Model-based Control of Rapid Thermal Processing for Semiconductor Wafers

J. L. Ebert, D. de Roover, L. L. Porter II, V. A. Lisiewicz, S. Ghosal, R. L. Kosut, and A. Emami-Naeini

***Abstract-* This paper describes the application of model-based control system design techniques to Rapid Thermal Processing (RTP). The paper considers all aspects of the distributed temperature control problem from physics-based modeling to implementation of the real-time embedded controller. With its exceptionally stringent performance requirements (low non-uniformity of wafer temperature, high temperature ramp rates), RTP temperature control is a challenging distributed temperature control problem. Additionally, it is an important problem in the semiconductor industry because of the progressively smaller 'thermal budget' resulting from ever decreasing integrated circuit dimensions. Despite the emphasis on faster cold wall, single-wafer processing RTP chambers, the approach described here for solving distributed temperature control problems is equally applicable to slower distributed thermal systems such as, hot-wall batch-processing furnaces. For the physical model, finite volume techniques are used to develop high-fidelity heat transfer models that may be used for both control design and optimal chamber design. Model-order reduction techniques are employed to reduce these models to lower orders for control system design. In particular, principal orthogonal decomposition (POD) techniques have been used to derive low order models. Multivariable techniques such as LQG, H_2 / H_∞ methods are employed for feedback control design. These methods have been successfully implemented on commercial RTP chambers.**

I. Introduction

Many processes in semiconductor manufacturing require precise control of temperature across a wafer or a stack of wafers [1]. Generally, the equipment associated with thermal processing of (mostly silicon) wafers fall into two broad classes – batch furnaces and single-wafer systems. In batch furnaces, multiple wafers are loaded into quartz wafer holders, called "boats", and the entire stack of wafers is placed inside the furnace. In single wafer systems such as rapid thermal processing (RTP) systems [2], one wafer is processed at a time. In addition, furnace and single-wafer systems can be further classified as either hot-wall or cold-wall systems. Hot-wall systems maintain the walls of the chamber at a very high temperature, close to or above the processing temperature. In contrast, the chamber walls of cold-wall systems are water cooled, although it is not unusual to have some hot walls (usually quartz, silicon carbide, alumina, or graphite) in the chamber.

In the past, furnaces were often built with thick thermal insulation that minimized heat loss through furnace walls. The goal of such a design is to create a nearly isothermal environment within the furnace. A typical process would involve placing a boat of wafers inside the furnace, raising the temperature slowly (and isothermally) to process temperature, holding for a specified time, then slowly cooling the furnace. While such processes are still common today, increasing demands for better temperature uniformity and greater yield are driving equipment makers to address complications related to the dynamics of the heating and cooling processes. A furnace designed for temperature uniformity in steady-state operation will not, in general, have temperature uniformity during ramp. Also, since the wafer stack is predominantly heated from the outer edge, the wafer temperature uniformity is very dependent on factors such as ramp rate, wafer spacing, and other chamber design details. Furnace makers have had to group the heaters into multiple zones to maintain good temperature uniformity across the boat during ramp-up and ramp-down. The problem becomes one of multivariable distributed temperature control. In an effort to increase throughput, cold-wall furnaces offer the promise of faster dynamics, including faster cool-down rates. These requirements place an even higher burden on the design and control of multiple heater zones.

There are several reasons for the increasing popularity of single-wafer systems. The longer a wafer is kept at an elevated temperature, the higher the probability of defects. Most thermal processes performed in a furnace can be done in a single wafer system in much less time by processing the wafer at a higher temperature. The integral of temperature over time is called the thermal budget, and is significantly lower for RTP systems. Other factors such as diffusion of impurities and defects make control of the thermal budget important. With the advent of 300 mm diameter wafers, single wafer systems are becoming even more popular because a batch of fifty to hundred wafers,

We gratefully acknowledge the early support of DARPA, the Applied Computational Mathematics Program (ACMP), under the direction of Dr. James Crowley and Dr. Anna Tsao.

The authors are with SC Solutions, Inc., Sunnyvale, CA 94085. Corresponding author: J. L. Ebert, phone: 408 617 4520, fax: 408 617 4521, e-mail: ebert@scsolutions.com.

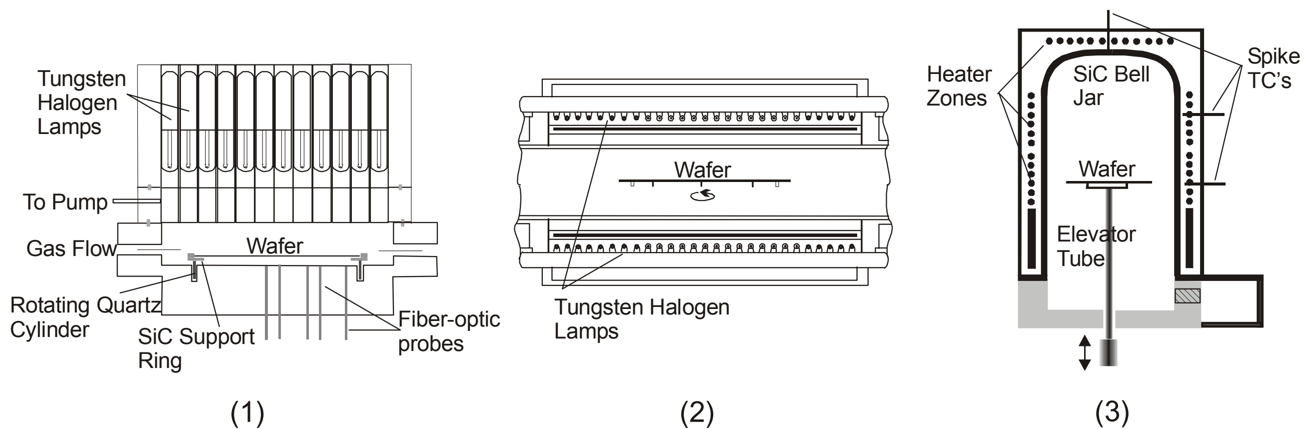

Figure 1: Three common commercial single-wafer RTP chamber configurations: (1) lamp heating from top, (2) lamp heating from two sides, and (3) hemispherical heating in hot wall chamber.

populated with complex integrated circuits, represents a substantial investment. Batch processing in a furnace incurs the risk of significant loss in case of an error or system failure. With proper monitoring, a single wafer system failure represents a smaller risk.

Because furnaces process many wafers, the net throughput of the single wafer system faces stiff competition from furnaces. RTP attempts to increase throughput by ramping at very high rates, e.g., 100–300°C per second and provides the capability for rapid thermal annealing (RTA). In addition, the temperature across the wafer is held uniform to within a few degrees during ramp-up, and to within 1–2°C process temperature. Such high performance is possible only with multivariable distributed temperature control. In this paper, we describe the design and implementation of a model-based control system for distributed temperature control of a single-wafer RTP chamber. The RTP system is highly nonlinear and actuator saturation is a common occurrence needing careful consideration. We show that a model-based technique is required to design an advanced multivariable controller to meet the performance requirements of stringent wafer temperature uniformity at all times, and very high wafer temperature ramp-up and ramp-down rates.

There are three primary designs of single-wafer RTP systems based on the heating method as shown in Figure 1. The first two are lamp-heated systems with top-side heating (e.g., manufactured by Applied Materials, Inc., and shown in Figure 2) and dual-side heating (manufactured by Mattson Technology, Inc.). The power delivered to the lamps is controlled dynamically to track the recipe temperature for the process while maintaining good temperature uniformity (temperature non-uniformity of less than 1-2°C across the wafer at process temperature is a common specification). In addition, there are hot-wall furnaces that impose wafer temperature uniformity by adjusting the power to the segmented wall heaters (produced by Axcelis, Inc.).

