

TECHNICAL RESEARCH REPORT

Modeling and Model Reduction for Control and Optimization of Epitaxial Growth in a Commercial Rapid Thermal Chemical Vapor Deposition Reactor

*by A. Newman, P.S. Krishnaprasad, S. Ponczak,
P. Brabant*

T.R. 98-45



ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.

ISR is a permanent institute of the University of Maryland, within the Glenn L. Martin Institute of Technology/A. James Clark School of Engineering. It is a National Science Foundation Engineering Research Center.

Web site <http://www.isr.umd.edu>

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE 1998	2. REPORT TYPE	3. DATES COVERED -			
4. TITLE AND SUBTITLE Modeling and Model Reduction for Control and Optimization of Epitaxial Growth in a Commercial Rapid Thermal Chemical Vapor Deposition Reactor		5a. CONTRACT NUMBER			
		5b. GRANT NUMBER			
		5c. PROGRAM ELEMENT NUMBER			
6. AUTHOR(S)		5d. PROJECT NUMBER			
		5e. TASK NUMBER			
		5f. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Army Research Office, PO Box 12211, Research Triangle Park, NC, 27709		8. PERFORMING ORGANIZATION REPORT NUMBER			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)			
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)			
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES The original document contains color images.					
14. ABSTRACT see report					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified		63	

Modeling and Model Reduction for Control and Optimization of Epitaxial Growth in a Commercial Rapid Thermal Chemical Vapor Deposition Reactor

Andrew J. Newman *

Institute for Systems Research and
Electrical Engineering Department
University of Maryland
College Park, MD 20742
newman@isr.umd.edu

Sam Ponczak

Northrop Grumman Corp.
Electronic Sensors and Systems Division
Baltimore, MD 21203

P. S. Krishnaprasad

Institute for Systems Research and
Electrical Engineering Department
University of Maryland
College Park, MD 20742
krishna@isr.umd.edu

Paul Brabant

Northrop Grumman Corp.
Electronic Sensors and Systems Division
Baltimore, MD 21203

Original Version: October 28, 1997

Current Version: September 8, 1998

Abstract

In December 1996, a project was initiated at the Institute for Systems Research (ISR), under an agreement between Northrop Grumman Electronic Sensors and Systems Division (ESSD) and the ISR, to investigate the epitaxial growth of silicon–germanium (Si–Ge) heterostructures in a commercial rapid thermal chemical vapor deposition (RTCVD) reactor. This report provides a detailed account of the objectives and results of work done on this project as of September 1997. The report covers two main topics - modeling and model reduction. Physics–based models are developed for thermal, fluid, and chemical mechanisms involved in epitaxial growth. Experimental work for model validation and determination of growth parameters is described. Due to the complexity and high computational demands of the models, we investigate the use of model reduction techniques to reduce the model complexity, leading to faster simulation and facilitating the use of standard control and optimization strategies. Some of the contents of this report are contained in [34].

*This research was supported by grants from the Northrop Grumman Foundation, the National Science Foundation's Engineering Research Centers Program: NSFD CDR 8803012 and NSF Grant EEC-9527576, and the Army Research Office under the ODDR&E MURI97 Program Grant No. DAAG55-97-1-0114.

Contents

1	Introduction	1
1.1	Objectives	1
1.2	Participants	1
1.3	Equipment	1
1.4	Project Evolution	2
1.5	Report Outline	4
2	Physics Based Models	4
2.1	Modeling Overview	4
2.2	Modeling Issues	5
2.2.1	Complications	5
2.2.2	Simplifications	7
2.3	Process and Equipment Models	7
2.3.1	Implementation	8
2.3.2	General Purpose and Specialized CFD	9
2.3.3	Wafer Heat Transfer	12
2.3.4	Lamp Heating	17
2.3.5	Chemical Reaction Kinetics	26
2.3.6	Preliminary Results	27
2.4	Experimental Validation	29
2.4.1	Deposition Kinetics	31
2.4.2	Lamp Heating	33
2.5	Feature Scale Models	36
3	Model Reduction	37
3.1	Motivation and Overview	40
3.2	Simplified Wafer Heat Transfer Model	40
3.3	Proper Orthogonal Decomposition	41
3.4	Balancing	45
3.5	Comparison and Remarks	47
3.6	Reduction from CFD Models	49
4	Conclusion	52
4.1	Future Directions	54
4.2	Summary	55
A	Physical Constants	56
B	Dependent Variables	56
C	Independent Parameters	57
D	Balancing For Linear Systems	57

1 Introduction

In December 1996, a project was initiated at the Institute for Systems Research (ISR), under an agreement between Northrop Grumman Electronic Sensors and Systems Division (ESSD) and the ISR, to investigate the epitaxial growth of silicon–germanium (Si–Ge) heterostructures in a commercial rapid thermal chemical vapor deposition (RTCVD) reactor. This report provides a detailed account of the objectives and results of work done on this project as of September 1997. Some of the contents of this report are contained in [34].

