TECHNICAL RESEARCH REPORT

Model Reduction for a Tungsten Chemical Vapor Deposition System

by Hsiao-Yung Chang, Raymond A. Adomaitis

T.R. 98-28
Model Reduction for a Tungsten Chemical Vapor Deposition System

Hsiao-Yung Chang and Raymond A. Adomaitis

Department of Chemical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract: A model of a tungsten chemical vapor deposition (CVD) system is developed to study the CVD system thermal dynamics and wafer temperature nonuniformities during a processing cycle. We develop a model for heat transfer in the system's wafer/susceptor/guard ring assembly and discretize the modeling equation with a multiple-grid, nonlinear collocation technique. This weighted residual method is based on the assumption that the system's dynamics are governed by a small number of modes and that the remaining modes are slaved to these slow modes. Our numerical technique produces a model that is effectively reduced in its dynamical dimension, while retaining the resolution required for the wafer assembly model. The numerical technique is implemented with only moderately more effort than the traditional collocation or pseudospectral techniques. Furthermore, by formulating the technique in terms of a collocation procedure, the relationship between temperature measurements made on the wafer and the simulator results produced with the reduced-order model remain clear.

Keywords: Model reduction, Chemical vapor deposition, Weighted residual methods

1. Introduction

This paper presents a distributed parameter system model reduction technique developed in the context of simulating a commercial Chemical Vapor Deposition (CVD) system. In the system studied, single semiconductor wafers are processed in a cold-walled chamber with wafer heating provided by one ring-shaped incoherent lamp bank.

Numerous simulation studies aimed at understanding or controlling wafer temperature uniformity in radiant-heated systems have been conducted (Breedijk et al., 1993, Cole et al., 1994, Dilliac et al., 1995; Merchant et al., 1996; Sorrell et al., 1992). Because of the complexity of some of the simulators developed and the computational expenses of performing simulations for process optimization and control, interest in developing reduced-order models from the high fidelity simulations has emerged (Aling et al., 1996, 1997; Theodoropoulos et al., 1998). These models are based on collecting “snapshots” of the RTP system at points in time during processing cycles designed to excite as many spatial temperature modes as possible, and processing them with the Proper Orthogonal Decomposition method (Sirovich, 1987) to determine the dominant temperature spatial modes. Because the original modeling equations are projected onto these modes with Galerkin’s method, the reduced-order models have a predictive capability not found in models generated from purely empirical model identification schemes.

In this paper, we develop a model representing the thermal dynamics of a commercial CVD system. We propose a multiple-grid, nonlinear collocation technique in the context of developing a reduced-order simulator, where the goal of the reduction procedure is to minimize the number of dynamic modes required for accurate dynamic simulations of this system. The proposed method is implemented using numerical procedures similar to the traditional collocation and pseudospectral techniques. An additional advantage of the technique is that the collocation formulation retains a direct connection between temperature measurements made on the wafer and simulation results produced with the reduced-order model.

2. ULVAC System for Tungsten CVD

Tungsten (W) is used both as a contact plug material and as a first-level metal for interconnects in microelectronic devices. In the selective W CVD processes, W selectively grows on Si but not on SiO₂ sur-
faces, thus can be used to fill via contacts in a single step. In blanket deposition processes, a metal nucleation layer is first deposited on the entire wafer. This allows W to be blanket-deposited over the wafer surface, filling the contact holes. A subsequent reactive ion etchback step is then used to complete the plugs.

Our modeling research focuses on the BTU-ULVAC ERA-1000 selective tungsten deposition system, shown schematically in Fig. 1. Reactant gases are introduced into the reactor from two sources: a gas mixture of SiH$_4$ and WF$_6$ is injected through a slit-like nozzle on the side wall, and H$_2$ is pumped in through a showerhead at the top of reactor chamber. Gases mix in the chamber and react at the surface of a 4 inch diameter wafer, which is supported by a slowly rotating quartz susceptor. A portion of the wafer near its outer edge is covered by a quartz guard ring. The wafer is heated to 600 K by a ring of incoherent tungsten-halogen lamps through the transparent showerhead. The CVD runs last for approximately 5 minutes after the operating temperature is reached.

