C(\mathbf{T}_n^{\text{ref}})\right)^{-1}$$

where in turn, from (10), neglecting $D_{\mathbf{T}}K^{\text{cond}}$,

$$D_{\mathbf{T}}\mathbf{q}(\mathbf{T}) = 4K^{\text{rad}}\operatorname{diag}\left((\mathbf{T})^3\right) + K^{\text{cond}}(\mathbf{T}) + K^{\text{conv}}.$$

Define the error vector $\mathbf{E}_n$ at sample time n as

$$\mathbf{E}_n = \mathbf{T}_n - \mathbf{T}_n^{\text{ref}}$$

and the vector $\mathbf{T}_n^{\text{cool}}$ as

$$\mathbf{T}_n^{\text{cool}} = \mathbf{T}_n^{\text{zero}}\left(t_o + (n+1)\Delta t\right) - \mathbf{T}_n^{\text{ref}}.$$

Then solution of (26) yields

$$\mathbf{E}_{n+1} = \Phi_n\mathbf{E}_n + \Gamma_n\mathbf{P}_n + \mathbf{T}_n^{\text{cool}} + \mathbf{T}_n^{\text{ref}} - \mathbf{T}_{n+1}^{\text{ref}}, \tag{27}$$

where

$$\Phi_n = \exp(F_n\Delta t) \quad \text{and} \quad \Gamma_n = \int_0^{\Delta t} \exp(F_n\tau)\, d\tau,$$

with $\exp(\cdot)$ being the matrix exponential function.

The solution to the recursion (27) is, for $n = 2, \ldots, N-1$,

$$\begin{aligned} \mathbf{E}_{n+1} \quad = \quad & \left(\prod_{m=0}^{n} \Phi_m\right)\mathbf{E}_0 \\ & + \sum_{l=0}^{n-1}\left\{\left(\prod_{m=l+1}^{n} \Phi_m\right)\left(\Gamma_l\mathbf{P}_l + \mathbf{T}_l^{\text{cool}} + \mathbf{T}_l^{\text{ref}} - \mathbf{T}_{l+1}^{\text{ref}}\right)\right\} \\ & + \Gamma_n\mathbf{P}_n + \mathbf{T}_n^{\text{cool}} + \mathbf{T}_n^{\text{ref}} - \mathbf{T}_{n+1}^{\text{ref}}. \end{aligned} \tag{28}$$

Define

$$A = \begin{bmatrix} A_{1,1} & 0 & \cdots & 0 \\ A_{2,1} & A_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{N,1} & A_{N,2} & \cdots & A_{N,N} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_N \end{bmatrix},$$

where for each n,

$$A_{n,n} = \Gamma_{n-1};$$

for each n and $p < n$,

$$A_{n,p} = \left(\prod_{m=p}^{n-1} \Phi_m \right) \Gamma_{p-1};$$

for each $n \geq 2$,

$$\begin{aligned} \mathbf{b}_n &= \left(\prod_{m=0}^{n-1} \Phi_m \right) \mathbf{E}_0 \\ &\quad + \sum_{l=0}^{n-2} \left\{ \left(\prod_{m=l+1}^{n-1} \Phi_m \right) \left(\mathbf{T}_l^{\text{cool}} + \mathbf{T}_l^{\text{ref}} - \mathbf{T}_{l+1}^{\text{ref}} \right) \right\} \\ &\quad + \mathbf{T}_{n-1}^{\text{cool}} + \mathbf{T}_{n-1}^{\text{ref}} - \mathbf{T}_n^{\text{ref}}; \end{aligned}$$

and

$$\mathbf{b}_1 = \Phi_0 \mathbf{E}_0 + \mathbf{T}_0^{\text{cool}} + \mathbf{T}_0^{\text{ref}} - \mathbf{T}_1^{\text{ref}}.$$

Using (27) and (28) the problem (25) can now be expressed as a constrained minimax error problem:

$$\begin{aligned} &\text{Minimize} \quad \left\| \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_N \end{bmatrix} \right\|_\infty = \left\| A \begin{bmatrix} \mathbf{P}_0 \\ \vdots \\ \mathbf{P}_{N-1} \end{bmatrix} + \mathbf{b} \right\|_\infty \\ &\text{subject to} \quad \mathbf{0} \leq \begin{bmatrix} \mathbf{P}_0 \\ \vdots \\ \mathbf{P}_{N-1} \end{bmatrix} \leq \begin{bmatrix} \mathbf{P}^{\max} \\ \vdots \\ \mathbf{P}^{\max} \end{bmatrix}. \end{aligned} \tag{29}$$

5.5 Optimal transients for System S

This subsection presents results from application of the method given in Section 5.4 to two different problems concerning System S. The first problem is that of checking the performance of the suboptimal, heuristic scheme used in Section 5.3 for following Trajectory T. The second problem is an investigation of the trade-off between nonuniformity and heating or cooling rate when the wafer temperature is between 1000°C and 1100°C. For both problems the gas ambient is assumed to be N_2 at 1 atm.