Figure 2: Applied Materials' RTP system (Courtesy Applied Materials).

The first step in the design of a model-based controller is the development of a thermal model which accurately captures the actual physical behavior of the system to be controlled. This high-fidelity thermal model is based on the application of the dynamic heat transfer equations to the system. The standard method of finding sufficiently accurate numerical solutions to these dynamic partial differential equations is by discretizing the spatial model state variables into many lumped elements or 'control volumes', and then iterating for the solutions of ordinary differential equations (ODE) while maintaining the boundary conditions at the desired values. This discretization is performed on all the components of the chamber (wafer, showerhead, walls, etc.) resulting in a large number of state variables that may add up to well over hundred. The model may contain physical variables whose values are not known in advance (e.g., heat transfer coefficients) and are identified from experimental data. A comparison of the model response with the actual system output provides a measure of model accuracy.

As a consequence of the discretization, the model often contains many state variables that are almost linearly dependent on each other. This dependence suggests that lower-order models may be found which approximate the model behavior quite well. Here, we have used the principal orthogonal decomposition (POD) method to reduce the number of state variables by a factor of four while retaining very good agreement with the high-fidelity model for all the important state variables (e.g., the wafer temperatures).

The next step in the design cycle is the development of a model-based controller. Using the model, we examined several advanced MIMO feedback control designs. The closed-loop system was then simulated using nonlinear simulation software (e.g., MATLAB®) to assess the relative merits of various candidate controllers. The optimal performance for the set of specifications (e.g., temperature uniformity, ramp-rate, overshoot) was obtained using the Linear Quadratic Gaussian (LQG) control design technique extended with frequency shaping. In addition, in some cases there is a need for addition of run-to-run control to deal with system nonlinearities.

Finally, an identical approach was used for developing a controller for a commercial chamber. The model-based controller for this commercial chamber was implemented on SC Solutions' control design and simulation software, which generated C-code for implementation on a computer which controls the equipment directly using a real-time operating system. The controller's performance on the actual equipment was then determined, and further tuning was carried out until desired performance was obtained.

The organization of this paper is as follows. In Section 2 we discuss the idea of the "generic RTP" system which is representative of the systems used in industry. Physical model development is discussed in Section 3. In Section 4 we describe the use of model order reduction methods. Section 5 details our model-based control design approach as well as controller implementation issues. Section 6 contains concluding remarks.

I. THE GENERIC RTP SYSTEM

As critical dimensions decrease, the temperature uniformity requirements become increasingly tight (~ ±1°C or less) and system design becomes increasingly coupled with the feedback control strategy. An accurate physical model of the system is valuable for evaluating system designs. The model also allows one to test control strategies and to evaluate the effect of design decisions on closed-loop performance.

The physical elements and important dimensional parameters for this generic RTP chamber are illustrated in Figure 3, and Table 1 and Table 2 [3]. It consists of a water-cooled cylindrical cavity with radius r_{wall} = 130mm and height $y_{top} - y_{exit} \approx$ 70mm. Five independently powered lamps are located near the top wall, and are modeled as axisymmetric rings at radii $r_1,\ldots,r_5$. A thick quartz window below the lamps divides the lamp cavity from the wafer cavity. A thinner quartz plate below the quartz window has thickness dy_{sh} and serves as a showerhead. Perforations in the plate at radii less than r_{sh} allow gas flow into the system. The silicon wafer is located below the showerhead at position y_{waf}. Only 200mm diameter standard wafers (r_{waf} 100mm) are considered here, with the SEMI M1.9-90 standard centerline thickness of $dy_{waf} = 725\mu$m ± 20μm and maximum thickness variation for a single wafer of 15μm. A guard ring near the edge of the wafer improves the temperature uniformity by limiting non-uniform edge losses. Finally, gas flows exit through a hole of radius r_{exit} through the bottom wall of the cavity. Here, we consider low-pressure operation where convection heat transfer due to gas flow is less important than radiation and conduction.

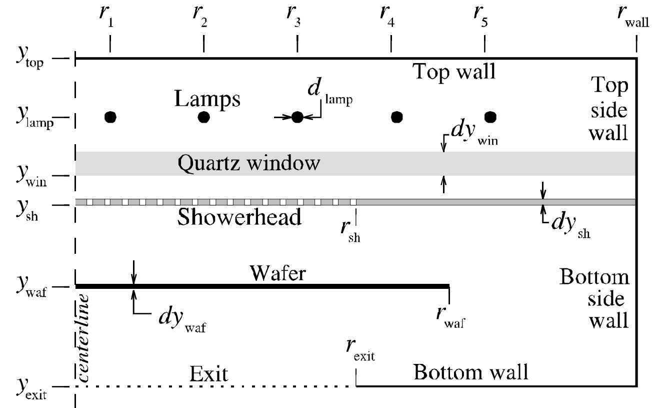

Figure 3: Schematic of the axisymmetric generic RTP system.

Table 1: Dimensions for the generic RTP system.

Dimension	Value(mm)	Dimension	Value(mm)
y_{exit}	-39.0	r_1	5.0
y_{waf}	0.0	r_2	32.5
y_{sh}	5.7	r_3	60.0
y_{win}	11.7	r_4	87.5
y_{lamp}	23.05	r_5	115.0
y_{top}	28.05	r_{wall}	130.0
dy_{waf}	0.725	r_{sh}	75.0
dy_{win}	6.35	r_{waf}	100.0
dy_{sh}	1.0	r_{exit}	75.0
d_{lamp}	2.0	$r_{grd,1}$	101.0
dy_{wall}	4.0	$r_{grd,1}$	111.0

Under normal operation the five lamps are independently powered to very high temperatures (1500–3000K) and emit radiation (predominantly at wavelengths < 4μm) that is transmitted through the quartz window and showerhead, and absorbed by the silicon wafer. As the wafer is heated, it loses power by conduction through the gas to the showerhead, and by radiation. Emission at wavelengths longer than approximately 4μm is absorbed by the showerhead. As the showerhead heats up it begins to emit longer wavelength radiation that slowly heats the thicker quartz window. Convective heat transfer

to the neighboring gas provides some cooling of the quartz window and showerhead. Additional cooling results from radial conduction through the quartz window and showerhead to the water-cooled cavity walls.

Table 2: Baseline input parameters for generic RTP model.

Parameter	Value	Units
Maximum power Lamp 1	500	W
Maximum power Lamp 2	3250	W
Maximum power Lamp 3	6000	W
Maximum power Lamp 4	8750	W
Maximum power Lamp 5	11500	W
Lamp filament solid fraction	0.3	
Conductivity from `WAF` to `SHR`	0	*a*
Cooling *h* for top of `WIN`	17.5	W/m^2K
Cooling *h* for bottom of `SHR`	17.5	W/m^2K
T for connective losses	300	K
Wafer emissivity	0.70, 0.70	*b*
Guard ring emissiveity	0.70, 0.70	
Lamp emissivities	0.30, 0.10	
Bottom wall emissivity	0.35, 0.35	
Top wall emissivity	0.05, 0.05	
Window emissivity	0.00, 0.95	*c*
Showerhead emissivity	0.00, 0.95	*c*
Temperature of surroundings	300	K

[a] `WAF`, `SHR`, and `WIN` refer to wafer, showerhead, and quartz window, respectively

[b] ($\lambda < 4\mu m$, $\lambda > 4\mu m$)

[c] Showerhead and window are invisible at $\lambda < 4\mu m$

II. Physical Model

A physical model for the temperature of the various components of the RTP system described in the previous section can be mathematically represented by the partial differential equation (PDE).

$$\frac{\partial}{\partial t}[\rho\, c_p(T)\, T(r,t)]\, dV = q_c(T) + q_h(T) + q_r(T) + f(T,u). \tag{1}$$

This model describes the rate of increase in thermal energy in a differential volume element dV with temperature $T(r,t)$ at position vector r and time t owing to heat transfer by conduction, $q_c(T)$, convection, $q_h(T)$, and radiation, $q_r(T)$, and to volumetric heat addition, $f(T,u)$. The variables ρ and $c_p(T)$ represent the density and temperature dependent specific heat capacity. The input vector, u, is the normalized electrical power into each of the five lamp filaments. The function $f(T,u)$ accounts for the distribution of this power in the filaments and is zero for elements with no external heating.