Initial modeling of the ULVAC tungsten CVD system has focused on the gas phase flow field and heat transfer, and the wafer thermal dynamics. Because of the short gas residence time relative to the wafer thermal dynamics, the gas phase flow field and temperature can be described by the steady-state conservation of total mass, momentum, and energy, and take the form

\[ \nabla \cdot (\rho \mathbf{v}) = 0 \]
\[ -\nabla \cdot \mu (\nabla \mathbf{v}) + \nabla P = 0 \]
\[ \nabla \cdot (\rho C_p \mathbf{v} T_g) - \nabla \cdot (\kappa \nabla T_g) = 0 \]

where \(\rho\), \(\mathbf{v}\), \(P\), \(\mu\), \(C_p\), \(\kappa\) represent gas density, velocity, pressure, viscosity, heat capacity, and thermal conductivity, respectively (Kleijn and Werner, 1993). A Galerkin discretization technique was used with globally defined trial functions to solve the gas flow field equations under the assumption of fully-developed, laminar flow and subject to no-slip boundary conditions at the chamber walls. A two-dimensional eigenfunction expansion method combined with a one-point spanwise collocation discretization was used to compute the gas temperature profile, subject to the flow field calculated above and gas temperature boundary conditions defined by representative operating conditions (Chang and Adomaitis, 1997). Typical gas temperature contours and wafer/gas energy transfer rates are displayed in Fig. 2.

3. Wafer Assembly Temperature Model

The wafer/susceptor/guard ring assembly model is based on assuming the assembly is thin relative to its diameter (\(\Delta_r \ll \) guard ring radius) and is derived by a differential element energy balance. This assumption implies a uniform temperature distribution across assembly thickness. Since there are different geometric combinations of the guard ring, wafer, and susceptor in the radial direction, an assembly
thickness parameter ($\Delta_z$) and mass-averaged, effective material properties are defined for each section (see Figure 3) making these parameters functions of $r$. Assuming no azimuthal temperature variations exist, the wafer temperature governing equation in cylindrical coordinates is

$$\Delta_z(r)\rho(r)C_p(r)\frac{\partial T_w}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( k(r)\Delta_z(r)r \frac{\partial T_w}{\partial r} \right) - \Delta Q$$

where the energy flux from the top and bottom of the assembly is defined as

$$\Delta Q = Q_{cd} + Q_{cv} + Q_{em} - Q_{lp}$$

These quantities are listed in Table 1 for the different regions. Note that the $q_{lp}$ function represents the lamp radiant flux distribution and the lamp power input is $u(t)$. The temperature of chamber floor and electrode surface (located under the wafer and used as part of the chamber plasma cleaning cycle) is assumed to be equal to the wall temperature by the design of the cooling system.

Because the quartz showerhead will absorb most of the infrared component of the light emitted by the heating lamp, we assume the primary component of radiant energy that reaches the assembly is in the visible portion of the spectrum, meaning that little will be absorbed by the susceptor or guard ring. However, the radiation emitted by the heated wafer will have primarily longer-wavelength components and so will be readily absorbed by the quartz guard ring, susceptor, and chamber walls.

The lamp radiation intensity at specific points on the wafer surface will depend on the lamp power output, the distance from the wafer point to each bulb, and the angle of incidence the ray makes with the wafer surface. The procedures for computing the flux intensity profile generated by the 6kW bulb ring is given in Chang (1997).

The temperature equation is made dimensionless by dividing the radial coordinate by the guard ring radius $R_{gr}$; the assembly thickness function $\Delta_z(r)$ by the wafer thickness, $Z_{wafer}$; and time by the time constant $\tau = (\rho SiCPSiR_{gr})/k_Si$. The effective property functions (except for the heat transfer coefficients) are made dimensionless by dividing each function by the corresponding thermophysical property of silicon evaluated at ambient conditions (see the curves in Fig. 3 and the constants listed in the center figure legend). The constants $C_{cd}$, $C_{cv}$, $C_{em}$, and $C_{lp}$ result from the nondimensionalization. The resulting wafer assembly temperature model is as follows

$$\rho(r)C_p(r)\frac{\partial T_w}{\partial t} = \frac{1}{\Delta_z(r) r} \left( k(r)\Delta_z(r)r \frac{\partial T_w}{\partial r} \right) + C_{cd}\frac{h_{cd}(r)}{\Delta_z(r)}(T_{wall} - T_{w})$$

subject to homogeneous boundary conditions

$$\frac{\partial T_w}{\partial r} = 0 \quad \text{at } r = 0, 1 \quad (2)$$

with initial condition $T_w(r, t = 0) = T_0(r)$ at $r < 1$. The dimensionless physical and geometric constants used in calculations are plotted in Figure 3. The dimensional values in SI units are obtained by multiplying corresponding factors listed in the legend box.