To check the results of Section 5.3 the LP's (29) were set up and solved for the following two cases:

Table 2: Results with optimal lamp power scheme for Trajectory T.

start temp.	ramp rate	ramp time	LP prediction		simulation result	
°C	°C/s	s	WAE	WNU	WAE	WNU
600	100	5.0	1.6	3.2	1.7	3.2
1100	−40	7.5	1.5	3.0	1.8	3.1

WAE (worst absolute error): $\max_{i,t} \left| T_i(t) - T^{\mathrm{ref}}(t) \right|$, in °C

WNU (worst nonuniformity): $\max_{i,j,t} \left(T_i(t) - T_j(t) \right)$, in °C

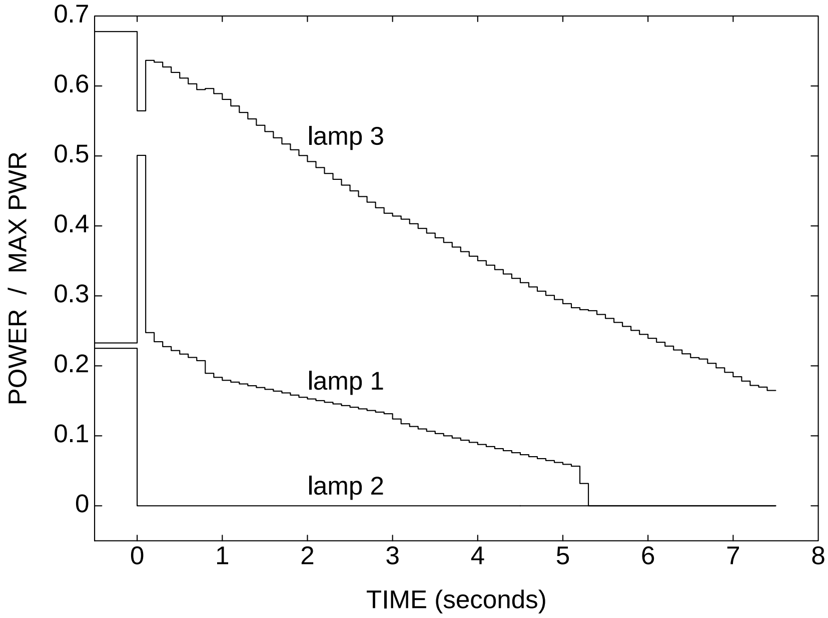
- ramp-up at 100 °C/s for 5 s beginning with optimal steady-state hold at 600°C
- cool-down at 40 °C/s for 7.5 s beginning with optimal steady-state hold at 1100°C.

The sample rate in each case was 10 Hz.

To check the accuracy of the linearization (26), nonlinear simulations were run with the lamp power sequences generated by the LP's. Results are summarized in Table 2. It can be seen that the predictions of nonuniformity obtained from the LP solutions are very accurate. The LP predictions of worst absolute error are too low by a few tenths of a degree. These few tenths of a degree result from a disagreement of that size between the LP-predicted temperature profiles and the corresponding temperature profiles from nonlinear simulation. The disagreement is nearly uniform across the wafer — relative to the hundreds of degrees of temperature change undergone by the wafer the bias is less than 0.1%.

Comparison of Tables 1 and 2 shows that the ramp-up using the heuristic scheme is within 0.5°C of being optimal but that the nonuniformity during cool-down achieved using the heuristic scheme is almost twice as big as necessary. The problem with the heuristic cool-down is likely the *ad hoc* approach used to deal with the constraint that all lamp powers be nonnegative.

Figure 13 shows the lamp settings used for optimal cool-down along with a mesh plot of temperature error as a function of time, t (with $t = 0$ at the beginning of the cool-down; lamp settings for $t < 0$ are for optimal steady-state hold at 1100°C), and radial position, r. Until $t = 5.3$ s the wafer center is one of the two hottest radial positions on the wafer. At that point the center lamp power, already heading for zero, is cut off entirely and the wafer center begins to cool relative to the rest of the wafer; the timing is such that at $t = 7.5$ s the wafer center is one of the two coldest radial positions on the wafer. Minimal error from the reference trajectory is obtained by using foresight about the impending non-availability of the center lamp.



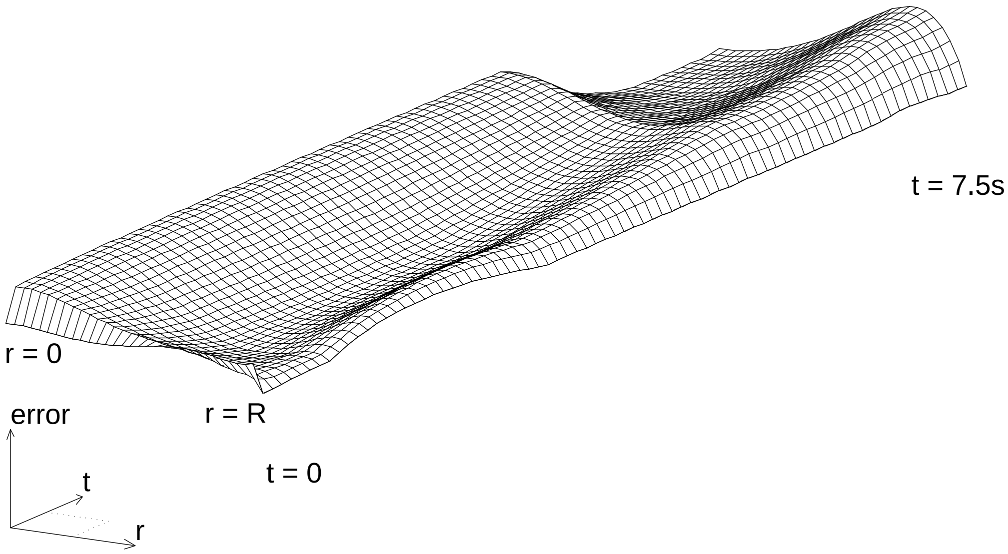


Figure 13: Optimal cool-down from 1100°C to 800°C at 40 °C/s.