The net conductive flux, $q_c(T)$ is described by

$$q_c(T) = \nabla \cdot (\, k(T, r)\; \nabla T\,)\, dV \tag{2}$$

where $k(T,r)$ is the temperature dependent thermal conductivity of the medium at position r. For materials common in semiconductor equipment and for the associated temperature ranges, the thermal conductivities vary quite significantly over the range of operating temperatures (300K to 1400K) and this variation must be modeled.

A precise thermal model of the system would include complex computations of the convective heat transfer, $q_h(T)$, involving buoyancy and pressure driven flows with three-dimensional patterns owing to wafer rotation. However, at atmospheric or lower pressures and for the temperatures of concern in RTP systems, it is found that convection through the gases plays only a minor role in the heat transfer. This observation is not completely general and has to be considered for any specific system configuration. In this system, as in many RTP systems, the radiation heat transfer increases as T^{α}, where α is typically 3.5–4.5, depending on radiative properties, while convection tends to increase more-or-less linearly with temperature. Also, since the gases have very low thermal mass, the dynamics of gas temperature is very fast. Therefore, it is a good approximation to only model the temperature of the solid parts of the chamber and include conduction through the gases for components that are in close proximity. For surfaces that are exposed to incoming gas flows, we model the associated heat transfer from those surface using a spatial (and perhaps temperature) dependent convective heat transfer coefficient. For example, if a gas is injected at temperature T_∞, across a surface of temperature T, then the heat flux from that surface is

$$q_h(T) = h(T,r)\; (T - T_\infty) \tag{3}$$

Here $h(T,r)$ is the convective heat transfer coefficient which can be a function of both temperature and position.

For RTP systems and many other thermal processing systems in semiconductor manufacturing, radiative heat transfer is the dominant mode of heat transfer. It is important to model radiation as accurately as possible. We can define a radiative exchange factor for radiation at a specific wavelength λ as $R_\lambda(dV_I, dV_j)$ such that

$$q_{\lambda\, i j} = R_\lambda(dV_i, dV_j)\; [e_{b\lambda}(T_j) - e_{b\lambda}(T_i)] \tag{4}$$

is the net radiative power exchange from differential volume dV_j at temperature absorbed by differential volume dV_i at temperature T_j. The quantity $e_{b\lambda}(T)$ is the spectral emissive power of at wavelength λ and temperature T, and is given by Planck's law

$$e_{b\lambda} = \frac{C_1}{\lambda^5\left(e^{(C_2/\lambda T)} - 1\right)} \tag{5}$$

where C_1 and C_2 are standard radiation constants.

In practice, the best known approach to modeling the radiative transfer is to divide the geometry into a number of

finite volumes with specified surfaces and to model the radiation using a Monte-Carlo style ray-tracing method [4]. With this method a "photon" or ray is emitted from a surface or volume and tracked as it interacts with the other surfaces or volumes in the system until it is finally absorbed by one. The record of the sources and destinations of such rays comprises the radiative exchange matrix. Complicated directional radiative properties (e.g., Fresnel relations) that closely model real radiative properties can be easily incorporated this method. By emitting many rays from each element one can produce an exchange matrix with an accuracy that is dependent on the number of rays. The error for any given element of the exchange matrix decreases as the inverse square root of the number of rays, and one may trace several million rays from each element, to obtain a sufficiently accurate exchange matrix.

Typically the radiative properties of the materials will vary considerably over the wavelength range of interest for heat transfer (e.g., $0.5\mu m \le \lambda \le 1.0\mu m$). In principle one would compute the radiative exchange matrix at multiple wavelengths throughout this range and integrate the spectral fluxes (q_λ) over all wavelengths to obtain the radiative flux, q_r. However, this approach would require performing a large number of Monte-Carlo ray-trace calculations, which can be very time consuming (hours or days of computational time). Instead, it is more efficient to judiciously select a few wavelength bands at which to perform the ray tracing calculations, and assume the radiative properties are constant over each band. For the generic RTP system described above, a two-band model was used. One band was chosen for $\lambda < 4\mu m$, where the quartz windows are transparent, and the second band was for $\lambda > 4\mu m$, where the quartz is opaque. The radiative properties of the wafer, walls, and lamp filaments are specified as constant within each remaining band. Thus, for the generic RTP model the radiative flux is

$$q_r = R_1\, e_{b1}(T) + R_2\, e_{b2}(T) \tag{6}$$

Here R_1 and R_2 and are the radiative exchange matrices for band 1 and 2, respectively. The quantity $e_{b1}(T)$ is the integral of Eq. (5) over band 1 wavelength $\lambda < 4\mu m$) and $e_{b2}(T)$ is similarly the integral over band 2 ($\lambda > 4\mu m$).

Another complication associated with the radiative transfer calculation is that in many cases the radiative properties of the surfaces and volume elements can vary considerably with temperature. For example, at wavelengths longer than about $1.1\mu m$, the transmissivity of silicon is very temperature dependent. At low temperatures (e.g., $T < 500K$) pure silicon is quite transparent, while at high temperatures ($T > 700K$) it becomes nearly opaque. This effect is compounded by the fact that the transmissivity also depends strongly on the level of impurity in the silicon (doping level). Heavily doped silicon wafers are nearly opaque in this wavelength range. For typical process wafers the doping level is often in the intermediate range where the low temperature absorption is more than pure silicon but the wafer is not opaque. To accurately deal with this one would need to, at the very least, add another band in the range $1.1\mu m \le \lambda \le 4\mu m$ and perform a ray trace for a range of wafer absorptivities (α). This would result in a table of exchange matrices, $R_3(\alpha)$, that could be interpolated to give the exchange matrix for a prescribed α. In practice, one does not know *a priori* what α will be for a given wafer since it is not typically measured. Our solution to this problem is to account for this model uncertainty in the control design, and design a robust controller.

A. Discretized Model

A simulation model of the thermal system is developed by first dividing the components of the system into a number of finite control volumes or nodes, as illustrated in Figure 4 [3]. The state of the system can be described by the vector of node temperatures $\mathbf{T} = [T_1, T_2, \ldots, T_n]^T$, , where n is the number of nodes. An energy balance on each volume of the system results in the governing equation

$$\mathbf{M}\dot{\mathrm{T}} = \mathbf{Q}(\mathbf{T}, \mathbf{u}) \tag{7}$$

Here, $\mathbf{M} = \mathrm{diag}\{(m\, c_p)_1, \ldots, (m\, c_p)_n\}$ the diagonal matrix of node thermal masses. The right hand side of Eq. (7) is the net power into each node, $\mathbf{Q}(\mathbf{T},\mathbf{u})$, and is generally a non-linear function of the temperature of all nodes and all input powers $\mathbf{u}$.