1.1 Objectives

Epitaxial growth of Si–Ge heterostructures on a silicon substrate is an area of great current interest [9, 15, 17, 32]. This is mainly due to the superior electrical performance and manufacturing economies associated with Si–Ge devices [9, 49].

One such epitaxial growth process of immediate interest to Northrop Grumman ESSD is Si–Ge co-deposition on a patterned silicon substrate. The deposition is performed using a commercial RTCVD reactor manufactured by Advanced Semiconductor Materials (ASM), Inc. and located at Northrop Grumman ESSD. The patterned substrate in this case is a silicon wafer upon which has been deposited a geometrically patterned layer of silicon oxide. One important property of the oxide pattern geometry is *pattern pitch*, which is the distance between peaks in the oxide layer pattern. Pattern pitch can be thought of as a measure of how densely packed the pattern geometry is. Its importance stems from the depletion phenomenon associated with tightly packed geometries which affects the uniformity of the deposited film.

The overall objective of this project is to improve the manufacturing effectiveness of the epitaxial growth process and the RTCVD reactor by

- improved understanding of the processes and equipment via physical and mathematical modeling, and
- using the resulting validated models for optimization of process conditions.

More specifically, the project seeks to improve product quality as measured by deposition uniformity. Currently, the process engineer operating the ASM Epsilon–1 reactor can achieve an acceptable level of deposition uniformity (1% variation) on bare and patterned wafers, but the uniformity control task is undertaken on mainly a trial–and–error basis. Furthermore, the limits to achieving deposition uniformity in the presence of microfeatures, or oxide patterns, on the wafer surface are not understood. An improved understanding of the processes and equipment will provide a method to improve or optimize process settings and conditions.

Thus, a main goal is to develop physics–based models for thermal, fluid, and chemical mechanisms in order to predict deposition characteristics under different process settings and conditions such as pressure, flow rate, and temperature. The modeling effort takes into account lamp characteristics, reactor geometry, chemical reaction kinetics, and optical and thermal properties of the materials involved.

Due to the complexity and high computational demands of these models, we are investigating the use of model reduction techniques to reduce the model complexity, leading to faster simulation and facilitating the use of standard control and optimization strategies. An important objective is to develop new model reduction techniques for a class of nonlinear systems that are exemplified by the models described in this report.

1.2 Participants

The participants in this project are as follows: the ISR team consists of Andrew Newman and Prof. P. S. Krishnaprasad, and the Northrop Grumman team consists of Sam Ponczak, Michael O’Loughlin, and Paul Brabant. The ISR participants gratefully acknowledge the assistance of the Northrop Grumman participants and the use of their equipment for experimental purposes.

1.3 Equipment

The equipment used for thin film deposition in this project is the Epsilon–1 Reactor System manufactured by ASM Inc., Phoenix, AZ. One of these reactor systems is currently used as a production tool by Northrop

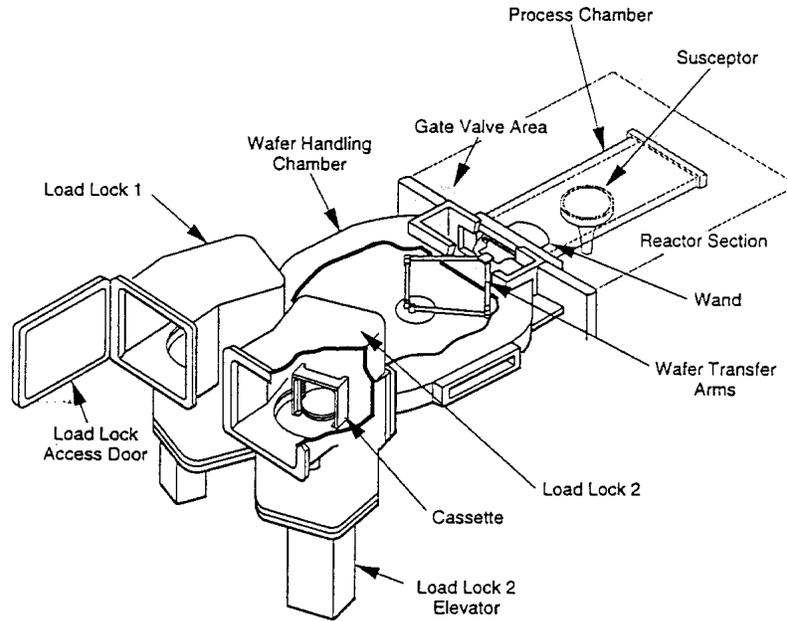


Figure 1: ASM Epsilon-1 RTCVD reactor wafer handling components. Process chamber is physical area of interest for deposition process. (Source: ASM Epsilon-1 Reactor Manual.)

Grumman ESSD, Linthicum, MD. The equipment and process models described in this report are specifically tailored for this particular reactor, although the physical principles on which they are based hold for similar reactors and mechanisms. Also, all model validation and other experiments performed for this project have taken place and will take place in the future using this reactor.

The ASM Epsilon-1 is a single-wafer, lamp-heated, RTCVD reactor. Figure 1 shows a schematic of the wafer handling components. The process chamber is the physical area of interest for the deposition process. For equipment details see [5].