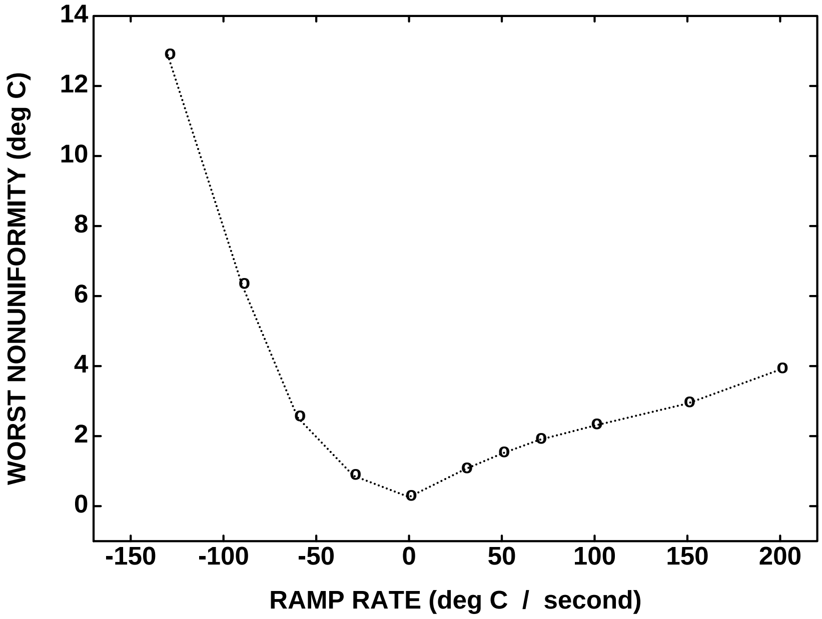


Figure 14: Uniformity as a function of ramp rate, from Table 3.

To investigate the trade-off between rate of temperature change and best possible nonuniformity, minimum temperature error lamp trajectories were computed for the following temperature trajectories:

- six ramp-ups, at rates from 30 °C/s to 200 °C/s, all starting near 1000°C and ending near 1100°C
- four constant rate cool-downs, at rates from 30 °C/s to 130 °C/s, all starting near 1100°C and ending near 1000°C.

All initial conditions were optimal steady-state hold. The temperature trajectories resulting from the LP solutions were obtained by nonlinear simulation. Results are summarized in Table 3. For purposes of comparison, data for optimal steady-state hold at 1050°C also appears in the table. The comments made earlier about the slight disagreements between LP-predicted numbers and the corresponding numbers determined by simulation apply here too.

Figure 14 shows the last column of Table 3 plotted against the second column. It can be seen that optimal nonuniformity increases much more quickly with cooling rate than it does with heating rate. The asymmetry arises because for fast cooling very little heat can be supplied to the wafer or else the entire wafer temperature profile will not be able to fall as quickly as is required — from a mathematical point of view, negative lamp power settings would be useful, but such negative settings are of course physically impossible. This result indicates that while good wafer temperature uniformity in System S is achievable during fast heating, it cannot be achieved during fast cooling.

Table 3: Results with optimal lamp power schemes for various ramp rates.

start temp.	ramp rate	ramp time	LP prediction		simulation result	
°C	°C/s	s	WAE	WNU	WAE	WNU
1000	200	0.5	2.0	3.9	2.0	3.9
995	150	0.7	1.4	2.9	1.5	2.9
1000	100	1.0	1.1	2.2	1.2	2.3
995	70	1.5	0.9	1.8	1.1	1.9
1000	50	2.0	0.7	1.4	1.0	1.5
995	30	3.4	0.5	1.0	0.9	1.1
1050	0	(∞)	—	—	0.13	0.26
1102	−30	3.4	0.4	0.8	0.8	0.9
1102	−60	1.7	1.3	2.5	1.5	2.5
1099	−90	1.1	3.2	6.3	3.3	6.3
1104	−130	0.8	6.4	12.8	6.5	12.9

WAE (worst absolute error): $\max\limits_{i,t} \left| T_i(t) - T^{\mathrm{ref}}(t) \right|$, in °C

WNU (worst nonuniformity): $\max\limits_{i,j,t} \Big(T_i(t) - T_j(t) \Big)$, in °C

6 Control Schemes Incorporating Feedback

In Sections 4 and 5 methods for obtaining minimal steady-state and transient temperature error for systems that are essentially disturbance-free. It is unlikely that an actual RTP system will be so free of disturbances that purely open-loop control schemes will work as well as they do for System S. This section will briefly discuss the use of feedback from temperature sensors to ensure small temperature errors in the face of substantial disturbances. A thorough treatment of feedback control for RTP is beyond the scope of this paper but more extensive discussion of the topic will appear in other publications by the author and colleagues.

