Nonlinear Model Reduction Strategies for Rapid Thermal Processing Systems

Suman Banerjee, J. Vernon Cole, and Klavs F. Jensen

***Abstract*—We present a systematic method for developing low order nonlinear models from physically based, large scale finite element models of rapid thermal processing (RTP) systems. These low order models are extracted from transient results of a detailed finite element model using the proper orthogonal decomposition (POD) method. Eigenfunctions obtained from the POD method are then used as basis functions in spectral Galerkin expansions of the governing partial differential equations solved by the finite element model to generate the reduced models. Simulation results with the reduced order models demonstrate good agreement with steady state and transient data generated from the finite element model, with an order of magnitude reduction in execution time.**

***Index Terms*— Eigenfunctions, eigenvalues, Galerkin method, rapied thermal processing, reduced order systems.**

I. Introduction

PROCESSES used to manufacture semiconductor devices are becoming increasingly complex, while competition demands that these devices be brought to market more quickly and manufactured more reliably. This calls for reduction in the large number of cut-and-try iterations in developing processes, process equipment, or process control software. In order to speed up this development process, one needs to understand the complicated physical rate processes governing each fabrication step. Such an understanding is best expressed in terms of a detailed, physically based, mathematical model. However, the solution of such a model is often time consuming and requires the use of hardware and software resources beyond those available to typical manufacturing organizations because of the complex time dependent and three-dimensional nature of the production equipment. Simulations of these processes using the existing computational models can take hours to days to yield results. Therefore, techniques are required for deriving low-order, physically based models for semiconductor manufacturing processes. These models could be used to study on-line process variations or to answer "what-if" type of questions under a limited range of conditions. The reduced complexity and smaller computational storage requirements imply that the reduced models can be simulated on desktop computers (such as PC's), besides workstations. Hence, process engineers and operators could use these models for a better understanding of semiconductor manufacturing processes. A well-designed reduced model could help in cutting down the number of experiments required in designing a process recipe and thus reduce the transition time in bringing a process from the research to the manufacturing stage in a fabrication line. Another use for such a model would be in advanced model based control strategies.

In this paper, we have studied a model reduction technique using a rapid thermal processing (RTP) system as a test vehicle. RTP is an emerging technology in chip manufacturing processes and has shown promise in a wide variety of applications. A typical fabrication process may consist of as many as 26 different RTP steps of oxidation, annealing, nitridation and chemical vapor deposition [1]. The demand for submicron device sizes have placed severe constraints on the thermal processing of silicon wafers. To minimize solid state diffusion of dopants, the amount of time spent by the wafer close to processing temperature needs to be considerably reduced. RTP provides a viable alternative to existing thermal processing techniques. RTP systems are, in general, single wafer reactors [2], [3]. The wafer is heated by tungsten halogen filament lamps or by water-cooled arc lamps. The primary mode of heat transfer to the wafer is by radiation from the lamps. The wafer is typically supported by quartz pins, so that the wafer temperature may be ramped at very high rates ($\sim$100 K/s). After processing, the wafer is ramped down quickly and the process gases are purged from the reactor using inert gases. The wafer processing time in a RTP reactor is very short, which minimizes diffusion lengths and preserves already formed dopant profiles from previous steps. The fast dynamics and transient nature of a RTP system make it a good choice for exploring the capabilities of the model reduction procedure.

The emerging nature of RTP technology, drives the need for models, both reduced and complex, which would lead to a better understanding of the process. A number of model based control studies of RTP systems have been developed [4]–[8], but, further advances in control design would need accurate models capable of simulating the process in real time (or faster). In this work we have developed nonlinear low order models without approximating the physical conservation equations describing the process, thereby making them more accurate compared to conventional linear models over a wider range of conditions. Such models show promise for application in the development of model based control schemes [9].

Manuscript received October 9, 1997; revised December 10, 1997. This work was supported by the Office of Naval Research under Contract N00014-94-C-0187 and the Semiconductor Research Corporation.

S. Banerjee and K. F. Jensen are with the Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: suman@mit.edu; kfjensen@mit.edu).

J. V. Cole was with the Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with Motorola, PEL, Austin, TX 78721 USA (e-mail: vernon_cole-RA5179@email.sps.mot.com).

Publisher Item Identifier S 0894-6507(98)02939-X.

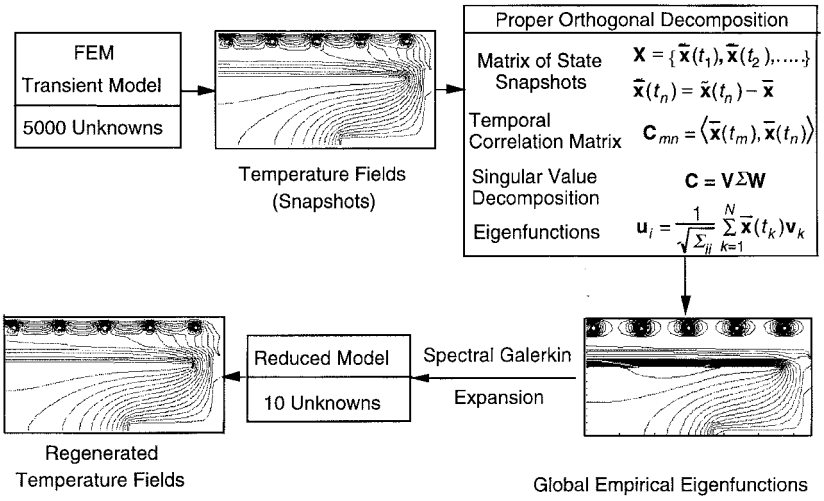

Fig. 1. Schematic representation of the model reduction method.

II. Model Reduction Approach

The model reduction strategy is shown schematically in Fig. 1. A detailed, physical model of a generic two-dimensional (2D) RTP system [10], [11] with features relevant to the next generation of RTP systems serves as the base case for the model reduction study. The modeling strategy used to generate this detailed model is similar to that used in previous simulations of RTP systems [10], [11]. It is based on a finite element (FEM) solution of the general equations representing conservation of mass, momentum and energy. The boundary condition of the energy equation describes the radiation heat transfer, which separates the thermal radiation into multiple wavelength bands and includes the effect of multiple reflections. In the present case studies the velocity field is assumed to be constant through the RTP cycle, fixed at a steady state solution at the process hold conditions, a reasonable approximation at low pressures [10], [11]. At higher pressures transient flow effects must be included, but the general model reduction strategy will remain the same.