Figures 2 and 3 show cross-sectional views of the reactor section and the lamp assembly. Process gases flow horizontally through the process chamber from inlet slits to the exhaust. The wafer is heated by an upper and lower array of linear tungsten-halogen lamps as well as four spot lamps directed at the center of the wafer. The upper lamp array illuminates the top side of the wafer while the lower array and the spot lamps illuminate the bottom side of the susceptor. Four thermocouples at the center, side, front (upstream), and rear (downstream) of the susceptor measure temperature. Thermocouple readings are used in proportional-integral-derivative (PID) control loops for temperature control. The reactor can operate at both atmospheric and reduced pressure (e.g., 20 Torr). The graphite susceptor rotates, typically at 35 rpm. Radiation from the wafer edge is reduced by a guard ring. The chamber walls are quartz and are air cooled.

Northrop Grumman provided two tools for measuring film thickness in our experiments, nanospec and ellipsometer. The Nanometrics 210 XP Scanning UV Nanospec/DUV Microspectrophotometer, or nanospec, is an instrument for measuring the thickness of optically transparent thin (10 to 4000 nm) films on silicon wafers. The speed of the nanospec makes it an attractive tool for taking a large number of measurements. A thickness measurement is recorded within 2 seconds using the nanospec, compared with approximately 15 seconds for the ellipsometer. However, the nanospec does not measure polysilicon thickness that is less than 100 Angstroms. For films less than 100 Angstroms, we used the ellipsometer.

1.4 Project Evolution

RTCVD of Si-Ge heterostructures on a patterned wafer is a complicated process requiring models and analyses that focus on different aspects of the problem. Therefore, the project has evolved gradually by first considering the least complex aspects and tasks and then building upon this foundation to attack the more complicated problems.

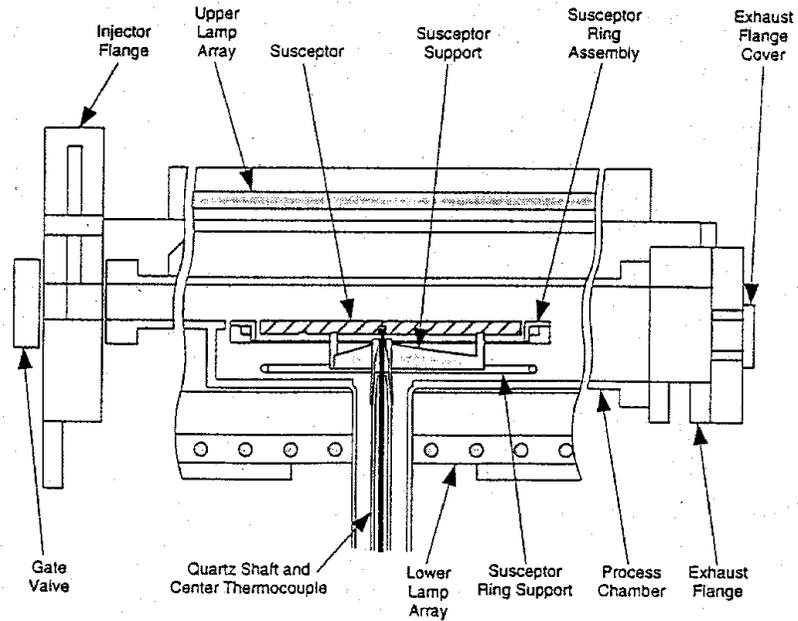


Figure 2: Cross-sectional view of reactor section of ASM Epsilon-1 including process chamber. Gases flow from inlet injector (left) to exhaust (right). (Source: ASM Epsilon-1 Reactor Manual.)

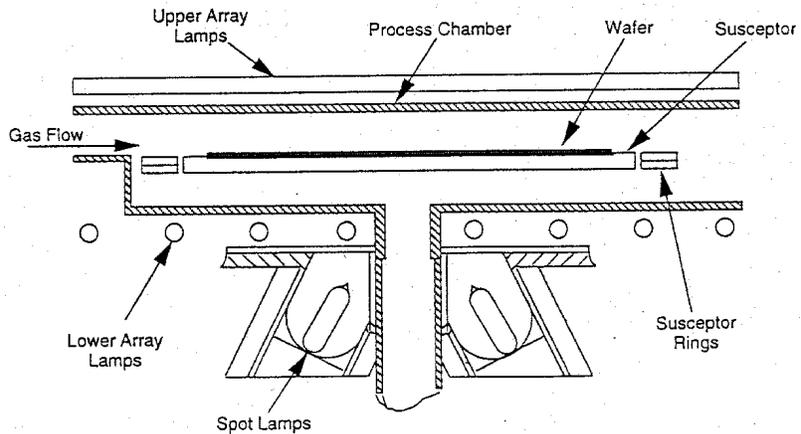


Figure 3: Lamp assembly of ASM Epsilon-1. Upper array illuminates top surface of wafer. Bottom array and spot lamps illuminate bottom surface of susceptor. (Source: ASM Epsilon-1 Reactor Manual.)