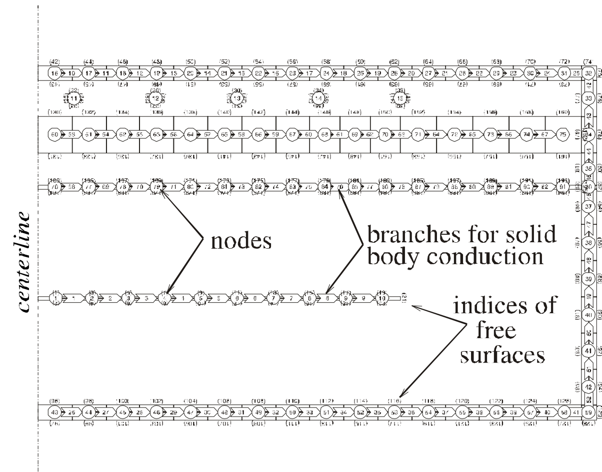

Figure 4: Discretized geometry for solution of the governing PDEs.

Eq. 7 establishes the basic structure of the physical model. As illustrated in Figure 5, the net power into each node is the sum of power contributions due to radiation, conduction, and convection heat transfer, as well as contributions by external inputs (e.g., lamp power) and possible contributions by other effects such as thermochemical reactions or mass transfer. Thus the net power into the cells can be divided into four main contributions,

$$\mathbf{Q}(\mathbf{T}, \mathbf{u}) = \mathbf{Q_r}(\mathbf{T}) + \mathbf{Q_c}(\mathbf{T}) + \mathbf{Q_{bc}}(\mathbf{T}) + \mathbf{Q_u}(\mathbf{T}, \mathbf{u}). \tag{8}$$

Here $\mathbf{Q_r(T)}$ is the radiative power into the nodes due to radiative exchange between all nodes of the system, $\mathbf{Q_c(T)}$ is the net power into each node due to conduction with neighboring nodes, and $\mathbf{Q_u(T,u)}$ is the net power into the nodes due to external inputs **u** such as lamp power. Finally, the net power into each node due to convective transfer due to gases flowing past solid boundaries is $\mathbf{Q_{bc}(T)}$. For completeness, other modes of energy transfer such as thermochemical reactions or mass transfer could also be included into this term. As the nomenclature suggests, $\mathbf{Q_{bc}(T)}$ is treated as a boundary condition (*bc*) for the solid elements of the system. In the following sections a brief description of each of the power terms in Eq. (8) is given.

1) Radiation Heat Transfer

As mentioned above, an accurate model for radiative transport is important because it dominates the thermal balance equation in high-temperature RTP systems operating in the 700–1400 K temperature range. For the generic RTP model we have used a number of simplifying assumptions that may not be applicable to real systems, but are adequate for the study of control design described here. For example, in the generic RTP system we assume that the surfaces are diffuse, i.e., the emission and reflection is independent of direction.

The tungsten halogen lamps provide the power for heating the silicon wafer. These lamps operate at temperatures in the range from 1500 to 3000 K, and radiate power that is absorbed by the wafer. At these high lamp temperatures, the spectral distribution of radiant power is significantly shifted to shorter wavelengths [4]. For example, more than 90% of blackbody radiation is at wavelengths shorter than 4μm for temperatures above approximately 2350 K. Additionally, the spectral emissivity, ε_λ, of tungsten decreases with increasing wavelength, falling from approximately 0.45 at visible wavelengths to less than 0.1 at wavelengths longer that 4μm [4]. The combined effect of wavelength on emissivity and on blackbody radiation results in a total emissivity:

$$\varepsilon(T) \equiv \frac{1}{\sigma T^4}\int_0^\infty \varepsilon_\lambda\, e_{b\lambda}(T)\, \mathrm{d}\lambda \qquad (9)$$

that increases from less than 0.1 at 500K to approximately 0.31 at 3000K.

At the lower temperatures of the silicon wafer ($T <$ 1400K) the blackbody radiation is shifted to longer wavelengths. For example, only about half of the radiative power of a blackbody at 1000K occurs at wavelengths shorter than 4μm. Aside from reflection (≈6%) and a few rather narrow absorption bands in the infrared due to impurities (particularly OH), quartz is very transparent at wavelengths shorter than approximately 4μm. At longer wavelengths, it is quite opaque due to the very strong Si–O–Si vibrational absorption band at 8–10μm. Therefore, while the quartz does not absorb significant amounts of the lamp emission, it does absorb a significant portion of the wafer emission. Further details of the physical model can be found elsewhere [3].

2) Model Parameters and Implementation

The model was implemented using a commercial graphical modeling and dynamic simulation software (Xmath/SystemBuild™) package using functional blocks with a structure similar to the schematic in Figure 5 [5]. The radiative exchange matrices were computed using a software package developed by the authors. Most of the simulation results presented here are for the baseline set of model parameters shown in Table 1 and Table 2. The thermal properties (e.g., c_p and k for the solid nodes are temperature dependent and are implemented using polynomial approximations.

3) Dynamic Results

The dynamic simulation results in Figure 6 show the centerline temperature histories for the wafer, lamp, showerhead, and quartz window for step changes in the power commands, u. The power command steps were selected from the optimal steady state results at T_{ref} = 500, 700, 900, and 1100°C. As shown, the lamp temperatures change very quickly with time constants on the order of 1 s. The wafer temperatures are the next fastest, with time constants on the order of 10 s. Finally, the showerhead and window time constants are of order 70 s and 500 s, respectively. Of course, the actual values for the time constants are dependent on the temperature, typically proportional to $\frac{1}{T^3}$.

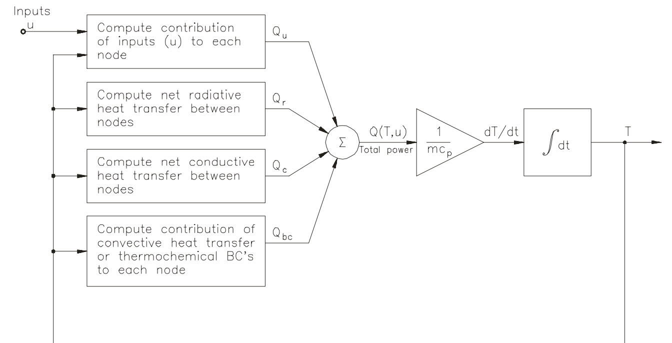

Figure 5: Block diagram of the physical model.

To quantify the response times and their temperature dependence consider a simplified scalar model for the centerline wafer temperature. Ignoring second order coupling between the wafer, lamp, and showerhead, suppose we fit a model of the form

$$\dot{T} = -A_c T - A_r T^4 + Bu + C \qquad (10)$$

Here A_c is proportional to the conductive or convective heat transfer coefficients, and A_r is a radiative loss coefficient. From static analyses, we know that the relative magnitudes of the five lamp commands are approximately constant. Hence, we can let $u = u_1$ without introducing significant

errors. The constant C is small, and is proportional to the loss coefficients multiplied by the gas temperature T_g.

Despite the simplicity of this approach, the least squares fit to he wafer temperature response obtained from the full model is quite good yielding $A_c \approx 0.0128\text{s}^{-1}$, $A_r \approx 1.81\times10^{-11}\text{K}^{-3}\text{s}^{-1}$, $B \approx 311\text{K/s}$, and $C \approx 5.26$ K/s. The error of approximately ±5°C results primarily from the omission of the coupling between the lamp and showerhead dynamics.