6.1 General comments on feedback schemes for RTP

Figure 15 is a block diagram of an RTP temperature control scheme that incorporates both open-loop transient control as described in Section 5 and feedback control. The lamp power signal vector sent to the lamp power supplies is the sum of $\mathbf{P}^{\mathrm{OL}}$, the lamp power signal vector generated by the open-loop controller, and $\mathbf{P}^{\mathrm{FB}}$, the lamp power signal vector generated by the error controller. The intended effect of adding the error controller is to nearly eliminate the drift of the wafer temperature profile away from T^{ref} that would probably occur if open-loop control only was used. This kind of control architecture is widely used in such fields as robotics and aerospace guidance. For RTP

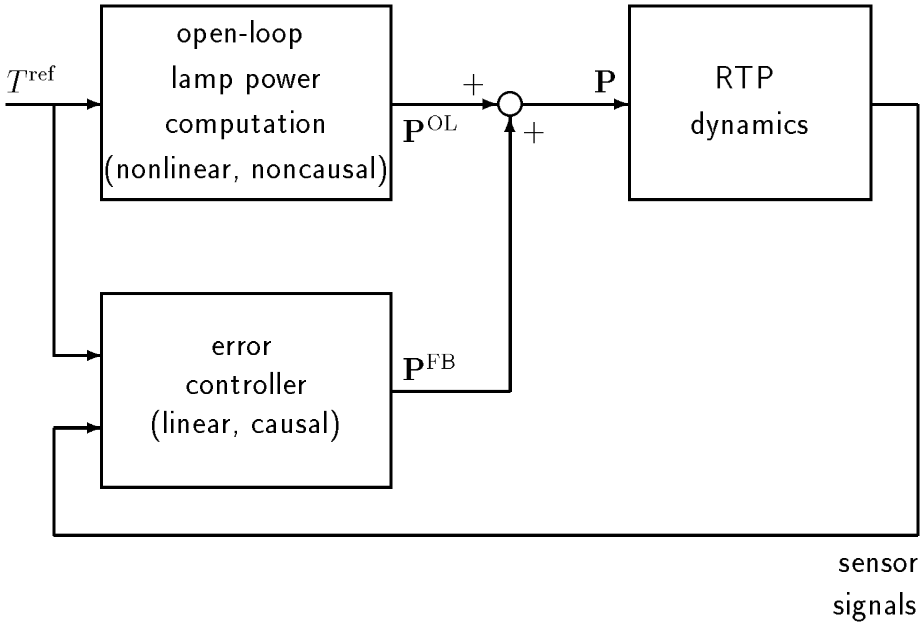


Figure 15: Feedback added to system for disturbance rejection.

temperature control, the open-loop lamp power computation might be based on the heuristic scheme of Section 5.2 or it might be done some other way. The design of the feedback controller will be discussed briefly later in this subsection.

The scheme outlined above can be contrasted with the scheme illustrated in Figure 16. The latter scheme could be called *feedback-only* control; in this scheme, the most common one in many industrial process control applications, a controller K processes the differences between $T^{\rm ref}$ and sensor signals and generates the lamp power signal vector $\mathbf{P}$. There are two main reasons why

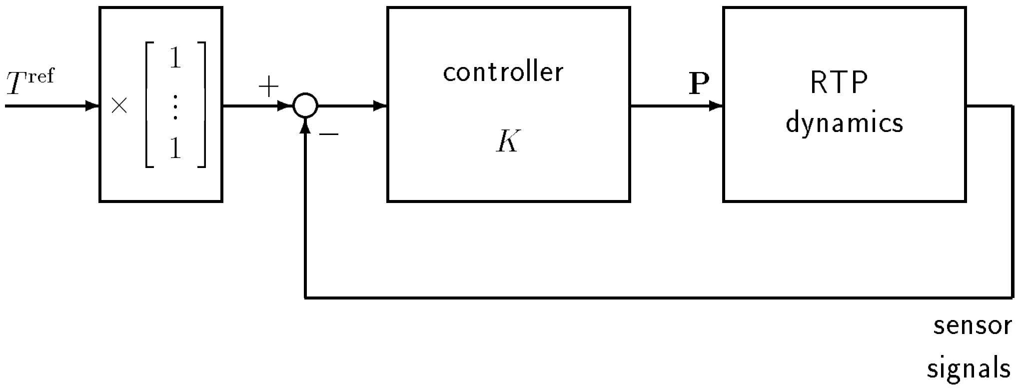


Figure 16: Feedback-only control.

the performance obtained with an RTP temperature control implementation of such a scheme would probably be inferior to that obtained with a well-designed system incorporating both open-loop and feedback control.