The modeling equations are solved by the Galerkin finite element method [10], [11]. In this method the unknown flow and temperature fields are approximated by expansions in piecewise, low order polynomials. This approach has the advantage of being general and flexible, but the large number of coefficients required leads to large nonlinear matrix problems. The numerical solution of this problem therefore requires workstations and special computational algorithms. The number of coefficients involved in representing the temperature fields could, in principle, be reduced if the approximating functions were similar in form to the actual solution. One approach for obtaining better approximating functions is the proper orthogonal decomposition (POD) method [12], [13] (also known as the Karhunen–Loève procedure). This method was first suggested by Lumley [14] as a rational procedure for the extraction of *coherent structures* [12]. In this method, empirical eigenfunctions can be extracted from either experimental observation or detailed model predictions of temperature fields ("snapshots") for the entire reactor at discrete time intervals. The method of eigenfunction extraction starts with a matrix of transient temperature fields generated by the finite element model at discrete time intervals:

$$\boldsymbol{X} = \{\tilde{\boldsymbol{x}}(t_1), \tilde{\boldsymbol{x}}(t_2), \cdots\} \tag{1}$$

$$\tilde{\boldsymbol{x}}(t_n) = \boldsymbol{x}(t_n) - \overline{\boldsymbol{x}} \tag{2}$$

where $\tilde{\boldsymbol{x}}(t_n)$ is the transient temperature field extracted at time t_n and $\overline{\boldsymbol{x}}$ is the steady state temperature field. For generating these temperature fields, the transient FEM model of the RTP reactor is run with a set of lamp powers till the wafer attains a steady temperature, $\overline{\boldsymbol{x}}$. After the wafer has attained a steady state, the lamp powers are individually perturbed to generate variations in the wafer temperature. The temperature fields obtained from these lamp power perturbations are then stored in the matrix $\boldsymbol{X}$. A temporal correlation matrix is subsequently constructed from the snapshots as follows:

$$\boldsymbol{C}_{mn} = \langle \tilde{\boldsymbol{x}}(t_m), \tilde{\boldsymbol{x}}(t_n) \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the inner product in the ℓ_2 norm. The eigenfunctions $\boldsymbol{u_i}$ are obtained from a singular value decomposition of the temporal correlation matrix,

$$\boldsymbol{C} = \boldsymbol{V}\Sigma\boldsymbol{W} \tag{4}$$

$$\boldsymbol{u_i} = \left(\frac{1}{\sqrt{\Sigma_{ii}}}\right) \sum_{k=1}^{N} \tilde{\boldsymbol{x}}(t_k)\boldsymbol{\nu}_k \tag{5}$$

where $\boldsymbol{V}$ is a matrix whose columns are the left singular vectors of $\boldsymbol{C}$ and Σ is a diagonal matrix with the singular values of $\boldsymbol{C}$ on the diagonal. Therefore the eigenfunctions are admixtures of the snapshots [15], [16]. The number of eigenfunctions determined from this technique is equal to the dimension of the square temporal matrix, $\boldsymbol{C}$. These eigenfunctions form an optimal basis set for the given series of snapshots [18]. The remaining eigenfunctions, for the series of snapshots, are not uniquely determined. The only requirement on them is that they be orthogonal to the already determined set, and hence orthogonal to the snapshots $\boldsymbol{\nu}_k$. The empirical eigenfunction set generated by this technique can be used to regenerate a series of temperature fields by

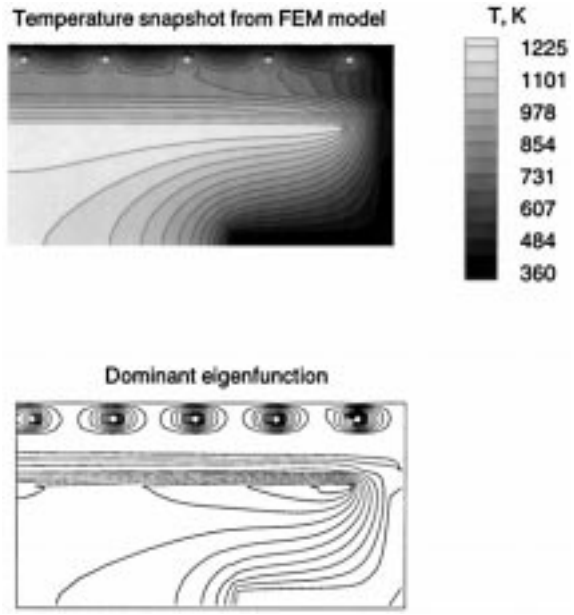

Fig. 2. Comparison of a typical temperature snapshot obtained from the FEM model with the dominant eigenfunction extracted by the POD procedure.

projecting a suitable set of temporal coefficients, $a_i(t)$, on the eigenfunction basis set as

$$\hat{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}} + \sum_{i=1}^{N} a_i(t)\boldsymbol{u_i}. \tag{6}$$

In order to solve for this set of temporal coefficients, $a_i(t)$, we have to integrate an initial value problem for a group of ordinary differential equations (ODE's). This set of ODE's, which is the low order reduced model, is obtained by the procedure discussed in the following section.

III. Method of Generating Nonlinear Reduced-Order Models

Fig. 2 compares a typical temperature snapshot obtained from the transient FEM model to the dominant eigenfunction extracted from the snapshots by the POD method discussed above. The dominant eigenfunction, i.e., the eigenfunction corresponding to the largest singular value, has most of the qualitative information about the temperature field. This can be seen from the figure where the contours of the temperature field closely match those of the dominant eigenfunction. Therefore, the empirical eigenfunctions, whether determined from computation or experiments, may be viewed as ideal fitting functions to be used in a pseudospectral [17] Galerkin procedure [18].

In general, for a set of differential equations on one variable, $\boldsymbol{y}$, expressed as

$$\frac{d\boldsymbol{y}}{dt} = F(\boldsymbol{y}) \tag{7}$$

the pseudospectral Galerkin procedure is given by

$$\left\langle \boldsymbol{u_i}, \frac{d\boldsymbol{y}}{dt} \right\rangle = \langle \boldsymbol{u_i}, F(\boldsymbol{y}) \rangle, \qquad i = 1, \cdots N \tag{8}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, $\boldsymbol{u_i}$ represents the eigenfunctions, and N is the number of eigenfunctions used in the pseudospectral Galerkin procedure. This procedure, using empirical eigenfunctions, has been applied to modeling turbulence and large-scale problems in fluid mechanics [16], [19]–[22].

The general method of expressing the FEM model in a form amenable to model reduction is given below. The important idea is to separate the terms linear and nonlinear in temperature so that they can be handled separately. The conservation of energy equation in the FEM model takes the following form:

$$\rho C_p \left[\frac{\partial T}{\partial t} + \boldsymbol{v} \cdot \nabla T \right] = \nabla \cdot [k^f \nabla T] \tag{9}$$

where

- ρ density;
- C_p specific heat;
- $\boldsymbol{v}$ velocity vector;
- T temperature;
- k^f fluid thermal conductivity.

The density and specific heat of the gas phase are modeled as temperature dependent properties in the FEM model. The solid thermal properties, except for the thermal conductivity of the silicon wafer, are constant in the model. The boundary condition for (9) takes the following form [10], [11]

$$k^s \nabla T_i \cdot \boldsymbol{n} = k^f \nabla T_i \cdot \boldsymbol{n} + \alpha_i \sum_{l=1}^{N_{\text{lamp}}} R_{il} P_l + \sigma \sum_{k=1}^{N_{\text{bands}}} \cdot \left[\alpha_i^k \sum_{j=1}^{N_{\text{SW}}} \phi_{\lambda^k - T_j} \varepsilon_j^k R_{ij}^k T_j^4 - \phi_{\lambda^k - T_i} \varepsilon_i^k T_i^4 \right] \tag{10}$$

The left-hand side represents the conduction into the solid. This is balanced on the right hand side by conduction in the gas, energy input from the lamps, and energy transfer with other surfaces in the system. In (10), k^s is the solid thermal conductivity, α is the absorptance of the solid surface, P_l is the radiation intensity of lamp l, σ is the Stefan–Boltzmann constant, $\varepsilon_j^k R_{ij}^k$ is the percentage of radiation, in band k, leaving surface i which is absorbed by surface j (by direct viewing and all intervening reflections). The exchange factors, R_{ij}^k, are assumed to be temperature independent based on the high temperature opaque silicon properties [10], [11]. This has been shown to be a reasonable approximation for RTP processes [10], [11].