![Figure 3: Detailed sketch of the wafer/guard ring/susceptor assembly (top), the corresponding physical properties of the different regions (center), and radiant energy flux distribution (bottom).](image)
The transformation array is computed as solutions to (4) to obtain the same number of ordinary differential equations in time subject to (7) defines the nonlinear Galerkin procedure (Aling, et al. 1997; Deane, et al., 1997). Solutions are sought in the form of the truncated trial function expansion \( T_w(r,t) = \sum_{j=1}^{\infty} a_j(t) \psi_j(r) \). (3)

where the trial functions \( \psi_j(r) \) are chosen as the normalized Bessel functions of the first kind of order 0:

\[
\psi_j(r) = \frac{J_0(\gamma_j r)}{\sqrt{J_0(\gamma_1 r)J_0(\gamma_J r)}}.
\]

The \( \gamma_j \) are computed as solutions to \( J_1(\gamma) = 0 \); this choice of trial functions means that the guard-ring edge no-flux boundary condition and the condition of symmetry at \( r = 0 \) are satisfied by each trial function. The trial functions form an orthonormal sequence under the inner product \( \langle f , g \rangle = \int_0^1 f g r dr \).

4. Interior Collocation Formulation

We now write the wafer temperature in terms of the orthonormal function series \( \psi_j \), truncated after the \( N \)th term, in vector notation \( T_w^N = \psi^1 \times N a^{N \times 1} \). The transformation array \( Q^N \times N \) is defined by \( Q_{i,j} = \psi_j(r_i) \) which equates mode amplitudes \( a^{N \times 1} \) to \( N \) discrete temperature points on the wafer \( T_w^N \) by \( T_w = Qa \). Temperature derivatives and the heat equation operator \( \nabla^2 \) can be written in terms of these discrete temperature values through the use of the differentiation arrays \( A \) and \( B \) defined by

\[
A = \frac{dQ}{dz} Q^{-1} \quad B = [\nabla^2 Q] Q^{-1}.
\]

The \( N \) interior collocation points are chosen as the roots of the \( \psi_{N+1} \) trial function (Villadsen and Stewart, 1967), and the discretized nonlinear modeling equation can be written as

\[
\frac{dT_w}{dt} = (A\Theta_j + \kappa_j \Delta_j B_j T_w + C_{em} \varepsilon(T_{wall}^4 - T_{w_j}^4) + C_{cd} h_{cd}(T_{wall} - T_{w_j}) + C_{cv} h_{cv}(T_g + 1 - T_{w_j}) + C_{lp} q_{lp}(u(t)) \} / \rho_j C_p \Delta_j \)
\]

subject to initial conditions \( T_{w_j} = 1, j = 1, \ldots, N \). Note that \( [A\Theta_j] = [A_j(\kappa_\Delta_S)] [A_j T_w] \), \( j = 1, \ldots, N \) in the above equation, and that \( A_j \) denotes the \( j \)th row of the discretization equation \( A \).

5. Nonlinear Weighted Residual Methods

The wafer temperature modeling equation (1) can be rewritten as

\[
\dot{T}_w(t, r) = G(T_w).
\]

By writing out (6) explicitly, we obtain

\[
\dot{a}_l^s = g_l(a^s, a^f) \quad l = 1, \ldots, L, \quad L < N
\]

\[
\dot{a}_n^f = h_n(a^s, a^f) \quad n = L + 1, \ldots, N
\]

and can make a distinction between the relatively “fast” dynamical modes \( a_n^f \) and the slower modes \( a_n^s \). The fast modes are associated with the highly-oscillatory (spatial) trial functions \( \psi_n \) which dissipate more quickly by heat conduction through the wafer relative to the slower, less spatially-oscillatory modes. If we assume the system’s dynamics are governed by the slower modes, the fast modes can be slaved to the slow by setting \( \dot{a}_l^f = 0 \). This means solutions \( T^N(t, r) \) are constrained to the manifold in phase space defined by the nonlinear equations

\[
0 = h_n(a) \quad n = L + 1, \ldots, N.
\]

Integrating the remaining ordinary differential equations in time subject to (7) defines the nonlinear Galerkin procedure (Aling, et al. 1997; Deane, et al., 1997).

At this point we note there are no obvious connections between the collocation-discretized equations and the combined slow and fast dynamics sets of ordinary differential equations in time produced by the Galerkin procedure. In other words, there is no guarantee that the fast modes would be slaved to the slow if \( \hat{T}_{w_n} \) were set to zero at \( N - L \) interior collocation points. Moreover, no obvious selection procedure for choosing these points is apparent in this formulation.

5.1. Nonlinear Collocation

At the start of each time step during numerical integration of the \( N \)-mode discretized wafer temperature equation, we would like to take a smaller number \( L \) of points on the solution curve \( T_w(r, t) \) and compute the time derivatives at these points based on the temperature profile \( T_w \) defined on the intermediate grid \( r \) that is generated when the (spatially) higher-frequency modes are slaved to the slow in a quasi steady-state manner (see Fig. 4). These dynamic equations are defined using the most-coarse grid \( \hat{r}_l \) (typically determined as the roots of \( \psi_{L+1} \)) with \( l = 1, \ldots, L \). The discretized equations defining the time derivatives at the coarse-scale grid points are