1. In RTP the entire trajectory $T^{\rm ref}$ is known before execution starts, but since a feedback-only scheme is entirely causal it can do nothing to anticipate the future behavior of $T^{\rm ref}$. An analogy is this: Feedback-only control is like driving a car looking not through the front windshield but only in the rear-view mirror.[9]

2. Feedback-only control requires the design of a controller that deals sensibly both with errors of hundreds of degrees and with errors of one or two degrees. Clearly designing such a controller is a more difficult task than designing an error controller that never sees errors larger than a few degrees.

Before discussing error controller design, the topic of temperature sensors should be mentioned. The problem of wafer temperature measurement in RTP systems is a difficult one because it is desirable to have as much of the temperature sensing equipment outside of the chamber as possible. Many measurement schemes have been proposed [RP90]. The leading candidate is pyrometry, but the problems of calibration for changing wafer emissivity and of rejection of light pollution from lamps have not yet been solved for all RTP systems and processes. The discussion that follows assumes that near-perfect temperature measurement will eventually be available, but it may turn out in practice that error controller designs will have to take sensor imperfections into account.

Nonuniformity cannot be corrected unless it is known whether points at various positions on the wafer are too hot or too cold relative to the wafer center, so measurement of temperature at multiple points on the wafer is necessary for effective feedback control of uniformity. A change in the power setting of any particular lamp zone will affect the entire wafer temperature profile. Thus it makes sense that a temperature error detected in a certain region on the wafer cause changes in *all* of the lamp zone settings. For example, if the wafer were too cold near its edge, power to lamp zones preferentially heating the edge region should of course be increased, but in addition power to lamp zones preferentially heating the region near the wafer center should be simultaneously decreased to prevent the entire wafer temperature profile from drifting upward. Modern multivariable control design techniques can be used for the synthesis of such control schemes — there is no need to pair each sensor with a single lamp zone. If such pairing is not done there is no reason for the number of sensors to be equal to the number of lamp zones. Information about the wafer temperature profile increases with the number of sensors until there are enough sensors to infer the exact profile; investigation

[9]The author thanks G. Franklin for bringing this analogy to his attention.

of the trade-off between sensor cost and system performance is well worth the effort.

6.2 A Simple Error Controller for System S

This subsection gives some evidence that open-loop-plus-feedback control can be used to ensure small temperature errors if some model parameters are incorrect. To do this, simulations were run on a variation of System S in which

- the emissivity, and hence as well the absorptivity, of the side of the wafer facing the lamps is 0.5 instead of 0.6; and
- convective loss 33% higher than that in System S

Two simulations were run using the same open-loop lamp setting sequence; one with an error controller running and one without. In each case the goal was to follow the first two parts of Trajectory T — ramp-up from 600°C and 1100°C and steady-state hold at 1100°C. The open-loop lamp sequence used was the appropriate part of the sequence in Section 5.3 — a sequence heuristically designed for System S with its original parameters — with the steady-state hold portion extended five seconds.

The dashed lines in Figure 17 are taken from the no-error-controller simulation and show the quantities $\max_i \left(T_i(t) - T^{\text{ref}}(t)\right)$ and $\min_i \left(T_i(t) - T^{\text{ref}}(t)\right)$. The wafer temperature profile for $t \leq 0$, *i.e.*, the initial condition for the simulation, is the steady-state temperature profile obtained in the altered-parameter system when the optimal 600°C steady-state setting for the original-parameter system is applied. It is clear from the plot that gross errors in trajectory-following result from the extra convective cooling and reduced absorption of lamp radiation.

To simulate a system with an error controller it was assumed that perfect measurements of the four element temperatures T_2, T_9, T_{15}, and T_{19} would be available at sample times $t_o + n\Delta t$, for $n = 0, 1, 2, \ldots$, and that $\mathbf{P}_n^{\text{FB}}$, the value of the feedback lamp setting signal vector over the time interval $t_o + n\Delta t \leq t_o + (n+1)\Delta t$, could be computed virtually instantly at time $t_o + n\Delta t$. The feedback control law used was

$$\mathbf{P}_n^{\text{FB}} = -\begin{bmatrix} 230 & -17 & -150 & -40 \\ -280 & 690 & 230 & -410 \\ -530 & -2260 & 420 & 2970 \end{bmatrix} \left(\begin{bmatrix} T_2 \\ T_9 \\ T_{15} \\ T_{19} \end{bmatrix} - T_n^{\text{ref}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right).$$

The numbers are presented to demonstrate that an error at any one sensor location has an immediate effect on all of the lamp settings. Details on the development of the control law will appear in a paper currently in preparation by the author and colleagues.