The gas in the lamphouse and in the region between the showerhead and the quartz window is treated as stagnant. Therefore the gradients in temperature in these regions are determined by gas phase conduction. There are additional boundary conditions at the fluid–solid interfaces on the exterior walls of the reactor which represent heat transfer to the surrounding ambient. Using the Galerkin finite element method, (9) is transformed to a set of algebraic equations as follows:

$$\int_D \rho C_p \left[\frac{\partial T}{\partial t} + \boldsymbol{v} \cdot \nabla T \right] \Phi^i \, dV = -\int_D k^f \nabla T \cdot \nabla \Phi^i \, dV + \int_{\partial D} k^f \nabla T \cdot \boldsymbol{n} \Phi^i \, dS \tag{11}$$

where Φ^i are the piecewise continuous basis functions used in the finite element method, D represents the volume of the domain and ∂D is the boundary of the domain [10], [11]. The boundary condition shown in (10) is evaluated as part of the boundary integral in (11).

In order to make the conservation of energy equation (9) amenable to the model reduction technique, the terms in the finite element expansion (11) are lumped together and expressed in the following matrix form

$$\boldsymbol{M}(\tilde{\boldsymbol{x}})\frac{d\boldsymbol{x}}{dt} = \boldsymbol{C}(\tilde{\boldsymbol{x}})\tilde{\boldsymbol{x}} + RD(\tilde{\boldsymbol{x}}) + \alpha_i \sum_{l=1}^{N_{\text{lamp}}} R_{il}P_l \tag{12}$$

where $RD(\tilde{\boldsymbol{x}})$ is the nonlinear radiation heat transfer contribution to the reduced model and can be written as

$$RD(\tilde{\boldsymbol{x}}) = \sigma \sum_{k=1}^{N_{\text{bands}}} \cdot \left[\alpha_i^k \sum_{j=1}^{N_{\text{SW}}} \phi_{\lambda^k - T_j}\varepsilon_j^k R_{ij}^k T_j^4 - \phi_{\lambda^k - T_i}\varepsilon_i^k T_i^4\right]. \tag{13}$$

$\boldsymbol{M}(\tilde{\boldsymbol{x}})$ is obtained by lumping all the dynamic contribution from the energy conservation equation and $\boldsymbol{C}(\tilde{\boldsymbol{x}})$ is obtained by lumping all the convection and conduction terms from the energy conservation equation.

This separates the nearly linear conduction and convection terms in the matrix $\boldsymbol{C}(\tilde{\boldsymbol{x}})$ from the highly nonlinear radiation terms in the matrix $RD(\tilde{\boldsymbol{x}})$. Thus, the temperature dependence of material properties, such as the gas phase thermal conductivity, k^f, gas phase specific heat, C_p, etc., can be linearized and included in the matrices $\boldsymbol{M}(\tilde{\boldsymbol{x}})$ and $\boldsymbol{C}(\tilde{\boldsymbol{x}})$. In the actual FEM model, these material properties are expressed as power law fits which are weakly nonlinear compared to the terms in the radiation heat exchange. Since the reduced model is extracted using deviation eigenfunctions, the models extracted would be exact around the given steady state and would differ from the FEM model around other operating conditions depending on the nonlinear effects of the material properties.

The method of extracting nonlinear reduced order models is implemented in deviation variables, i.e., the steady state temperature field is subtracted from the transient temperature fields and the eigenfunctions are extracted from the deviation fields. This eliminates any steady-state offset completely, if one generates the reduced model from small perturbations about a given steady state. The T^4 nonlinearity in the radiation heat exchange term prevents a linearization of the model equations from being valid over a broad range of conditions. Therefore, this contribution to the reduced model has to be evaluated by reconstructing the temperature fields, generating the T^4 term explicitly in absolute temperatures, and then evaluating the radiation contribution to the reduced model at every time step.

The empirical eigenfunctions obtained from the POD method are used in a pseudospectral Galerkin expansion of (12). The resulting low order $(N < 10)$ system of ordinary differential equations takes the form

$$\left(\boldsymbol{u}_m^T\boldsymbol{M}(\overline{\boldsymbol{x}})\sum_{i=1}^{N}\boldsymbol{u}_i\frac{da_i}{dt}\right) = \left(\boldsymbol{u}_m^T\boldsymbol{C}(\overline{\boldsymbol{x}})\sum_{n=1}^{N}a_n(t)\boldsymbol{u}_n\right) + (\boldsymbol{u}_m^T\boldsymbol{R})[\hat{\boldsymbol{x}}(t)]^4 + \left(\boldsymbol{u}_m^T\left[\alpha_i\sum_{l=1}^{N_{\text{lamp}}}R_{il}\tilde{P}_l + \boldsymbol{K}\right]\right) \tag{16}$$

where $\boldsymbol{R}$ is the nonlinear radiation exchange term. The nonlinearities which arise from the temperature dependence of the emissivity, thermal conductivity, density and heat capacity are not explicitly accounted for at each time instant. Instead these properties are evaluated at the given steady state resulting in the matrices $\boldsymbol{M}(\overline{\boldsymbol{x}})$ and $\boldsymbol{C}(\overline{\boldsymbol{x}})$. The matrix $\boldsymbol{K}$ arises from the steady state contribution of the heat transfer to the ambient boundary conditions and the steady state part of the radiation term $(RD(\tilde{\boldsymbol{x}}))$.

Equation (16) can be reformulated in matrix notation as

$$[\boldsymbol{U}^T\boldsymbol{M}(\overline{\boldsymbol{x}})\boldsymbol{U}]\frac{d\boldsymbol{a}}{dt} = [\boldsymbol{U}^T\boldsymbol{C}(\overline{\boldsymbol{x}})\boldsymbol{U}]\boldsymbol{a} + [\boldsymbol{U}^T\boldsymbol{R}]\{\hat{\boldsymbol{x}}(t)\}^4 + [\boldsymbol{U}^T\boldsymbol{G}]\tilde{P} + [\boldsymbol{U}^T\boldsymbol{K}] \tag{17}$$

where $\boldsymbol{U}$ is the matrix of eigenfunctions, $\boldsymbol{a}$ is the temporal coefficient vector, and $\boldsymbol{G}$ is the lamp power transformation matrix. This set of ordinary differential equations is then integrated using an initial value solver. In order to calculate the contribution, $\{\hat{\boldsymbol{x}}(t)\}^4$, the term, $\hat{\boldsymbol{x}}(t)$, is calculated at each time instant using (6).

IV. Steady-State Performance of Reduced Models

The method outlined above was used to obtain reduced models with ten unknowns from the FEM model with 5060 unknowns. The reduced models showed excellent agreement with the FEM model at steady state operating conditions and for local perturbations around those operating conditions. The FEM model uses a two-band approximation for the partial transmission by quartz in different wavelength ranges. The quartz is treated as transparent for wavelengths shorter then 4 μm and opaque for wavelengths longer than 4 μm [10], [11]. The principal source of deviation between the reduced and FEM models proved to be the nonlinear function which decides the fraction of radiation in each of the two wavelength bands.