Initial efforts in this project focused on thermally activated epitaxial growth of silicon on bare silicon wafers. Simulations were accomplished by hard coding stand-alone programs using an appropriate programming language and by programming commercially available general purpose and specialized computational fluid dynamics (CFD) software packages. Both of these software implementation approaches numerically integrate high-order ordinary differential equation (ODE) approximations of partial differential equation (PDE) models for the relevant balance equations. Models that incorporate more complicated mechanisms associated with Si-Ge co-deposition, patterned wafers, and deposition in the mass transport controlled regime are under investigation. Different software implementations of models are being integrated to work together effectively. Model reduction techniques are being used to find low-order ODE models that provide good approximations to the original high-order ODE and PDE models.

The following illustrates a summary of the project evolution.

Si Epitaxy	→	Si-Ge Epitaxy
Bare Wafers	→	Patterned Wafers
Thermally Activated	→	Mass Transport Controlled
Stand Alone, Hard Coded Simulations	→	Integrated With CFD
PDE and High Order ODE Models	→	Low Order ODE Models

1.5 Report Outline

This report is organized into sections as follows. Section 2 describes in detail the modeling effort including the process and equipment models we have developed and experimental work for validation and determination of growth parameters. Section 3 discusses our work in model reduction as it is applied to the models used in this project. Section 4 outlines future directions in optimization and control, and presents some concluding remarks.

2 Physics Based Models

As stated in Section 1, one goal of our research is to model and understand the phenomena that influence deposition uniformity in the presence of microfeatures or patterns on the silicon substrate. The first task we undertake in achieving this goal is the development of physics-based and other types of models for processes and equipment. These models will provide an improved understanding of the process and equipment dynamics and a foundation for subsequent efforts in simulation, optimization, and control.

Development of sophisticated models will progress in stages of increasing generality, fidelity, and complexity. Ultimately, the models must retain high fidelity for a wide range of operating conditions for temperature, pressure, species concentration, and flow rate. In addition, it is hoped that they will accurately predict deposition characteristics for Si-Ge thin films on patterned wafers. Initially, however, we make simplifying assumptions, restrict models to only the most important processes, and consider less complicated deposition tasks. For example, we begin by considering deposition of thin epitaxial layers of silicon on a bare silicon substrate under the assumption that kinetics are strongly surface-reaction controlled. This modeling effort provides a foundation for a series of minor and major enhancements.

2.1 Modeling Overview

The general modeling scheme is illustrated in Figure 4 (see [26] for a similar approach). The two main components of the modeling effort are a macroscopic level process and equipment model, and a microscopic level feature scale model. These two components cooperate to achieve the overall modeling goals, but separately they accomplish qualitatively different tasks. Therefore, they have different inputs, outputs,

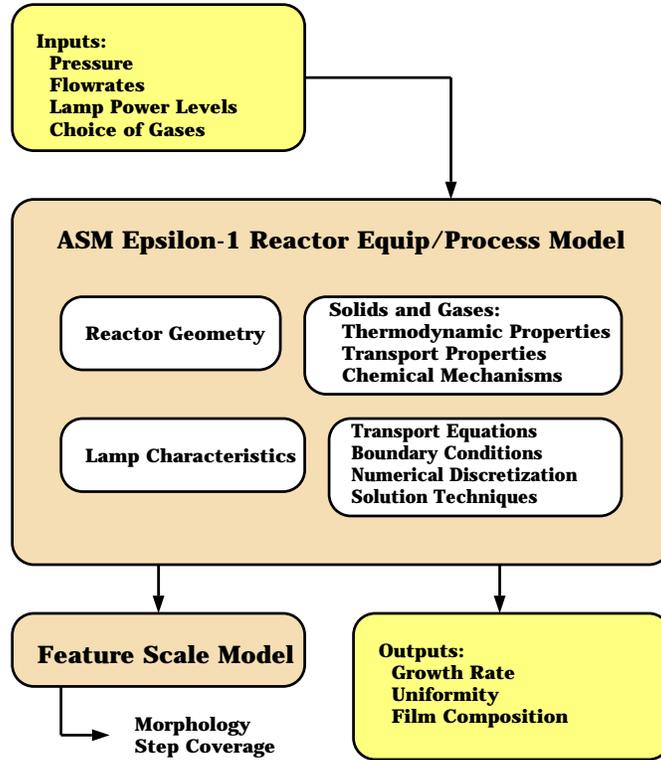


Figure 4: Overview of modeling effort for epitaxial growth in ASM Epsilon-1 reactor.

implementation strategies, and underlying physics. Cooperation is accomplished by using certain outputs of the macroscopic level model as inputs to the microscopic level model.

The process and equipment model describes phenomena that occur at a macroscopic level. This refers to continuum transport of energy, mass, and momentum in the solids and gases that comprise the reactor and process materials, and chemical mechanisms viewed in the aggregate sense. The motion or deposition of individual atoms is not considered.

The feature scale model describes phenomena that occur at a microscopic level. We are interested in deposition of atoms and formation of alloys on a patterned substrate. Therefore, the microscopic model predicts atom motion on the wafer surface, and in the boundary layer just above the surface, under given process conditions, as provided by the macroscopic level model.