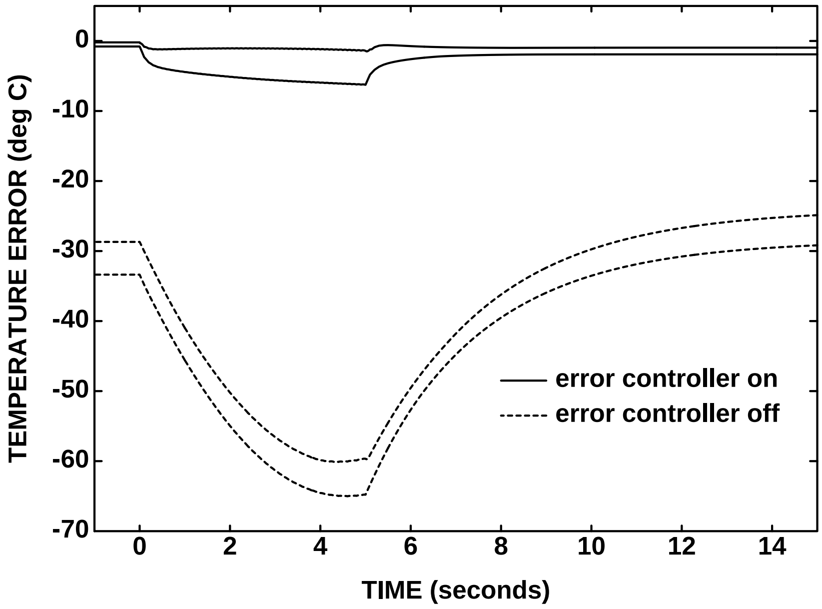


Figure 17: Maximum and minimum differences between wafer temperature and T^{ref} for System S with altered parameters.

The initial condition for the error-controller-on simulation was the steady-state temperature profile obtained with the open-loop lamp signal vector constant at the optimal 600°C steady-state setting for the original-parameter system and the error controller running. The values of $\max_i \left(T_i(t) - T^{\mathrm{ref}}(t)\right)$ and $\min_i \left(T_i(t) - T^{\mathrm{ref}}(t)\right)$ for this simulation are plotted as solid curves in Figure 17. The worst temperature error and the worst temperature nonuniformity during ramp-up are -6.2°C and 4.9°C; in steady-state these values are -1.9°C and 1.0°C. Simulations of the same type with various different changes in the parameters of System S produced similar results.

This example is intended only to demonstrate that open-loop-plus-feedback control will possibly be very effective in assuring uniformity and good trajectory-following even when there are large errors in model parameters. Obviously a single idealized example cannot stand as a thorough proof of concept. Nevertheless it is worth noting that a very simple linear feedback control law exhibits very good performance and no stability problems over a wide range of wafer temperatures, despite the nonlinearity of RTP temperature dynamics.

6.3 Extensions

Practical error controller designs will be somewhat more complex than the simple constant matrix used in Section 6.2. The following features might be added to improve performance:

- Integral action — integrating one or more temperature errors (as in PID control) will assure zero steady-state offset from a temperature setpoint at one or more points on the wafer.
- Gain scheduling — radiative heat loss depends nonlinearly on temperature, so different control laws may work better in different temperature regions.
- Sensor noise filtering — sensor signals may be noisy.
- Clipping of components of the vector $\mathbf{P}^{\mathrm{FB}}$ — excessively large signals should not be used when temperature errors are large, which might be the case when the error controller is first turned on.

7 Discussion

7.1 Why solve optimal control problems?

After looking at Section 6, a practically-minded reader might ask, "Even with feedback the temperature errors due to modeling error are quite a bit bigger than the predicted optimal errors, so what is the point of using a detailed model and much computation to determine exactly optimal performance, as suggested in Section 5?"

An answer to this question with regard to the specific case of the System S simulation results is that the optimization results do give useful quantitative information. A control system designer might be disappointed with the 4.9°C of nonuniformity occurring during ramp-up with the error controller running in Section 6.2 if he or she thought that less than 1°C was achievable, but might be more satisfied if he or she were aware that the *best possible* nonuniformity for the same trajectory in a very similar system is 3.2°C.

A more general version of this answer is that optimization is extremely valuable because it can determine limits of system performance. If the RTP thermal model used to determine optimal steady-state and transient system performance comes reasonably close to describing the behavior of a real RTP system, the performance limits established with the model will almost certainly apply to the real system too, no matter how many sensors and no matter what control law is used for the real system.

Analysis of this type may save time and money in RTP system design. If a good model can be determined by measuring the responses of a prototype system, or, even better, if a good model can be developed without building anything at all, immediate decisions can be made about whether or not lamp or chamber design modifications are necessary. For example if it is determined with a model for a certain system that the smallest possible nonuniformity during steady-state hold at a given temperature setpoint and processing environment is 10°C, it is time to redesign the lamp array and/or the processing chamber. These kinds of decisions are much harder to make with a trial-and-error approach to control design — in a trial-and-error approach there is always the hope that the next adjustment to the control law is the one that will make the system work.

Determining limits of performance also helps in temperature trajectory design. Quantitative information about the trade-offs of ramp-up or cool-down rate against uniformity, such as that developed for System S in Section 5.5, would be valuable for process engineers trying to develop feasible recipes for specific RTP equipment.

An entirely different reason for finding optimal control strategies is that the results may provide nonintuitive suggestions for control system design. For example, consider the solution shown in Section 5.5 to the problem of optimal cool-down of the System S wafer from 1100°C to 800°C at 40 °C/s, in which an interesting strategy is used to accommodate the constraint that the lamp power settings always be nonnegative.

7.2 Extensions of open-loop optimization schemes

Many useful variations of the optimization problems formulated in Sections 4 and 5 are no more difficult to solve than the original problems. Space limitations prevent complete descriptions in this paper, but a brief listing is provided here.