In order to arrive at this conclusion, a reduced model was extracted using lumped band radiation properties. In this reduced model, referred to elsewhere in this paper as the "lumped band reduced model," there is a single matrix [$\boldsymbol{R}$ in (17)] which accounts for the total radiation contribution. In the other type of reduced model, the explicit two-band formulation from the FEM model is retained. In this type of reduced model, referred to as the "explicit two-band reduced

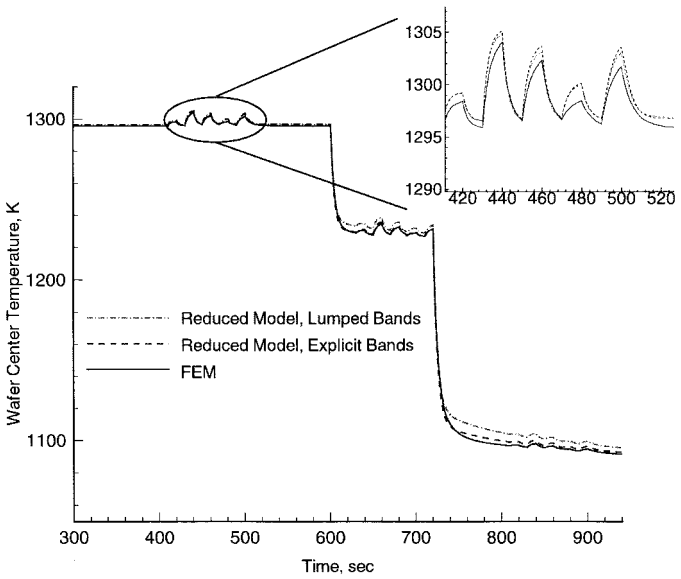

Fig. 3. Comparison of the wafer center temperature of FEM and reduced models. (a) Lumped radiation model. (b) Explicit two band radiation model.

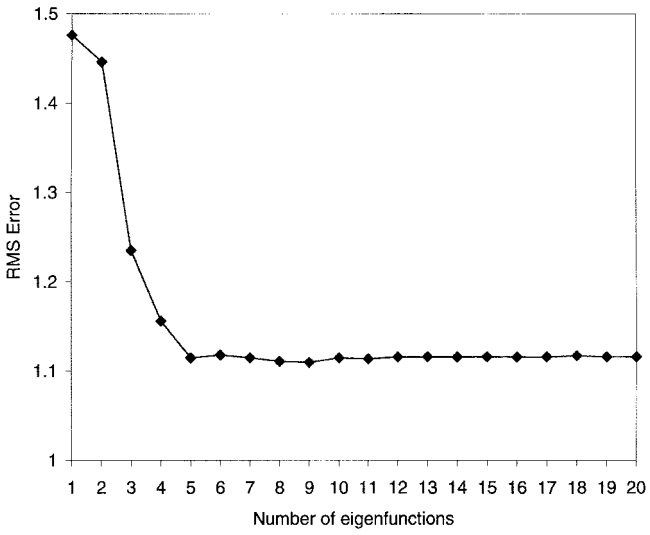

Fig. 4. Variation of rms error of wafer temperature with increasing number of eigenfunctions.

model," there are two matrices ($\boldsymbol{R}_1$ and $\boldsymbol{R}_2$) which separately account for the radiation contributions in the two wavelength bands. The fraction of the radiation contribution in each of the two bands is read dynamically from a look up table indexed to temperature. In the lumped band reduced model the fractions are the same as those at the steady state at which the reduced model is extracted. The wafer center temperatures predicted by the lumped band reduced model and the explicit two-band reduced model are compared to those predicted by the FEM transient model in Fig. 3. Both the reduced models were extracted at a steady state where the wafer temperature was at 1300 K, so that the properties in the matrices $\boldsymbol{M}(\bar{\boldsymbol{x}})$ and $\boldsymbol{C}(\bar{\boldsymbol{x}})$ in (16) were for that steady state. Both the reduced models predict the temperature perturbations at 1300 K steady state operating conditions with reasonable accuracy. As can be seen from the figure, the temperature difference between the reduced and FEM models is within 2 K. The explicit two-band reduced model predicts the wafer center temperature more accurately at other steady state operating conditions, when compared to the lumped band reduced model. Hence, in the rest of the study, the explicit two-band reduced model was used and is referred to as the "reduced model."

The most nonlinear term in the conservation of energy equation, other than the radiation boundary condition, is the inverse of temperature appearing in the gas density. In an attempt to further improve the accuracy of the explicit two-band reduced model, this term was linearized about the steady state. However, this change gave no improvement in the agreement of the reduced and FEM models because the wafer, quartzware, and walls provide the majority of the system mass, and these solids have a constant density in both the formulations. This leads to the conclusion that the deviation of the reduced model temperature trajectory from that predicted by the FEM model for other steady state operating conditions is due to the nonlinear variation of gas phase properties such as thermal conductivity and specific heat.

V. Variation of RMS Error with Increasing Model Order

An important issue in extracting reduced models is deciding upon the number of eigenfunctions to be used in generating the reduced model. The fewer the number of eigenfunctions, the less accurate the reduced model is going to be when compared to the FEM model. On the other hand, a larger number of eigenfunctions would increase the complexity of the reduced model to an extent that it might be too slow to be used in real time process control or other applications. To study this problem, the lamp power was perturbed around a given steady state operating condition, giving rise to local temperature perturbations similar to those shown in Fig. 3. The rms error of the wafer temperature between the reduced and FEM model was calculated as follows:

$$\text{Error} = \sqrt{\frac{\sum_{i=1}^{N}(T_{i,\text{Reduced}} - T_{i,\text{FEM}})^2}{N}} \tag{18}$$

where N denotes the number of points on the wafer surface over which the rms error is evaluated. The rms error was found over 15 points distributed over the wafer surface, and the results were plotted against the number of eigenfunctions as shown in Fig. 4. The error falls steeply till the introduction of the fifth eigenfunction. Following this, there are minor variations in the rms error till the introduction of the tenth eigenufnction. The rms error then settles down at approximately 1.1 after the tenth eigenfunction. The results show that a fifth order reduced model would be good enough for the model reduction strategy, however we chose a tenth order reduced model to perform a rigorous analysis of the technique.

VI. Transient Responses Using Reduced-Order Models

After having studied the response of the reduced models at steady-state operating conditions, the next issue addressed

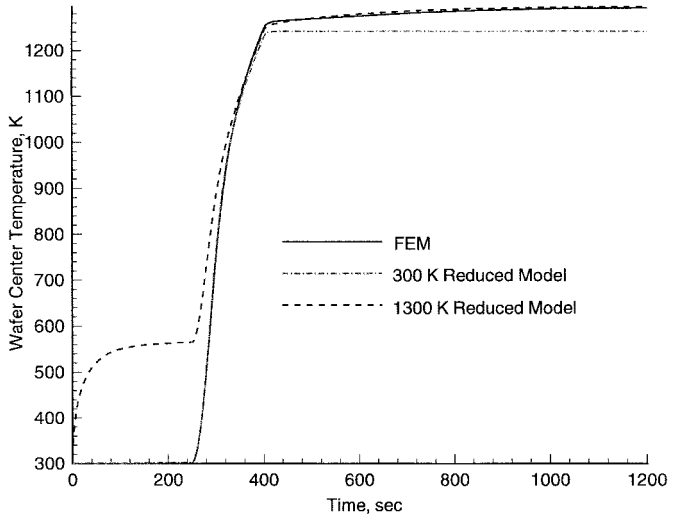

Fig. 5. Behavior of wafer center temperature of FEM and reduced models, during transient ramp up and hold phases.