\[
\frac{d\hat{T}_w}{dt} = \hat{Q}Q^{-1}[F(T_w)]
\]

with \( \hat{Q}_{i,j} = \psi_j(\hat{r}_i), i = 1, \ldots, L, j = 1, \ldots, N \), and \( F(T_w) \) defined by (4). The vector \( T_w \) corresponds to values of \( T_w^N(r, t) \) defined on the intermediate-resolution grid \( r_n, n = 1, \ldots, N \). The \( N \) solution points \( T_{w_n} \), defining this temperature profile are computed from the \( N \) equations:

\[\begin{align*}
\text{[L equations]} & \quad T_w^{N}(r, t) \text{ passes through the points} \\
& \quad (\hat{r}_l, \hat{T}_{w_l}), \ l = 1, \ldots, L; \\
\text{[N - L equations]} & \quad \hat{a}_n = 0 \text{ for } n = L + 1, \ldots, N.
\end{align*}\]

This can be done in the physical space by the following \( N \) equations for \( T_w \) at the start of each time-step:

\[
0 = \hat{T}_w - \hat{Q}Q^{-1}T_w \\
0 = \hat{P} [F(T_w)]
\]

In this set of nonlinear equations, \( \hat{F}(T_w) \) is the residual produced by the \( N \)-mode collocation solution evaluated on the finest grid \( \hat{f} \) and is defined as

\[
\hat{F}_j(T_w) = \left\{ [\hat{A}\hat{\Theta}]_j + \hat{\alpha}_j \Delta z_j \hat{B} \hat{T}_w + C_{cm} \epsilon_j (\hat{T}_{wall}^4 - \hat{T}_w^4) + C_{cd} h_{cd} (\hat{T}_{wall} - \hat{T}_{w}) + C_{CV} h_{cv} (\hat{T}_w + 1 - \hat{T}_{w}) + C_{LP} \hat{\alpha}_j q_{LP} u(t) \right\} / \hat{P}_j \hat{C}_P \Delta z_j
\]

with \( j = 1, \ldots, M \). The \( \hat{f} \) notation refers to the \( r \)-dependent parameter and state variables evaluated on the finest grid. The points \( \hat{r}_m \) are determined as the roots of \( \psi_{M+1}(r) \); \( [\hat{A}\hat{\Theta}]_j \) are elements of the \( M \times 1 \) vector produced by the term-by-term product of the vectors \( \hat{A}\hat{\alpha}\Delta z \) and \( \hat{A}T_w \).

Using this finer grid \((M > N) \) allows for residual calculations that are more accurate than would be possible using the intermediate-resolution grid. Therefore, a more accurate discrete analog to the Galerkin projection can be developed to satisfy the \( N - L \) equations: \( \hat{a}_n = 0 \text{ for } n = L + 1, \ldots, N \).

To compute the fine-grid residual values corresponding to a solution defined on the original, intermediate-resolution grid \( r \), we define the interpolation array \( \hat{T}_w^{M \times N} = \hat{Q}^{-1} \hat{T}_w \) with \( \hat{Q}_{m,n} = \psi_n(\hat{r}_m), n = 1, \ldots, N \), and \( m = 1, \ldots, M \). This defines the transformation array \( \hat{Q}Q^{-1} \) that takes the function \( T_w(r) \) defined by the intermediate-discretization grid \( r \) values of \( T_{w_n} \) and exactly interpolates the function to the fine grid. Similarly, we can generate the non-square discretization array necessary for computing the residual by

\[
\hat{A} = \left[ \begin{array}{c} d\hat{Q} \end{array} \right] Q^{-1} \quad \text{and} \quad \hat{B} = \left[ \begin{array}{c} 1 \ d\hat{Q} \end{array} \right] Q^{-1}.
\]

The discrete Galerkin projection is implemented by defining the nonsquare array \( \hat{P}^{(N-L \times M)} \) by rows \( L + 1 \) through \( N \) rows of \( \hat{P}^{-1} \); in other words, \( \hat{P}_{n,m} = \left[ \hat{P}^{-1} \right]_{n,m}, n = L + 1, \ldots, N \), and \( m = 1, \ldots, M \) with \( \hat{P}_{i,j} = \psi_j(\hat{r}_i) \) with \( i, j = 1, \ldots, M \).

6. Results and Conclusions

Representative results in the form of snapshots showing the wafer assembly temperature profiles during the startup and soak phases of a processing run are presented in Fig. 5. We conclude that the numerical technique developed produces accurate, low-order simulation results, especially at the collocation points of the coarse grid \( \hat{r} \) which is most relevant to comparisons with experimental measurements. Further
refinements to the numerical technique are currently under investigation, such as using the “acceleration” POD modes in place of the Bessel’s functions $\psi_n$ (Aling, et al. 1997; Sirovich, 1991).

**References**