Optimization of temperature error over part of the wafer subject to a constraint on temperature error over the rest of the wafer. If it is found to be impossible to obtain adequate temperature uniformity for uniform processing across the entire wafer it might be useful to solve a problem of the following type:

$$\begin{array}{ll} \text{Minimize} & \displaystyle\max_{0\le r\le 0.9R,\ t_o\le t\le t_f} \left|T(r,t)-T^{\text{ref}}(t)\right| \\ \text{subject to} & \displaystyle\max_{0.9R< r\le R,\ t_o\le t\le t_f} \left|T(r,t)-T^{\text{ref}}(t)\right| \le 5^\circ\text{C}. \end{array}$$

The numbers $0.9R$ and 5°C are of course only examples. A similar variation of the minimax steady-state error problem can be formulated too.

Temperature-weighted error cost. Since yield stress decreases as wafer temperature increases and since most processing occurs at the highest temperatures during an RTP cycle, having small temperature nonuniformity is more important at higher temperatures than at lower temperatures. Hence the problem

$$\text{Minimize} \quad \max_{0\le r\le R,\ t_o\le t\le t_f} w\left(T^{\text{ref}}(t)\right)\cdot\left|T(r,t)-T^{\text{ref}}(t)\right|,$$

where $w(\cdot)$ is a positive weighting function that would typically take larger values at higher temperatures.

Artificial bounds on power signals. Consider the following problem:

$$\text{Minimize} \quad \max_{0\le r\le R,\ t_o\le t\le t_f} \left|T(r,t) - T^{\mathrm{ref}}(t)\right|$$

$$\text{subject to} \quad 0.05 \begin{bmatrix} P_1^{\max} \\ \vdots \\ P_J^{\max} \end{bmatrix} \le \mathbf{P}(t) \le 0.95 \begin{bmatrix} P_1^{\max} \\ \vdots \\ P_J^{\max} \end{bmatrix}.$$

The constraints prevent the lamp power settings from pushing against their absolute lower and upper bounds — if the lamp power trajectory obtained from solving this problem were used as the open-loop part of an open-loop-plus-feedback control scheme, then it would be always be possible for the error controller to add small negative or positive corrections to the lamp power signals.

Incorporation of dynamic models for lamps. In this paper it has been assumed that the power radiated by each lamp can be changed instantly. This is a reasonable assumption for arc lamps, but not for incandescent lamps. A dynamic model for incandescent lamps is this:

$$\begin{aligned} \dot{T}_{\mathrm{lamp}} &= -\alpha_1 T_{\mathrm{lamp}}^4 - \alpha_2 T_{\mathrm{lamp}} + \beta P_{\mathrm{elec}} \\ P_{\mathrm{rad}} &= \gamma T_{\mathrm{lamp}}^4, \end{aligned} \tag{30}$$

where T_{lamp} is the filament temperature, P_{elec} the electric power supplied to the lamp, P_{rad} the power radiated by the lamp, and α_1, α_2, β and γ are positive constants; for tungsten-halogen lamps the settling time of P_{rad} in response to a step change in P_{elec} is a few tenths of a second. The model (30) can be merged with the thermal model of Section 2; using the merged model, P_{elec} trajectories for optimal transient temperature response can be determined using a suitably modified version of the method given in Section 5.

8 Conclusions

In this paper it has been shown that using a detailed thermal model for an RTP system, the problems of minimizing the worst error in the wafer temperature profile during steady-state hold and during temperature trajectory following can be formulated as easily-solved linear programming problems (LP's). Simulations using an idealized RTP system illustrated how these LP formulations can be applied. For the steady-state case it was shown how to set up an LP to minimize error without a detailed system model. A heuristic scheme for trajectory-following was presented and shown to perform near-optimally in the idealized system. The idea of using open-loop-plus-feedback control to ameliorate the effects of disturbances and modeling errors was introduced.

Probably the most useful practical application of the optimization techniques presented in this paper is in determination of limits of system performance. These techniques provide answers to questions such as, “How good is the performance of my RTP temperature control system compared to the best possible performance?" and, "Is the temperature control I need for my process achievable on this RTP system?" These are questions that a trial-and-error design method can answer only very slowly, if at all.

In addition to the points made explicitly in the paper, a significant result is implicit in the data from the simulations. For some axisymmetric RTP systems with the properties (a) that there are multiple independently-controlled zones and (b) that each lamp zone casts a broad, not-carefully-tuned pattern of radiation across the wafer, steady-state temperature nonuniformity less than 1°C and transient temperature nonuniformity less than about 5°C are achievable over a wide range of wafer temperature trajectories. Experimental confirmation for the steady-state case has been reported at Stanford [AWSM91], at Texas Instruments [MKY91], and probably at other sites as well. Experimental confirmation for the transient case is anticipated soon.

Acknowledgments

This work would not have been possible without help from many sources. The author gratefully acknowledges the assistance provided by the following people: Profs. S. Boyd, K. Saraswat, G. Franklin, and T. Kailath, for support and considerable guidance; Dr. C. Schaper and Mr. Y. M. Cho for lengthy and fruitful discussions of problems in RTP temperature control; Dr. M. Moslehi, Mr. P. Apte, and Mr. S. Wood for helpful remarks about RTP; and Mr. P. Gyugyi and Mr. L. Booth for experimental work on the Stanford Rapid Thermal Multiprocessor.