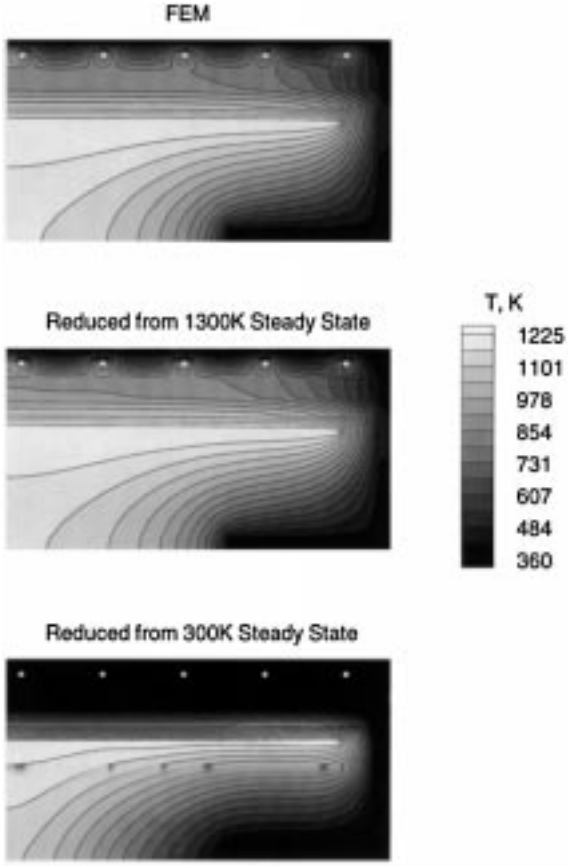

Fig. 6. Temperature flood plots of FEM and reduced models during hold phase.

was the performance of the reduced model in replicating the transient RTP cycle as generated by the FEM model. For this study, a suitable lamp power profile was designed so that the wafer temperature was ramped from 300 to 1300 K at approximately 10 °C/s. After an initial stabilization of the numerical simulations for 250 s, the lamps are turned on and the wafer temperature is ramped from 300 to 1300 K in 150 s and then held constant at 1300 K for 800 s. All the reduced models used to study the transient ramp response were explicit two-band reduced models. A typical RTP cycle is much shorter in duration than the present case study, but the larger cycle was chosen to explore the effect of any drifts which might be present in the reduced model, as they would be amplified over a length of time.

Fig. 5 shows the comparison of the ramp response as shown by the FEM transient model and the reduced model extracted at a wafer steady state of 1300 K. The reduced model attains a different steady state from that given by the FEM model when the lamps are kept at zero power. This is because the fluid properties in the reduced model are linear extrapolations from the 1300-K values instead of the nonlinear power law fits employed in the FEM model. As the lamp powers are ramped, the reduced model shows good agreement with the FEM model and finally attains the same steady state as the FEM model.

In order to further understand the performance of the reduced order modeling scheme, a model was extracted at a wafer steady state of 300 K. This reduced model was then used to study the ramp response. The results are also shown in Fig. 5. As seen from the figure, the initial steady state attained by the reduced model and the FEM model are the same. This is to be expected as the reduced model extracted at 300 K has the same set of properties as the FEM model at 300 K. This reduced model shows good agreement with the FEM model for the lower portion of the ramp, but deviates as the FEM model nears 1300 K. Finally at the higher steady state the reduced model attains a different steady state from the one attained by the FEM model because the linearization of the fluid properties is inadequate at this temperature.

The temperature fields throughout the reactor at the higher temperature steady state, at 800 s into the cycle, are compared in Fig. 6. The reduced model generated at 1300 K shows excellent agreement with the full FEM model, as expected. In the 300-K reduced model, the walls, showerhead, and the quartz window are cooler than both the 1300-K reduced model and the FEM model. This is because the thermal conductivity of the gas phase is over predicted by the 300-K reduced model. This effect is much more predominant in the lamphouse and in the region between the quartz window and the showerhead because the fluid is treated as stagnant in these regions. As a result, the temperature gradients in these regions are determined by radiation and gas phase conduction. In other parts of the reactor this effect is not so evident due to the forced convection in the gas.

These results show that at least two reduced order models need to be combined in order to replicate the FEM ramp response over the entire trajectory. The strategy devised in this regard was to start integrating with the reduced order model extracted at 300 K, then switch to the reduced model extracted at 1300 K when the wafer center temperature is at 1000 K. The switching temperature of 1000 K was chosen because this was the temperature at which the trajectories of both the reduced models intersected the FEM trajectory. Switching models forces the time integrator to restart, and initial values for the 1300-K reduced model coefficients are needed at the switching time. To obtain the coefficients, the transient

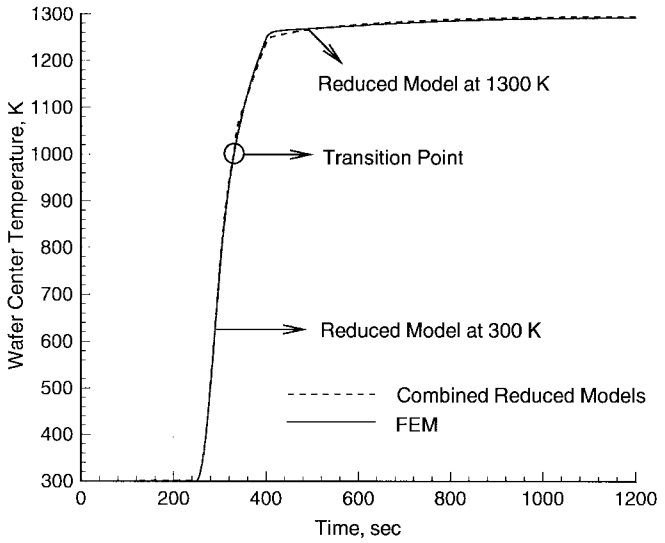

Fig. 7. Behavior of wafer center temperature of FEM and combined reduced models, during transient ramp up and hold phases of the RTP cycle.

temperature field at 1000 K was extracted and the inverse problem was solved in the lower dimensional eigenfunction space. This was done by using the QR-Transform method [23] to determine a least squares solution of (6) for the temporal coefficients.

The results of this strategy are shown in Fig. 7. As can be seen from the figure, the trajectories obtained from the reduced models and the FEM model coincide almost exactly. There is a deviation between the two trajectories immediately after the switch-over, because the integrator has to be reinitialized at the switch-over temperature. As a consequence, the integrator has to start with a new set of coefficients and lacks information about the time derivatives of the coefficients. Also, the least squares solution is not the exact initial state for the given temperature field, when the integrator switches between the two sets of eigenfunctions.

The difference between the FEM and the reduced model temperature trajectories are plotted in Fig. 8. Other than at the switch-over temperature, there is good agreement between the two trajectories. If we ignore the region of the switch-over between the reduced models, the temperature difference is within ± 10 °C for the center and within ± 15 °C for the edge.