2.2 Modeling Issues

Epitaxial growth of Si-Ge thin films in the ASM Epsilon-1 reactor is a complicated process due to several factors. The deposition kinetics are strongly dependent on the material to be deposited, source gases, operating conditions, reactor design, and wafer properties. Reactants are transported to the wafer via gas flows which may exhibit swirling or turbulent behavior under certain conditions. Heat energy is exchanged among lamps, wafer, chamber walls, and flowing gases in a complex manner. Properties of the materials involved depend on process variables that are changing, often in complicated ways. We state here some complicating issues which may be important to consider for development of high-fidelity models, and some simplifying assumptions we can reasonably adopt for dealing with them.

2.2.1 Complications

1. The wafer rests on a rotating susceptor. This rotation gives rise to gas flow vorticity and dynamics in the axial direction. This phenomenon together with asymmetries in the reactor geometry may cause

azimuthal symmetry of flow over the wafer surface to disappear. Thus, a full three-dimensional model may be needed to predict the gas flow field.

2. The ASM Epsilon-1 reactor can operate at both atmospheric pressure (AP) and reduced pressure (RP). Choice of pressure mode will affect deposition kinetics. Furthermore, flow can be turbulent in the AP regime [51], although it has been observed to always be laminar in the ASM reactor [29].
3. Transport of reactant species to the wafer surface occurs via multi-component diffusion and convection. Hence, complicated gas flows and boundary layer effects enter into the deposition kinetics equations [51]. Depletion effects (reduction in reactant species concentration at the downstream end of the flow) may also occur.
4. Deposition rates are affected by germanium content in the reactant gases. This phenomenon is the subject of much recent research [22, 23, 25, 31], which provides empirical data and rules of thumb but no explicit mathematical models. Effects of multiple reactant gases complicate the deposition kinetics models.
5. Reaction kinetics are sensitive to patterns (microfeatures) on the wafer surface. For example, it has been shown experimentally that thickness dependence on an oxide pattern is a strong function of pressure [21] and deposition rates and germanium fraction are affected by pattern pitch and closeness to pattern edges [23].
6. The type of growth (epitaxial, polycrystalline, or amorphous) is sensitive to the operating conditions [20, 30, 39, 44, 51], most importantly temperature, pressure, flow rate, and type and concentration of reactant species. It is also strongly influenced by the properties of the substrate surface [29]. The literature provides experimental results indicating some combinations of conditions that result in epitaxial growth but little in the way of mathematical models.
7. Deposition can be either mass transport or surface reaction limited, depending on wafer surface temperature. The transition temperature from one regime to the next is a function of silicon precursor. In the ASM reactor, it has been observed that for silane, mass transport limited deposition occurs above 850 C, but the transition temperature is higher for dichlorosilane, trichlorosilane, and silicetetrachloride [29].
8. Growth rate parameters often need to be determined experimentally. Moreover, it is likely that the measured parameters will only be useful for deposition of one particular material or a limited range of operating conditions. Ideal Arrhenius laws must be modified to account for leveling at high temperature and for pressure dependence.
9. The wafer is heated with upper and lower lamp arrays which are divided into ten lamp groups and four lamp heating zones, each driven with separate actuation. Finding the relationship between a particular lamp group or zone power setting and radiant heat flux intensity on the wafer surface requires either an empirically determined heat flux intensity profiles (using experimental data and further analysis), a ray-trace algorithm, or a view factor model (see, e.g., [16, 24]). Any analytical approach needs to be verified experimentally.
10. Heat transfer to and from the wafer includes conductive, convective, and radiative transport mechanisms. Heat energy is exchanged among lamps, wafer, chamber walls, reflectors, and flowing gas in a complex manner. High-fidelity models require incorporation of chamber wall geometry and material properties for computation of radiative terms due to reflections and nonuniformity in ambient temperature [50].
11. Material properties are functions of the process variables, whose time evolution in turn depend on the material properties. For example, heat transfer in the wafer depends on the mass density, heat capacity, thermal conductivity, and emissivity of the wafer. Meanwhile, mass density and thermal conductivity are nonlinear functions of the wafer temperature, and emissivity is a nonlinear function of deposition thickness. Imperfections in the wafer may cause spatial variation in properties. Furthermore, the wafer itself has a finite thickness so that there may exist heat transfer in the axial direction.

12. Thermocouple sensors are not in contact with the wafer, but instead are located at four locations on the susceptor ring and susceptor center. If temperature measurements are used in a model, this must be taken into account.

2.2.2 Simplifications

Homoepitaxy of silicon thin films on a silicon substrate is a less complicated and better understood process than heteroepitaxy of Si–Ge thin films. Therefore, the initial modeling effort focuses on homoepitaxy of silicon on a silicon substrate which provides a foundation for a progression of more accurate and complex models. Thus, initially we eliminate the need to consider effects of germanium and its source gases on the deposition kinetics. We choose silane as the source gas for homoepitaxy of silicon since this provides well known and simple chemical reactions, with reduction occurring at lower temperatures than for the other precursor gases such as silicon tetrachloride [51].