References

[ÅW90] K. J. Åström and B. Wittenmark. *Computer Controlled Systems: Theory and Design.* Prentice-Hall, 2nd edition, 1990.

[AWSM91] P. P. Apte, S. Wood, K. C. Saraswat, and M. M. Moslehi. Temperature uniformity optimization using three-zone lamp and dynamic control in a rapid thermal multiprocessor. In *Mater. Res. Soc. Proc.*, volume 224. Materials Research Society, 1991.

[BCD84] G. Bentini, L. Correra, and C. Donolato. Defects introduced in silicon wafers during rapid thermal annealing: Thermoelastic and thermoplastic effects. *J. Appl. Phys.*, 56(10):2922–2929, November 1984.

[CAK+91] S. A. Campbell, K.-H. Ahn, K. L. Knutson, B. Y. H. Liu, and J. D. 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.

[Chv83] V. Chvátal. *Linear Programming.* W. H. Freeman, 1983.

[Dan63] G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, 1963.

[FPW90] G. F. Franklin, J. D. Powell, and M. L. Workman. *Digital Control of Dynamic Systems.* Addison Wesley, 2nd edition, 1990.

[GHM+86] P. E. Gill, S. J. Hammarling, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for LSSOL (Version 1.0): A FORTRAN package for constrained least-squares and convex quadratic programming. Technical Report SOL 86-1, Operations Research Dept., Stanford University, Stanford, California 94305, January 1986.

[GRS91] R. S. Gyurcsik, T. J. Riley, and F. Y. Sorrell. A model for rapid thermal processing: Achieving uniformity through lamp control. *IEEE Trans. Semicond. Manufact.*, 4(1):9–13, February 1991.

[Gyu91] R. S. Gyurcsik. Real-time control of rapid thermal processing systems. In *Proc. SRC Workshop on Real-Time Tool Controllers*, pages 109–133. Semiconductor Research Corporation, February 1991.

[HJB88] C. Hill, S. Jones, and D. Boys. Rapid thermal annealing — theory and practice. In R. A. Levy, editor, *Reduced Thermal Processing for ULSI*, pages 143–180. Plenum Press, 1988.

[Kak91] R. Kakoschke. Is there a way to a perfect RTP system? In *Mater. Res. Soc. Proc.*, volume 224. Materials Research Society, 1991.

[KB89] R. Kakoschke and R. Bussmann. Simulation of temperature effects during rapid thermal processing. In *Mater. Res. Soc. Proc.*, volume 146, pages 473–482. Materials Research Society, 1989.

[KCL91] K. L. Knutson, S. A. Campbell, and J. D. Leighton. Thermal effects of gasses in rapid thermal processing. In *Mater. Res. Soc. Proc.*, volume 224. Materials Research Society, 1991.

[KS88] H.-Matsuo Kagaya and T. Soma. Specific heat of Si. In *Properties of Silicon*, EMIS Datareviews Series, pages 35–36. INSPEC, The Institution of Electrical Engineers, 1988.

[Lor88] H. A. Lord. Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven. *IEEE Trans. Semicond. Manufact.*, 1(3):105–114, August 1988.

[Lue84] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Reading, Mass., 2nd edition, 1984.

[Mat90] MathWorks, Inc., The. *PRO-MATLAB User's Guide*. The MathWorks, Inc., 1990.

[MKY91] M. M. Moslehi, J. Kuehne, and R. Yeakley. In situ fabrication and process control techniques in rapid thermal processing. In *Mater. Res. Soc. Proc.*, volume 224. Materials Research Society, 1991.

[NSB91] S. A. Norman, C. D. Schaper, and S. P. Boyd. Improvement of temperature uniformity in rapid thermal processing systems using multivariable control. In *Mater. Res. Soc. Proc.*, volume 224. Materials Research Society, 1991.

[RP90] F. Roozeboom and N. Parekh. Rapid thermal processing systems: A review with emphasis on temperature control. *J. Vac. Sci. Technol. B*, 8(6):1249–1259, November 1990.

[SC78] E. M. Sparrow and R. D. Cess. *Radiation Heat Transfer*. Hemisphere Publishing, 1978.

[SMG+89] K. Saraswat, M. Moslehi, D. Grossman, S. Wood, P. Wright, and L. Booth. Single wafer rapid thermal processing. In *Mater. Res. Soc. Proc.*, volume 146, pages 3–13. Materials Research Society, 1989.

[VMDK89] P. Vandenabeele, K. Maex, and R. De Keersmaecker. Impact of patterned layers on temperature non-uniformity during rapid thermal processing for VLSI applications. In *Mater. Res. Soc. Proc.*, volume 146, pages 149–160. Materials Research Society, 1989.

[Wyb88] M. N. Wybourne. Thermal conductivity of Si. In *Properties of Silicon*, EMIS Datareviews Series, pages 37–39. INSPEC, The Institution of Electrical Engineers, 1988.

[YM90] G. M. Young and K. A. McDonald. Effect of radiation shield angle on temperature and stress profiles during rapid thermal annealing. *IEEE Trans. Semicond. Manufact.*, 3(4):176–182, November 1990.