The reduced models extracted at wafer steady state temperatures of 1300 K and 300 K were then used to study the cool down phase of the RTP cycle. As seen from Fig. 9, the 1300-K reduced model reaches a higher steady state on cool down compared to the FEM transient model and the 300-K reduced model reaches a lower steady state. Therefore unlike in the ramp up phase, the cool down part of the RTP cycle cannot be replicated by switching between the two models. In order to understand the cool down behavior better, temperature snapshots obtained from the reduced models at the end of the cool down phase (1600 s) are compared to the snapshot obtained at the same time instant from the FEM model in Fig. 10. The 1300-K model shows a much hotter reactor compared to the FEM model. Whereas the 300-K model shows

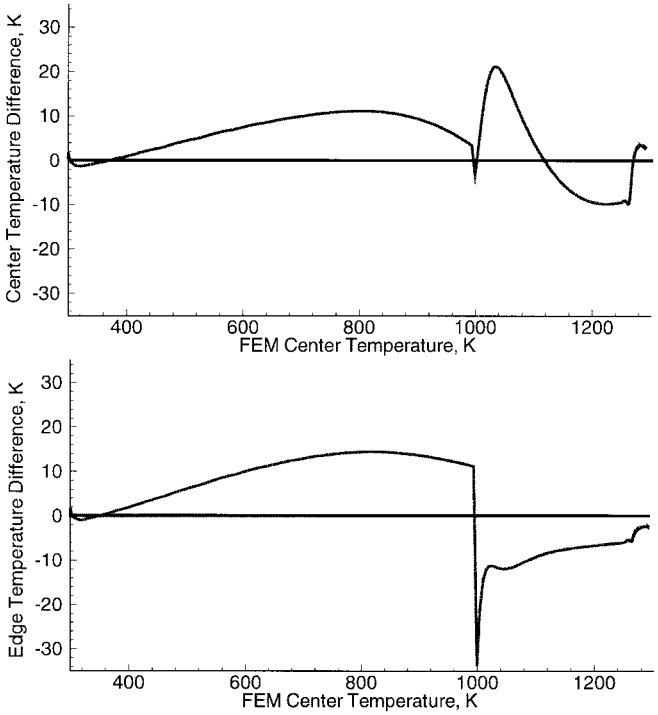

Fig. 8. Difference in wafer temperatures between FEM and combined reduced models during ramp up and hold phases of the RTP cycle.

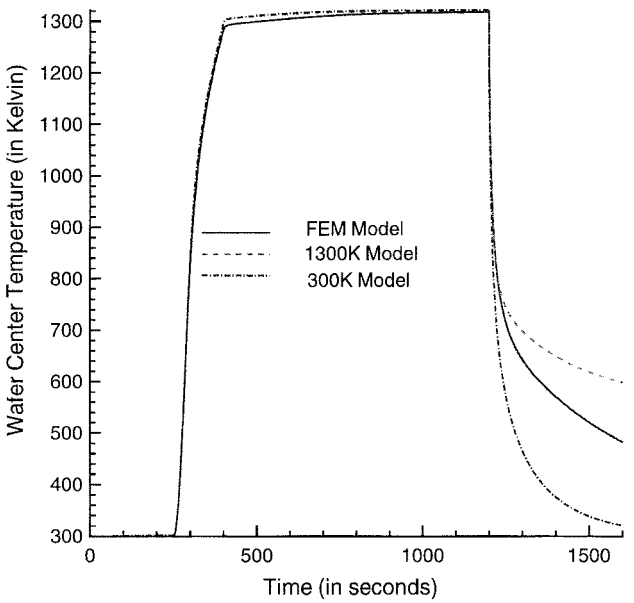

Fig. 9. Behavior of wafer center temperature of FEM and reduced models during ramp up, hold, and cool down phases of the RTP cycle.

a reactor which has reached a nearly uniform temperature of 300 K throughout the reactor. The hottest temperature zone in the 1300-K model is in the lamphouse and the region between the quartz window and the showerhead. These effects are again due to the linearization of the gas phase properties in the reduced models. The 1300-K reduced model underestimates the thermal conductivity, so there is less conductive cooling from the cold walls and the wafer region is warmer than the FEM. This effect is predominant in the lamphouse and in

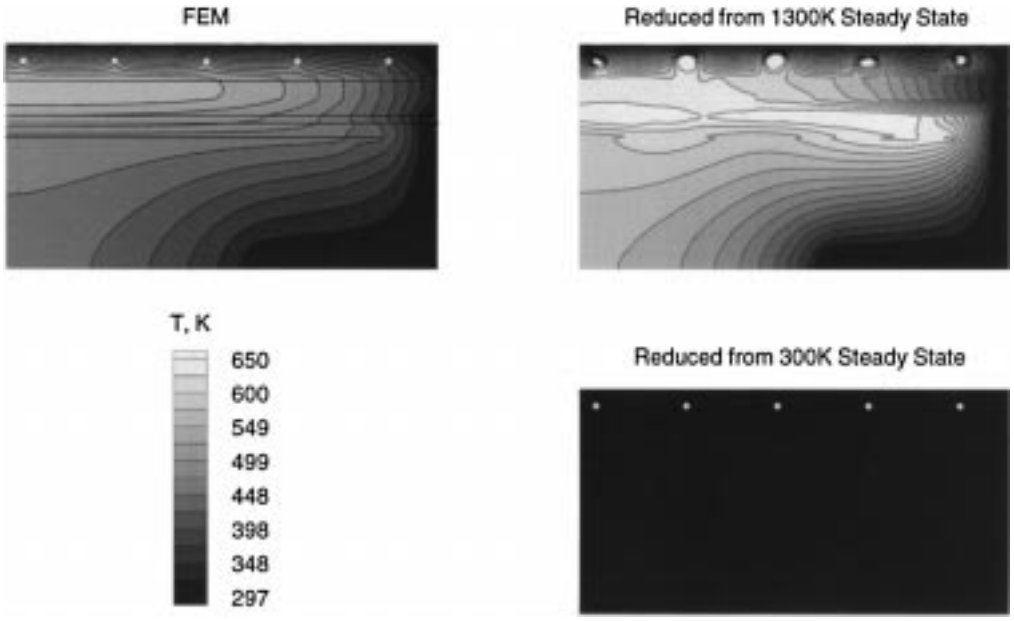

Fig. 10. Temperature flood plots of FEM and reduced models during cool down phase of the RTP cycle.

the region between the quartz window and the showerhead because the gas is treated as stagnant in these regions, leading to the temperature gradients in this region being determined by radiation and gas phase conduction. This leads to the occurrence of the hot-zones in these regions in the 1300-K reduced model. The 300-K model overpredicts the thermal conductivity at elevated temperatures. As a result, there is too much conduction coupling between the wafer and walls, leading to a cooler reactor. In both cases, the most dramatic difference between the FEM and the reduced models are in the lamphouse and in the region between the showerhead and window. In other regions of the reactor, the forced convection of the gas helps in removing some of these effects.

There are several ways of approximating the cool down dynamics using reduced models. One of them would be to use some kind of arithmetic average of the responses of the two reduced models (300- and 1300-K reduced models) to yield a cool down trajectory similar to the FEM model. Hence the strategy to replicate the cool down trajectory was to integrate both the reduced models simultaneously over the entire RTP cycle. For the ramp up phase, a linear interpolating function of wafer temperature was used to determine the contribution of each of the reduced models to the temperature trajectory. At the initial part of the ramp up phase, the temperature predicted by the 300-K reduced model is taken, and at temperatures close to the hold phase, the temperature predicted by the 1300-K reduced model is taken as the overall response of the combined reduced models. In between these two extremes, the interpolating function determines the contribution of the two reduced models in determining the overall temperature trajectory. In the cool down phase, an average of the predictions of the reduced models gives the overall response. The results of this strategy are shown in Fig. 11.