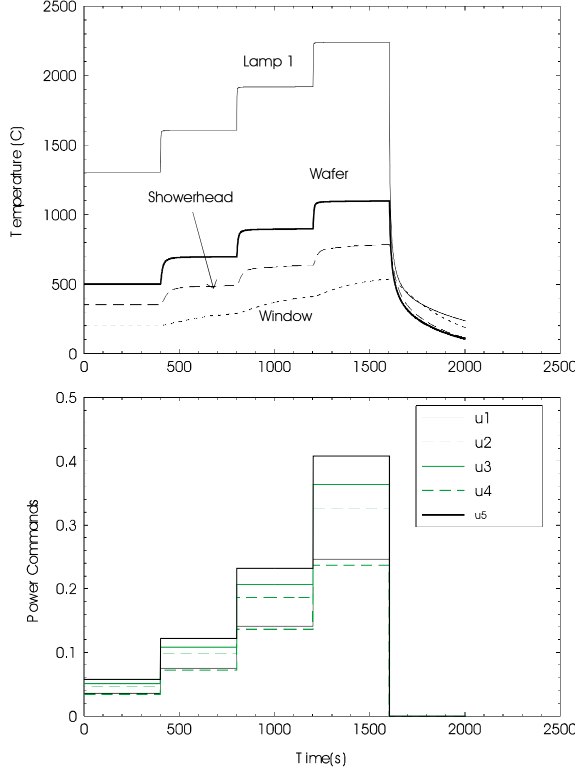

Figure 6: Centerline wafer, lamp, showerhead, and quartz window response to step changes in power commands.

Apart from estimating the lamp, showerhead, and the window temperatures, this simplified model can also be used to compute the approximate time constants of each element as a function of temperature from

$$\tau \approx \frac{1}{A_c + 4A_r T^3} \tag{11}$$

The temperature dependence of the gains may also be approximated from

$$G = \frac{B}{A_c + 4A_r T^3} = B\tau \tag{12}$$

For example, the peak wafer temperature gain due to Lamp 2 varies from approximately 1700°C/cmd (at 773 K) to approximately 350°C/cmd (at 1373 K), a ratio of ≈ 4.9. The simple model predicts a ratio of approximately 4.3. A more detailed dynamic analysis is given elsewhere [3].

III. Model Order Reduction

The need for efficient control design algorithms for real-time model-based control design calls for simple low-order system models that approximate the behavior of the full-order nonlinear models in sufficient detail. A variety of techniques are available for model order reduction including aggregation [8], Hankel singular value using Gramians [9], and principal orthogonal decomposition (POD) [10]. In this section we apply the POD method to models of RTP systems [11], [12].

The POD method is a nonlinear model-order reduction method where reduction of the size of the state space is achieved using a singular value decomposition of a matrix of snapshots of the state vector. There is an interesting physical interpretation of the POD method. The state trajectory is projected into a lower dimensional hyperspace. Also in the linear case, the POD is the same as the balanced model order reduction. The first step in the method is to perform model simulations resulting in an appropriate state matrix *X*. There are several possibilities; since we are interested in optimal control around a nominal open-loop optimal temperature trajectory, we used the variations around the nominal trajectory as the basis for constructing *X*. For this purpose, we first ran the nominal optimal model simulation by exciting the model with the optimal input sequence one hundred times in succession. Next, we added a sequence of independent Pseudo-Random Binary Sequences (PRBS) to the same optimal input sequence and used that as input for another simulation run. Each component of the input sequence consists of the sum of four independent PRBS sequences, chosen in such a manner that each of those sequences covers one of the four dominant time scales in the process (lamp, wafer, showerhead and window temperature). The fluctuation of the state around the nominal trajectory is then described by the difference of the perturbed sequence and the nominal one, which is our choice for the matrix *X.*

The basis for the POD method is formed by SVD of the snapshot matrix *X:*

$$X = U\Sigma V' = [U_1 \quad U_2]\begin{bmatrix}\Sigma_1 & 0 \\ 0 & \Sigma_2\end{bmatrix} = \begin{bmatrix}V_1' & V_2'\end{bmatrix}'. \tag{13}$$

Here, $X = (T_i, \ldots, T_i, \ldots, T_P)$ where the T_i is the state snapshots at time t_i and where U_i, V_i and Σ_i (i=1,2) are determined by a suitable truncation of Σ to the first n singular values. If the state dimension is N then X is a ($N \times P$) matrix. The full order thermal model is of the form,

$$\dot{T} = M^{-1}\left[A_c T + A_r T^4 + B u + C\right] \tag{14}$$

Assuming the snapshots are representative for the state vectors that occur during system operation, we can approximate T by $\hat{T} = U_1 z$ where z is obtained by substituting this term in Eq. (14),

$$\dot{z} = U_1' M^{-1}\left[A_c U_1 z + A_r (U_1 z)^4 + Bu + C\right]$$

$$= A^r{}_c \; z \; + \; A^r{}_r \; (U_1 \, z)^4 \; + \; B^r \, u \; + \; C^r \qquad (15)$$

when $\Sigma_2 = 0$, then $T = \hat{T}$. When Σ_2 is nonzero but small, $\hat{T}$ will be close to T.

We applied the method to the generic RTP system. The singular values of X were computed and the point where the singular values drop below 0.1% point was approximately for $n = 27$. We performed model validation on the nominal sequence first, for orders $n = 10$, 15, 20, 30, 35, and 40. Examining the RMS error between the full-order (116th order) and reduced order models, order 30 seemed to be a good choice. One of the most important measures is the wafer temperature which was found to be off by only one third of a degree. Other validation runs were carried out. An independent PRBS-perturbed sequence was used. The results are quite similar to that of validation on the nominal sequence. The computational details for model order reduction are presented elsewhere [12].

IV. Model-based Control Design

In this section we describe our strategy to design controllers for the generic RTP system. Precise temperature control is critical to obtaining required high performance as mentioned in Section 1. In an RTP chamber, many heaters affect the temperature at each location where it is measured. Multi-Input-Multi Output (MIMO) control that explicitly accounts for the influence of each heat source on each temperature sensor is needed for high performance. With such strong physical coupling, it is difficult to obtain high performance control of the temperature profile using single loop conventional controllers commonly used in industrial applications. Moreover, since previous approaches relied heavily on precise calibration, small changes in chamber design or wafer geometry can require substantial and time-consuming efforts in control re-design. The necessity for meeting extremely high performance specifications requires that the control system be optimal with respect to the specific process being controlled, and robust in order to cope with variations in the system components.

A. Control Problem Formulation

To be able to design temperature controllers that achieve the desired wafer quality, it is important to consider the performance specifications in terms of temperature control quality. The temperature control problem in an RTP system typically has the following demands to ensure uniform wafer properties:

1) Steady-state tracking, better than 1°C, preferably zero error;
2) Good temperature uniformity across the wafer during ramp, with little (only a few degrees Celsius) or no overshoot for temperature changes up to 600°C, varying ramp rates (50°C/sec to 300°C/sec), and set points up to 1100°C;
3) Insensitivity to sensor noise, process disturbance and variations, such as wafer-to-wafer variations (e.g., variation in wafer emissivity), changes in temperature setpoints, etc.

These demands pose a substantial challenge for controller design, since very high precision has to be obtained while retaining sufficient robustness in the design. It is noted that even if the controller has zero-tracking errors at points on the wafer where the temperatures sensed (usually five points or less), there could be large departures from the recipe temperature at several points where the temperatures are not measured. Our controller solves the problem using an estimate of the maximum error based on model prediction. This is a very important advantage in applying model-based control to RTP.