A portion of the research described in this report focuses on thermally activated growth, i.e., deposition for which reaction kinetics are strongly surface–reaction controlled. This assumption will not be valid when high operating temperatures are used. However, Northrop Grumman ESSD performs some low temperature epitaxial growth of undoped silicon at temperatures between 600 C and 800 C [38]. In that portion of our work where this assumption is invoked, we can essentially ignore the gas flow dynamics which control convective transport of species to the wafer surface. Instead, we can assume a fixed concentration profile of reactant species at the wafer surface as a parameter in the kinetics equations.

Furthermore, ignoring gas flow dynamics also has the effect that wafer rotation is no longer a complicating factor. In fact, wafer rotation helps to simplify matters by eliminating any azimuthal asymmetry in radiant heat transfer from the lamp banks to the wafer. Thus, the wafer thermal dynamics can be formulated solely in the radial direction. Moreover, since a susceptor ring is used we can realistically ignore wafer edge effects, i.e., lamp heat flux incident upon, or additional radiative losses from, the side of the wafer edge.

Initial models will be for a perfectly flat, perfectly cylindrical, homogeneous, unpatterned, silicon wafer of negligible thickness. Effects of microfeatures will be considered later. Physical constants such as mass density, heat capacity, thermal conductivity, and emissivity are initially assumed constant, i.e., no variation with temperature, film thickness, position, or time.

We employ a view factor model for lamp heating. Radiative heat energy exchange within the chamber is limited to only that between the wafer surface and a uniform ambient, in this case the chamber walls. Reflections off chamber walls and temperature nonuniformity throughout the chamber and walls is not included.

The physical mechanisms of other actuators and sensors such as mass flow controllers and thermocouples are not modeled. Instead it is simply assumed that they function as intended without error or corruption.

2.3 Process and Equipment Models

The macroscopic level process and equipment model predicts time evolution for

- heat transfer within and among the wafer, chamber walls, and process gases (temperature fields in solids and gases),
- momentum transport in the gas phase (gas flow velocity vectors),
- mass transport in the gas phase (species concentrations), and
- chemical reaction kinetics in the gas phase and at the wafer surface (reaction rates, deposition thickness).

Thus, it consists mainly of balance equations for conservation of energy (e.g., heat equation), and momentum and mass (e.g., Navier–Stokes), along with equations that describe the relevant chemical mechanisms (e.g., Arrhenius laws). Boundary conditions, source terms, required parameters, and any other details of the model are determined using available data regarding the reactor geometry, lamp characteristics, properties of the gases and solids involved, and chemical mechanisms. In the course of this project some of this data has been experimentally measured by the participants, some has been analytically determined via mathematical

modeling, and some has been gathered from appropriate references and databases. Each particular source of data will be indicated in this report in the appropriate section.

Inputs to the model consist of pressure, inlet flow rates, inlet injector slit widths, choice of gases, and lamp power settings. For thermally activated growth (surface reaction controlled), the heat lamps provide the control actuation via radiative heat transfer to the wafer. For mass transport controlled growth, control actuation is achieved by adjusting boundary conditions at the inlet.

The measured outputs of the model include aggregate growth rate, spatial uniformity, and film composition. These are the variables which we ultimately wish to control. We also assume, when convenient, access to other variables, such as the temperature at given points on the wafer surface, which is used as an input to the feature scale model. These variables may or may not be available as measured outputs. If *in-situ* measurement is not possible then for real-time control a method of estimating them would be required.

2.3.1 Implementation

As stated in the previous section, the macroscopic level process and equipment model consists of evolution equations and associated boundary conditions for transport of heat, mass, and momentum along with gas phase and surface chemical reaction kinetics. We take two approaches to implementation. One approach is to hard-code individual stand-alone software modules in a suitable programming language such as MATLAB¹ to simulate individual components of the model. These individual components can then be interfaced to simulate larger portions of the model or the overall model. The other approach is to use commercially available general purpose computational fluid dynamics (CFD) software packages with capability for simulation of complex fluid flow, heat transfer, and chemical reaction systems. In addition, supplemental code can be developed for integration with the CFD software for modeling phenomena specific to the equipment and processes of interest. We shall see that both (hard-coding and CFD) approaches are useful and necessary, and that the overall implementation requires some degree of redundancy between the two.

There currently exist commercially available general purpose CFD software packages that provide the required capabilities for implementing the model including

- body fitted grids for discretization of complicated physical domains such as those found in a CVD reactor,
- efficient algorithms for numerically integrating the fully coupled nonlinear versions of the transport equations, and
- Monte-Carlo ray-tracing and view factor algorithms for modeling radiative heat transfer among surfaces.

We are using two of these general purpose CFD packages, Fluent² and PHOENICS³, to provide high-fidelity simulations of the fluid flow, heat transfer, species transport, and chemical mechanisms in the ASM Epsilon-1 reactor. One attractive feature of PHOENICS is a specialized add-on program called PHOENICS-CVD for modeling CVD processes which includes specialized databases for optical and thermal properties of materials and gas phase and surface chemistry of many commonly used reactions. It is not our purpose here to provide a complete review of the capabilities of these software packages. We will indicate when and how they are used in this modeling effort as the need arises in this report.