Another strategy to replicate the cool down dynamics would be to extract a reduced model at an intermediate wafer steady

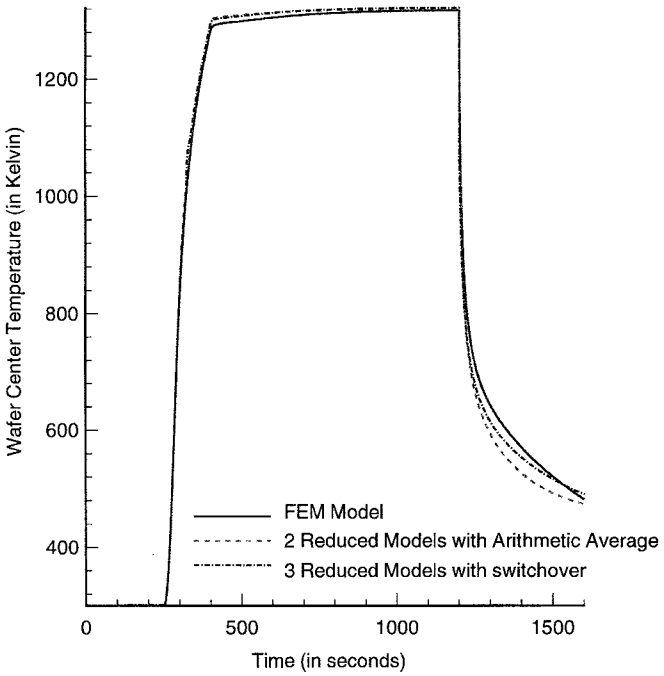

Fig. 11. Transient ramp up, hold, and cool down response of FEM and combined reduced models. (a) Combination of reduced models extracted at 1300 and 300 K with arithmetic averaging. (b) Combination of reduced models extracted at 1300, 1130, and 300 K with switch over.

state, viz. 1130 K, and switch over to this reduced model during the cool down phase. After studying the cool down temperature trajectories predicted by reduced models extracted at different wafer steady states, the reduced model extracted around a wafer steady state of 1130 K was found to have the best agreement with the cool down temperature trajectory as predicted by the FEM model. Therefore, in this strategy we start integrating using the 300-K reduced model and switch to the 1300-K reduced model at 1000 K during the ramp

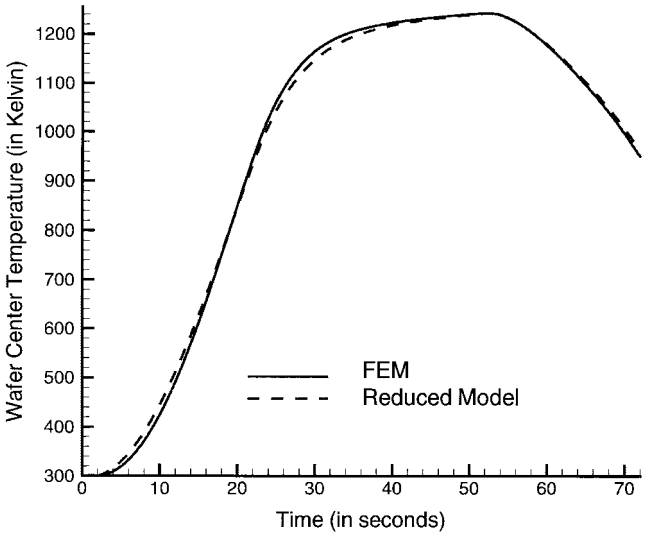

Fig. 12. Replication of RTP ramp cycle using FEM and reduced models.

up phase. The response of the 1300-K model is taken as the overall response of the reduced models during the hold phase. At the end of the hold phase (1200 s), we switch over to the 1130-K reduced model and use it to predict the response for the entire cool down phase. The results are shown in Fig. 11. Both the strategies show reasonable agreement with the FEM model, but the latter strategy is marginally better.

Finally, the reduced models were used to replicate an actual RTP cycle. In this cycle, the lamp powers were ramped from their initial switched-off state to the values corresponding to the wafer steady state of 1300 K in 20 s. The lamp powers were then held at the steady state values for 30 s and then ramped down to the switched-off state in 20 s. The effect of this power protocol on the wafer center temperature for both the FEM and reduced models are shown in Fig. 12. For the reduced model strategy, the 300-K reduced model was used to replicate the initial wafer steady state. At the start of the ramp up, the integrator was switched over to the 1300-K reduced model and this was used to replicate the entire trajectory from then on. The figure shows good agreement between the FEM and reduced model responses and further validates the efficacy of the reduced model strategy in replicating RTP transients.

VII. Reduction in Computation Time

The primary motivation for developing the technique of reduced order model extraction is to obtain reduced models with good predictive capabilities which have significantly less computation times compared to the FEM model. Therefore, timing runs were carried out both during steady state operating conditions and during transient ramp up to determine the reduction in computation time. For 200 s of real process time at the steady-state operating temperature of 1300 K, the following were the computation times for the FEM model and the reduced model extracted at 1300 K to perform local temperature perturbations of the kind shown in Fig. 3. The FEM model and the reduced model were simulated on a HP-735 workstation.

TABLE I

Models	Computation Time
FEM Transient Model	45 min
1300 K Reduced Model	0.75 min

TABLE II

Models	Computation Time
FEM Transient Model	22.17 min
Reduced Model with switch-over	1.8 min
Combined Reduced Models	3.73 min

As shown in Table I, the time required for execution of the reduced model is nearly two orders of magnitude less than the FEM model. Timing runs were also carried out to compare the reduction in computation time between the combined reduced models used for the cool down study, the reduced model with switch over used to study the ramp up phase and the FEM transient model. For a real processing time of 150 s for the ramp up between 300 and 1300 K the computation times on a HP-735 workstation are shown in Table II.

Both the reduced models show nearly an order of magnitude decrease in computation time when compared to the FEM model. The computation time doubles in the case of the combined reduced models, as two reduced models, and hence two sets of differential equations, have to be integrated simultaneously. The computation time for this combined model can be decreased by choosing lesser number of eigenfunctions in each of the reduced models, hence leading to a smaller number of differential equations in each of the two sets. The reduced model with switch-over integrates faster than real time and shows promise of being useful in model based control.

The main overhead in terms of computation time comes in the reduced model extraction stage. A typical snapshot generation and eigenfunction extraction run can take hours. In our case, generation of 220 temperature snapshots and eigenfunction extraction from the transient FEM model took ~ 8 h on a HP-735 workstation. Hence reduced models are good for applications in which the models have to be executed repetitively, as in model based controllers, or to study process changes under small perturbations. This would involve extracting a few reduced models at predetermined steady state operating conditions once, and then using them for the desired applications repetitively, thereby cutting down on the overhead.