We approach the control problem by using linear design techniques [8], [14], [15]. Hence, we have to derive a linearized model of the system from the nonlinear discretized model (see Section II.A). Two alternatives are possible. The first option is to directly linearize the reduced nonlinear model of the system obtained as described in Section III. The second option is to linearize the full nonlinear model, and then use the POD reduction algorithm, as described in Section III. In either case, denoting the nonlinear model by

$$\begin{cases} \dot{T} = \underbrace{A_r T^4 + A_c T + C_1 + B_1 u}_{f(T)} \\ y = h(T) \\ z = H\,T \end{cases} \qquad (16)$$

where y denote the pyrometer measurements and z denote wafer temperatures. After selecting a suitable linearization (operating) point (x_o, u_o), the linear model is obtained by computing

$$A = \left.\frac{\partial f}{\partial x}\right|_{x0} \qquad (17)$$

$$B = B_1 \qquad (18)$$

$$C = \left.\frac{\partial h}{\partial x}\right|_{x0} \qquad (19)$$

which yields the usual linear state-space equations

$$\begin{cases} \dot{x} = A\,x + B\,u \\ y = C\,x \\ z = H\,x \end{cases} \qquad (20)$$

For ease of notation, we use the same symbols x, u and y to denote state, input and output of both the nonlinear and the linear models. The input vector $u \in \mathrm{IR}^m$ is composed by the power commands that are applied to the lamps. The output vector $y \in \mathrm{IR}^p$ is the sensed temperatures. The components of the input signal u are limited by physical constraints, i.e., the power applied to the lamps cannot be negative nor can they exceed a maximum value. We normalize the power with respect to this maximum value,

such that every component of u is bounded between zero (lamps off) and unity (lamps full on). The correct choice of the linearization point (x_o, u_o) is important. The system is linearized about a few points spanning the relevant (600°C–1100°C) temperature range.

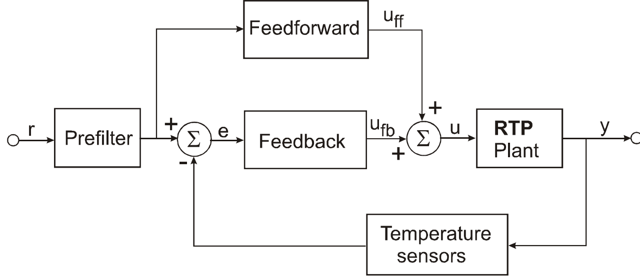

Figure 7: Controller structure.

B. Controller structure

The controller structure is shown in Figure 7. The *feedforward* controller takes advantage of the known reference temperatures to compute a suitable control signal that is injected in the closed-loop. Due to the relatively simple structure of a feedforward controller, it can be nonlinear and can be based directly on the nonlinear RTP model. An important practical consideration is whether the reference is known to the feedforward controller *a priori*, or if it is provided in real-time. The latter case is the most common in practice, but the first option allows a global optimization of the trajectory rather than point-to-point optimal commands. It is assumed here that the reference will be provided in real-time.

The *feedback* controller is based on a linear design, as dynamic output feedback is required. Its task is to address any mismatch that arises from the limited fidelity of the feedforward controller, and to deal with the process disturbances. The feedback controller includes logic to deal with integrator anti-windup due to lamp saturation nonlinearities [13].

The *prefilter* smoothes the temperature reference, the latter being piecewise linear and thereby having discontinuities in the rate of change. If the "raw" reference is tracked closely by the controller, it will inevitably result in overshoot, because finite lamp dynamics introduce delays between the feedback signal and the actuator (i.e., the system is at least of second order). In addition, the prefilter reduces excessive control action due to the sudden changes in rate. The following section discusses the feedback part of the RTP controller. Results on the feedforward design are given later in Section IV.D.

C. LQG Feedback

For feedback control, design we use Linear Quadratic Gaussian (LQG) control extended with frequency shaping. LQG is a standard controller design method that has been successfully applied to Multiple-Input Multiple-Output (MIMO) RTP control problems. Similar results may be obtained using H_∞ methods [18].

The basis for LQG control is the linear model of the RTP system as represented by Eq. (20). To be able to enforce zero steady-state tracking error, this model is augmented with integrators on the plant output. The resulting model is

$$\begin{cases} \dot{x}_{aug} = \begin{bmatrix} 0 & C \\ 0 & A \end{bmatrix} x_{aug} + \begin{bmatrix} 0 \\ B \end{bmatrix} u \\ y = \begin{bmatrix} I & 0 \end{bmatrix} x_{aug} \end{cases} \tag{21}$$

with $z = [\xi'\ x']'$. The design of an LQG controller is separated into the design of the state feedback gain K, and the design of the estimator gain L. The state feedback gain is found by minimizing the quadratic cost function:

$$J_k = \frac{1}{2}\int_0^\infty \left(x'_{\text{aug}}\, Q\, x_{\text{aug}} + u'\, R\, u \right) \mathrm{d}t \tag{22}$$

Here the symmetric positive-(semi-) definite matrices Q and R are the key design parameters. R is used to penalize the control effort, whereas Q is used to penalize tracking error. Typically we choose $R = \rho I$ where ρ is a scalar, and Q is selected in such a way to penalize temperature differences. Hence its structure is given by (shown for n = 4, see [13])

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix} \tag{23}$$

The resulting gain K can be partitioned according to:

$$K = [K_I \quad K_P] \tag{24}$$

where K_I is the integral gain associated with the integral state variables ξ, and K_P with the plant state x. The design of the estimator gain is similar, which is found by minimizing:

$$J_L = \frac{1}{2}\int_0^\infty \left(z' R_{\text{w}} z + y' R_{\text{v}} y \right) \mathrm{d}t$$

Here the symmetric positive-definite matrices R_w and R_v are the design parameters. Typically, R_w and R_v are used to characterize the statistical properties of Gaussian noise at state x_{aug} and output y. However, they can be chosen as diagonal matrices to provide design knobs for the estimator design.

By including the integral state variables in the controller, the control law now has the following structure

$$\dot{\gamma} = e \tag{25}$$

$$\dot{\hat{x}} = A\hat{x} + Bu + L\left(e - \hat{e}\right) \tag{26}$$

$$\hat{e} = C\hat{x} \tag{27}$$

$$u = -K_I\,\gamma - K_P\,\hat{x} \tag{28}$$

where $e = r - y$, with r the reference, such that the resulting controller state-space realization becomes:

$$\begin{cases} \dot{q} = \begin{bmatrix} 0 & 0 \\ BK_I & A - LC - BK_P \end{bmatrix} q + \begin{bmatrix} I \\ L \end{bmatrix} e \\ u = -\begin{bmatrix} K_I & K_P \end{bmatrix} q \end{cases} \quad (29)$$

with $q = \begin{bmatrix} \gamma' & \hat{x}' \end{bmatrix}'$. Note that the resulting controller is solely based on the error *e,* rather than using *y* for the estimator. The controller is augmented with appropriate integrator anti-windup logic to deal with lamp saturations [13].

The resulting performance of the designed controller as tested on the full nonlinear generic RTP model is presented in Figure 8. This figure shows the tracking response for a 50°C/ramp rate without the use prefilter or feedforward. The tracking performance is good since the tracking error is small and the overshoot is limited to 1°C. Hence, settling is achieved as soon as the response enters the band of ±1°C around the final temperature. Furthermore, wafer temperature non-uniformity is limited to approximately 3°C during ramp-up. The peak at 10 sec and the drop at 20 sec are both due to sudden changes in the reference ramp.

While the controller performance is good from a tracking point of view, it is likely to be sensitive to sensor noise, disturbances, and/or model uncertainty at high frequencies. To investigate this shortcoming, we considered a representative 2 Hz periodic measurement disturbance induced by a 120 rpm wafer rotation. We modeled this disturbance as a sinusoidal signal of frequency 2 Hz with random phase, and amplitude linearly increasing from 1°C at the center temperature measurement to 5°C at the edge measurement. Figure 9 shows the tracking response using the same controller used in Figure 8. Its performance is now unacceptable. The excellent tracking performance is overshadowed by the effect of the periodic disturbance: larger overshoot, no settling, and increased temperature non-uniformity. The reason for this performance degradation is the high sensitivity of the power input to the disturbance at frequency 2 Hz, which is displayed in Figure 9d. The high controller gain at high frequencies, which provided good tracking, also amplified the measurement disturbance.