The CFD approach does not do everything we need. Not only are general purpose CFD programs restricted to modeling at the macroscopic level, but the output data is not particularly suitable for integration into a model at the microscopic level. For patterned wafers, we are studying mechanisms at the feature scale. We need models which focus attention on what is happening at the wafer surface and which provide very high spatial resolution there. General purpose CFD simulations with a spatial resolution (discretization) suitably fine for feature scale modeling would be computationally prohibitive.

Furthermore, implementation of feedback control and optimization techniques with the reactor model as the plant would be cumbersome and impractical using the general purpose CFD model. Model reduction

¹The Mathworks, Inc., Natick, MA

²Fluent, Inc., Lebanon, NH

³CHAM Ltd., London, UK

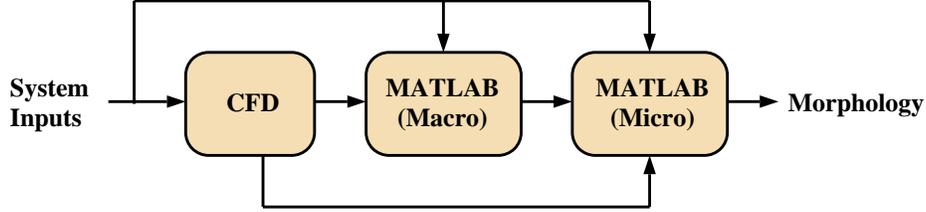


Figure 5: Hard-coded individual MATLAB programs are integrated with general purpose and specialized CFD models to produce the desired overall input-output model of epitaxial growth the ASM Epsilon-1 reactor.

strategies for development of low-order models can be applied to the hard-coded stand-alone programs to provide useful approximations of the original models for use in optimization and control.

There is a necessary degree of redundancy between the two approaches. For example, heat transfer in the solid wafer is implemented both via general purpose CFD software and via stand-alone MATLAB programs. In the CFD approach the spatial discretization of the solid wafer is relatively coarse but suitable for simulating the overall heat transfer in the reactor. In the MATLAB implementation the spatial discretization can be set arbitrarily fine without seriously degrading computational performance, and the resulting code is immediately available for use in model reduction, optimization, and control schemes.

In summary, the CFD approach yields more realistic solutions, since the sophisticated software packages solve the fully coupled nonlinear versions of the transport equations. It would be impractical for the authors to attempt to replicate all of these advanced features into our stand-alone hard-coded programs. Instead, the hard-coded programs invoke many simplifying assumptions to make their implementation practical. Both approaches are integrated to accomplish the project objectives. Figure 5 illustrates a strategy for integrating the two approaches to build a complete input-output model for describing morphology evolution for epitaxial growth in the ASM Epsilon-1 reactor.

2.3.2 General Purpose and Specialized CFD

As stated earlier, we are using two general purpose CFD software packages, Fluent and PHOENICS, to model fluid flow, heat transfer, and chemical reactions in the ASM Epsilon-1 reactor. These packages incorporate up-to-date modeling techniques and a wide range of physical models. In addition, we are using a specialized add-on program, PHOENICS-CVD, which has features and databases of special use for CVD applications. We cannot provide a full discussion of the equations, algorithms, discretization techniques, and other characteristics of these software packages. See [11] for a complete description of Fluent and various articles in [48] for a detailed description of PHOENICS-CVD and several example applications.

Here we demonstrate the CFD modeling approach as applied to the ASM Epsilon-1 reactor by presenting some results from steady-state simulations of certain phenomena of interest. These simulations assumed the following constant operating conditions: wafer temperature 725 C, chamber wall temperature 425 C, inlet gas temperature 25 C, and inlet flow rate 1.5 slm of 2% silane in hydrogen. Effects due to surface reactions are included.

A 2-dimensional model of the ASM Epsilon-1 reactor was developed using a Cartesian geometry and nonuniform grid. This model represents a 2-dimensional slice of the process chamber parallel to the direction of flow, i.e., from inlet to outlet. Symmetry is assumed in the direction perpendicular to the plane of the slice.

Figure 6 shows the steady-state condition for gas flow velocity. The gas velocity magnitude is large near the inlet, with some recirculation vortices. The velocity magnitude is small as it passes the wafer, with the expected parabolic flow profile. Figure 7 shows the steady-state condition for gas temperature. The gas is cold as it leaves the inlet and is heated as it passes the wafer. Figure 8 shows the steady-state condition for silane mole fraction. Surface reactions cause a depletion effect as the concentration of silane decreases from upstream to downstream side of the wafer. Figure 9 shows the mass flux of deposited silicon. Deposition is a maximum at the leading edge of the wafer due to the depletion effect.