VIII. Conclusion

A strategy for extracting lower dimensional physically based reduced-order models from complex finite element models has been developed. RTP was used as a test vehicle because of its dynamic nature, but the reduced model extraction procedure can be applied to any other process which can be described by similar fluid-thermal conservation equations. The reduced models (ten unknowns) showed very good agreement with the FEM model (5060 unknowns) not only around the steady state operating conditions from which they were extracted, but also at other steady operating conditions. This technique is superior to other strategies, such as lumping of nodes

within the FEM framework or assuming certain variables constant, because it does not simplify any of the physical conservation equations and the eigenfunction sets used to expand the equations carry qualitative information about the solution fields. A single reduced model can, therefore, be used for process optimization studies and answering "what if" type of process questions spanning a large window in process space ($\pm 100^{\circ}$C). We have shown how the entire RTP cycle (ramp up, hold and cool down) can be simulated using combinations of a few reduced models in real time on workstations. The reduced models have computation times which are an order of magnitude less than the FEM model. The reduced model strategy can be used in a combined feedforward and feedback control application. In such a strategy, the reduced models described in this paper would be used to provide the feedforward trajectory and a simple PID controller would be used to implement feedback control around this predicted trajectory. Due to the linearization of the gas phase thermal properties, the temperature response of the reduced models would tend to become more and more inaccurate as the range of operation is stretched beyond the conditions around which the linearization is done. Therefore, by explicitly accounting for these nonlinearities, the response of the reduced models can be improved. However, this would introduce further complexities in the reduced model and increase their computation time. Intelligent "model switching" could provide a viable alternative to circumvent this tradeoff problem.

ACKNOWLEDGMENT

The authors would like to thank H. Aling of Integrated Systems, Inc. and I. G. Kevrekidis of Princeton University for helpful discussions.

REFERENCES

[1] I. Calder, "Rapid thermal process integration," *Reduced Thermal Processing for ULSI*, R. A. Levy, Ed. New York: Plenum, 1989, pp. 181–226.
[2] F. Roozeboom, "Introduction: History and perspectives of RTP," in *Proc. NATO Advanced Study Institute, Advances in Rapid Thermal and Integrated Processing*, F. Roozeboom, Ed. Dordrecht, The Netherlands: Kluwer, 1996.
[3] P. Singer, "Rapid thermal processing: A progress report," *Semicond. Int.*, May 1993, pp. 64–69.
[4] T. Breedijk, T. F. Edgar, and I. Trachtenberg, "A model predictive controller for multivariable temperature control in rapid thermal processing," in *Proc. Amer. Control Conf.*, June 1993, pp. 2980–2984.
[5] C. Schaper, "Real time control of rapid thermal processing semiconductor manufacturing equipment," in *Proc. Amer. Control. Conf.*, June 1993, pp. 2985–2989.
[6] C. Schaper, M. Moslehi, K. Saraswat, and T. Kailath, "Control of MMST RTP: Repeatability, uniformity, and integration for flexible manufacturing," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 202–219, May 1994.
[7] C. Schaper, M. Moslehi, K. Saraswat, and T. Kailath, "Modeling, identification, and control of rapid thermal processing systems," *J. Electrochem. Soc.*, vol. 141, no. 11, pp. 3200–3209, Nov. 1994.
[8] G. Aral, T. P. Merchant, J. V. Cole, K. L. Knutson, and K. F. Jensen, "Concurrent engineering of an RTP reactor: Design and Control," in *Proc. RTP-'94*, 1994, pp. 288–295.
[9] H. Aling, J. Abedor, J. L. Ebert, A. Emami-Naeni, and R. L. Kosut, "Application of feedback linearization to model of Rapid Thermal Processing (RTP) reactors," in *Proc. RTP-'95*, 1995, pp. 356–366.
[10] T. P. Merchant, J. V. Cole, K. L. Knutson, J. P. Hebb, and K. F. Jensen, "A systematic approach to simulating Rapid Thermal Processing systems," *J. Electrochem. Soc.*, vol. 143, no. 6, pp. 2035–2043, June 1996.
[11] K. F. Jensen, T. P. Merchant, J. V. Cole, J. P. Hebb, K. L. Knutson, and T. G. Mihopoulos, "Modeling strategies for Rapid Thermal Processing: Finite element and Monte Carlo methods," in *Proc. NATO Advanced Study Institute, Advances in Rapid Thermal and Integrated Processing*, F. Roozeboom, Ed. Dordrecht, The Netherlands: Kluwer, 1996.
[12] L. Sirovich, "Turbulence and the dynamics of coherent structures: I, II and III," *Quart. Appl. Math.*, vol. XLV, no. 3, p. 561, 1987.
[13] P. S. Wyckoff, "Numerical solution of differential equations through empirical eigenfunction expansions," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1995.
[14] J. L. Lumley, *Transition and Turbulence*, R. E. Meyer, Ed. New York: Academic, 1981, pp. 215–242.
[15] L. Sirovich and H. Park, "Turbulent thermal convection in a finite domain: Part I. Theory," *Phys. Fluids A*, vol. 2, no. 9, pp. 1649–1658, Sept. 1990.
[16] H. Park and L. Sirovich, "Turbulent thermal convection in a finite domain: Part II, numerical results," *Phys. Fluids A*, vol. 2, no. 9, pp. 1659–1668, Sept. 1990.
[17] C. Canuto, M. Y. Hussaini, A. Quarderonic, and T. A. Zang, *Spectral Methods in Fluid Dynamics*. New York: Springer, 1988.
[18] L. Sirovich, Ed., "Empirical eigenfunctions and low dimensional systems," in *New Perspectives in Turbulence*. New York: Springer, 1991.
[19] L. Sirovich, J. D. Rodriguez, and B. Knight, "Two boundary value problem for Ginzburg Landau equation," *Physica D*, vol. 43, pp. 63, 1990.
[20] J. D. Rodriguez and L. Sirovich, "Low dimensional dynamics for the complex Ginzburg Landau equation," *Physica D*, vol. 43, p. 77, 1990.
[21] L. Sirovich, "Chaotic dynamics of coherent structures," *Physica D*, vol. 37, pp. 126–145, 1989.
[22] N. Aubry, P. Holmes, J. L. Lumley, and E. Stone, "The dynamics of coherent structures in the wall region of a turbulent boundary layer," *J. Fluid. Mech.*, vol. 192, pp. 115–173, 1988.
[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

Suman Banerjee received the B.Tech. degree in chemical engineering in 1993 from the Indian Institute of Technology, Madras, India. He received the M.S. degree in chemical engineering from the Massachusetts Institute of Technology, Cambridge, in 1994. He is currently pursuing the Ph.D. degree in chemical engineering under the supervision of Prof. K. F. Jensen at MIT.

His research interests include process development and modeling related to semiconductor manufacturing.

J. Vernon Cole received the B.S. degree in chemical engineering from Clemson University, Clemson, SC, in 1987 and the Ph.D. degree in chemical engineering from the University of Florida, Gainesville, in 1993.

After postdoctoral studies at the Massachusetts Institute of Technology, Cambridge, he joined the Motorola Advanced Products Research and Development Laboratory, Austin TX, in 1996. His professional career centers around the equipment scale simulation of thermal reactive processes in semiconductor device fabrication. This involves a variety of processes and equipment, including oxidation, chemical vapor deposition, and annealing in both furnaces and single-wafer equipment. Since July 1997, he has been continuing these activities in the Motorola Predictive Engineering Laboratory.

Klavs F. Jensen, for photograph and biography, see p. 107 of the February 1998 issue of this TRANSACTIONS.