To decrease the controller sensitivity to measurement disturbances, a frequency shaping filter was added to the LQG control design to improve high frequency roll-off, especially at 2 Hz. Figure 10 shows the same simulation as in Figure 9, but now with the improved controller. Clearly, this controller is much more insensitive to the 2 Hz measurement disturbance. The good tracking properties shown in Figure 8 are partly recovered. However, the overshoot has increased to approximately 2°C, and consequently the settling has increased. Also, wafer temperature non-uniformity has slightly degraded. However, for a 50°C/sec ramp rate, this performance is acceptable according to the requirements in Section IV.A.

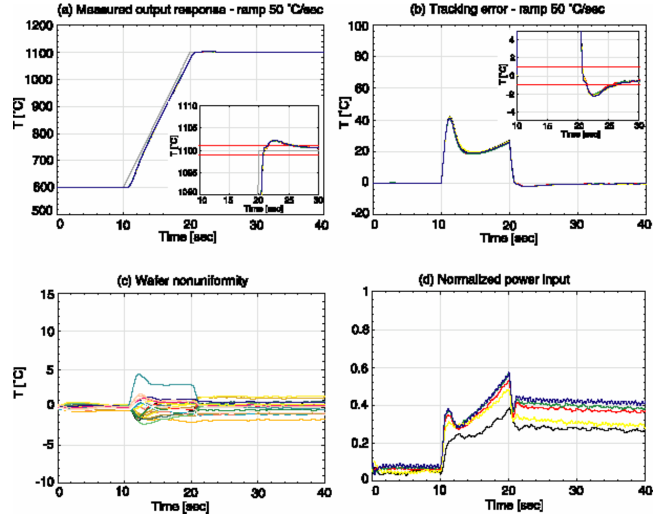

Figure 8: Simulated tracking response LQG controller for ramp rate 50°C/sec; (a) reference r and measured wafer temperature y; (b) tracking error $e = r - y$; (c) wafer temperature non-uniformity for 21 nodes on wafer. Each line represents the distance from the average wafer temperature; (d) power input u to RTO systems, normalized to maximum power of 65 W.

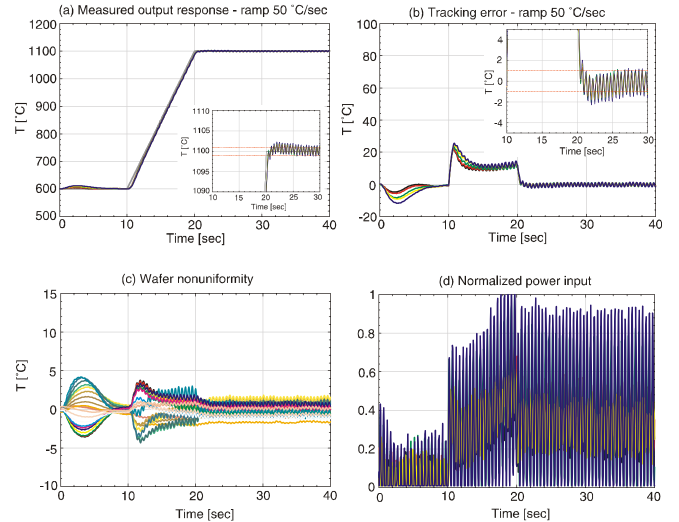

Figure 9: Simulated tracking response LQG controller for ramp rate 50°C/sec with 2 Hz periodic disturbance added to the measurements; see Figure 8 for an explanation of individual graphs.

D. Feedfoward Control

It is very difficult to independently achieve both good tracking, disturbance suppression, robustness to unmodeled dynamics, and stabilization with a single-degree-of-freedom (feedback) controller [15], [16]. By adding a *feedforward,* controller, as shown in Figure 7, one uses the reference temperatures to compute a suitable control signal that is injected into the closed-loop. Since we wish to

move the system from one operating point to another along a specified trajectory, we can approximately determine the input that is required for this. Consequently, we can apply this input directly to the system instead of the feedback controller computing the input based on the tracking error *e*.

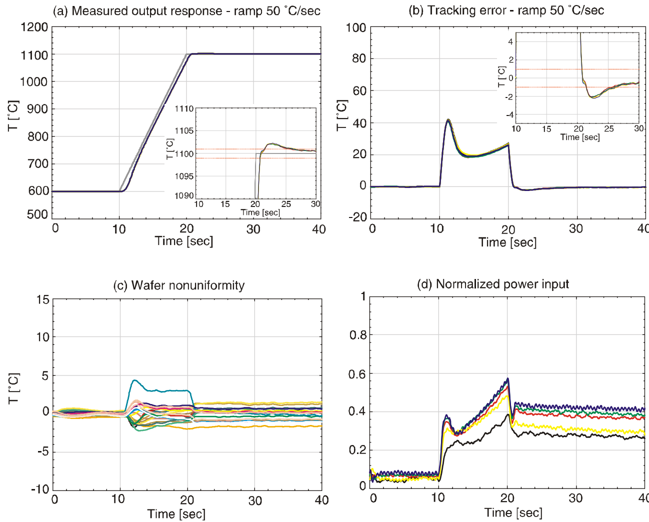

Figure 10: Simulated tracking response for controller with improved high frequency rolloff for ramp rate 50°C/sec with 2 Hz periodic disturbance added to the measurements. See Figure 8 for an explanation of individual graphs.

In the control structure of Figure 7, the feedforward filter should approximate the *inverse dynamics* of the RTP plant. We computed this inverse using the high-order linear plant model because inverting a low-order (approximate) plant model resulted in unstable feedforward filters due to the presence of non-minimum phase transmission zeros in the low-order approximation. The high-order inverse model may be reduced, if desired, although it was not done for these simulations. Figure 11 shows the simulation results for tracking a ramp with 50°C/sec with a controller scheme that includes feedback, feedforward and prefilter. The prefilter consisted of a second-order lowpass filter for each measured output channel. These figures show the merits of using prefilter and feedforward. The prefilter suppresses the overshoot shown in Figure 10, but also delays the response, whereas the feedforward speeds up the response. By exploiting the full freedom in the controller design, we are now able to achieve tight tracking, fast settling, very little overshoot, and robustness against high frequency model errors and measurement disturbances.

In addition to *in-situ* feedback/feedforward control, an off-line run-to-run control scheme may be added to the existing feedback control scheme in order to improve the product quality [17].

E. Controller Implementation

To implement this high-performance temperature controller on an actual RTP system, a computer with significant computational power is used. The controller software is executed on a real-time operating system (RTOS) because the controller may be run (i.e., the inputs sampled, and the outputs computed) at a fairly high rate (10 Hz or more). Depending on the actual RTOS and the hardware platform used, tradeoffs between controller performance and achievable sampling rate may have to be made. Once the controller is implemented, the actual system performance is compared with the simulated system performance. The differences are due to several factors including unmodeled plant dynamics, noise in the system, and un-modeled actuator and sensor dynamics. The process of controller design generally necessitates multiple iterations involving re-design and re-testing. Once a design is deemed satisfactory (i.e., meets specifications), there is a subset of system perturbations which the controller can accommodate by making minor "tuning" changes to the controller parameters. Figure 12 [14] shows the aomparison between closed-loop simulation results and actual sensor measurements for peak wafer temperatures and actuator commands during spike anneal. The agreement is excellent.

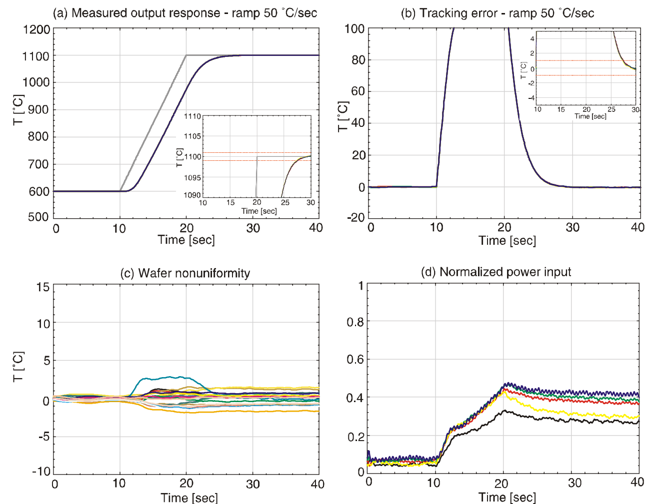

Figure 11: Simulated tracking response for controller included feedback, feedforward and prefilter for ramp rate 50°C/sec. See Figure 8 for an explanation of individual graphs.

V. Conclusions

In this paper we have described our model-based control system design methods for distributed temperature control for RTP systems. However, the modeling and control approaches are generic, and are applicable to a variety of thermal systems including furnaces. A physics-based, high-order, nonlinear model was developed for an axisymmetric RTP chamber with generic attributes. The governing equations of heat transfer for the components of the chamber are nonlinear, coupled PDE's. The control-volume discretization used to create the high-order model results in a set of coupled, non-linear ODE's. For model reduction, the POD approach, which is based on principal component analysis, was used to develop a low-order

nonlinear model. A model-based LQG controller was designed for the linearized low-order model and was shown to provide excellent temperature control on both the low-order and the full nonlinear simulations. While the details of the generic chamber were discussed here in illustrating the methodology, the same approach was used for real-time control of an actual commercial RTP system, and yielded excellent results. Controllers developed using this approach have been installed in commercial systems, and currently operating in the field.

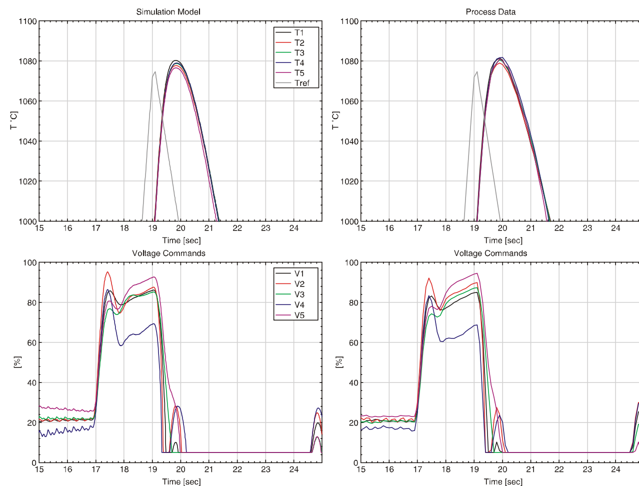

Figure 12: Comparison of model simulation (left column) with actual measurements (right column) on a 200 mm single-wafer RTP chamber during a fast-ramp process.

ACKNOWLEDGEMENTS

The authors also would like to acknowledge the contributions of G. van der Linden, H. Aling, Prof. K. F. Jensen, Dr. S. Banerjee, and Prof. I. G. Kevrekidis.

REFERENCES

[1] T. F. Edgar, S. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. S. Balakrishnan and J. Hahn, "Automatic Control in Microelectronics Manufacturing: Practices, Challenges, and Possibilities," *Automatica*, Vol. 36, No. 11, pp. 1567-1603, 2000.

[2] J. Nulman, "Rapid Thermal Processing with Reactive Gases," in Reduced *Thermal Processing for ULSI*, Ed. R. A. Levy, NATO ASI Series, Plenum, 1989.

[3] J. L. Ebert, A. Emami-Naeini, R. L. Kosut, "Thermal modeling of rapid thermal processing systems," *3rd International Rapid Thermal Processing Conference*, R. B. Fair, B. Lojek (eds), pp. 343-355, Amsterdam, The Netherlands, 1995.

[4] M. F. Modest, *Radiative Heat Transfer*, McGraw-Hill, 1993.

[5] S. Ghosal, J. L. Ebert, K. Chung, G. Aral, and A. Emami-Naeini, "A GUI-based Concurrent Software Tool for Thermal Modeling and Control System Design for RTP Chambers," *2nd Symp. Process Ctrl, Diag., Model. in Semicond. Manuf.*, 191st Meeting of the Electrochemical Soc., Montreal, May 1997.

[6] M. G. Kabuli, R. L. Kosut, and S. Boyd, "Improving static performance robustness of thermal systems," *Proc. 1994 IEEE CDC*, pp. 62-66, December 1994.

[7] R. L. Kosut and M. G. Kabuli, "Robust Control of Thermal Processes"" in *Proc. 2nd International Rapid Thermal Processing Conference*, *RTP'94*, pp. 296-297, August 1994.

[8] J. L. Ebert, A. Emami-Naeini, H. Aling, R. L. Kosut, "Thermal Modeling and Control of Rapid Thermal Processing Systems," *Proc. 34th IEEE Conf. Dec. Control*, pp. 1304-1309, December 1995.

[9] K. Zhou, J. C. Doyle and K. *Glover, Robust and Optimal Control*, Prentice-Hall, 1996.

[10] L. Sirovich, C. H. Sirovich, "Low dimensional Description of Complicated Phenomena," *Contemporary Mathematics*, Vol. 99, pp. 277-305, 1989.

[11] H. Aling, S. Banerjee, A. K. Bangia, V. Cole, J. Ebert, A. Emami-Naeini, K. F. Jensen, I. G. Kevrekidis, S. Shvartsman, "Nonlinear Model Reduction for Simulation and Control of Rapid Thermal Processing," in *Proc. Automatic Control Conference*, June 1997.

[12] H. Aling, A. Emami-Naeini, R. L. Kosut, J. L. Ebert, "Nonlinear Model Reduction with Application to Rapid Thermal Processing," in *Proc. 35th IEEE Conf. Decision Contr.*, December 1996.

[13] G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems,* 4th Ed.Prentice-Hall, 2002.

[14] D. de Roover, S. Ramamurthy, A. Mayur and J. L. Ebert, "Improved Performance of a Fast-Ramp RTA System through Recipe and Controller Optimization," *In Symp. Rapid Thermal And Other Short-Time Processing Technologies,* 197th ECS Meeting, Toronto, Ontario, Canada, Vol. 2000-9, pp. 445-452, 2000.

[15] D. de Roover, A. Emami-Naeini, J. L. Ebert, and R. L. Kosut, "Tradeoffs in Temperature Control of Fast-Ramp RTO and RTA Systems," *7th Intl Conf. on Advanced Thermal Processing of Semiconds*, *RTP 99*, Colorado Springs, Sept. 1999.

[16] D. de Roover, A. Emami-Naeini, J. L. Ebert, S. Ghosal, G. W. van der Linden, "Model-Based Control of Fast-Ramp RTP Systems,"*6th Int Conf. on Advanced Thermal Processing of Semiconductors, RTP 98,* pp. 177-186, Sept. 1998.

[17] R. L. Kosut, D. de Roover, A. Emami-Naeini, J. L. Ebert, "Run-to-Run Control of Static Systems," in *Proc. 37th IEEE Conf. Decision Contr.*, pp. 695-700, December 1998.

[18] G. W. van der Linden, A. Emami-Naeini, R. L. Kosut, "RTP Robust Control Design: Part II: Controller Synthesis,"in *Proc. RTP'96*, September 1996.