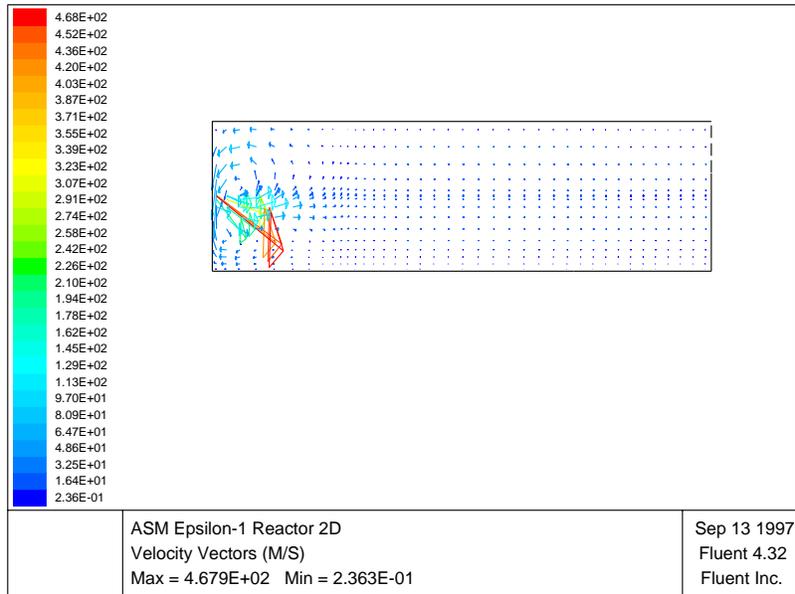


Figure 6: Steady-state gas velocity vectors (fluid flow) in 2-dimensional slice of ASM Epsilon-1 process chamber.

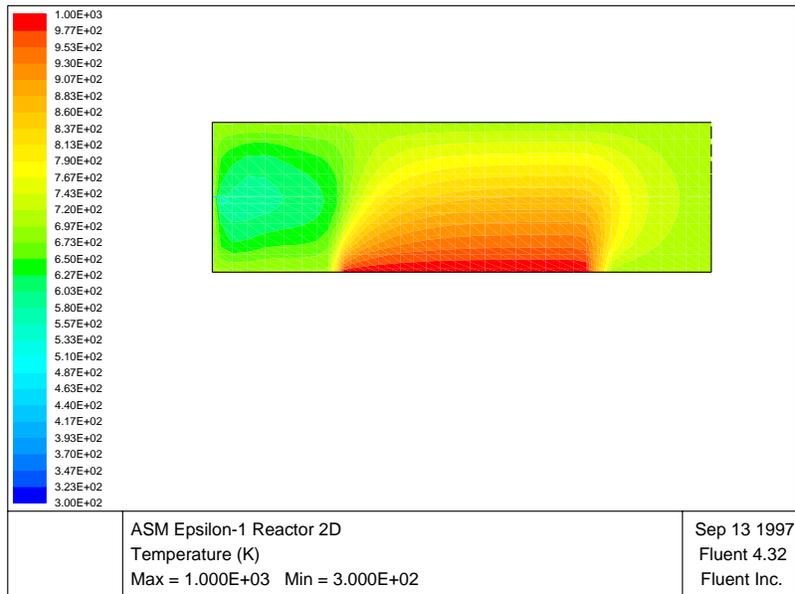


Figure 7: Steady-state gas temperature in 2-dimensional slice of ASM Epsilon-1 process chamber.

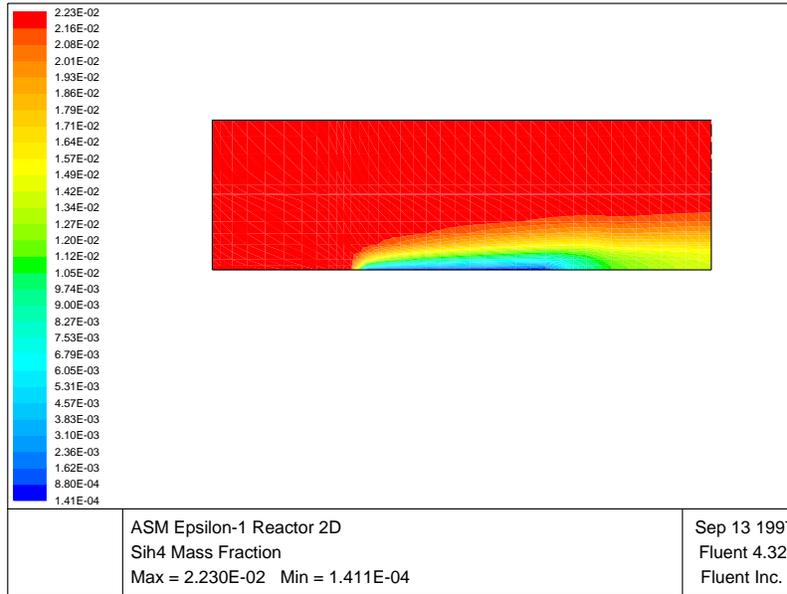


Figure 8: Steady-state silane mole fraction in 2-dimensional slice of ASM Epsilon-1 process chamber.

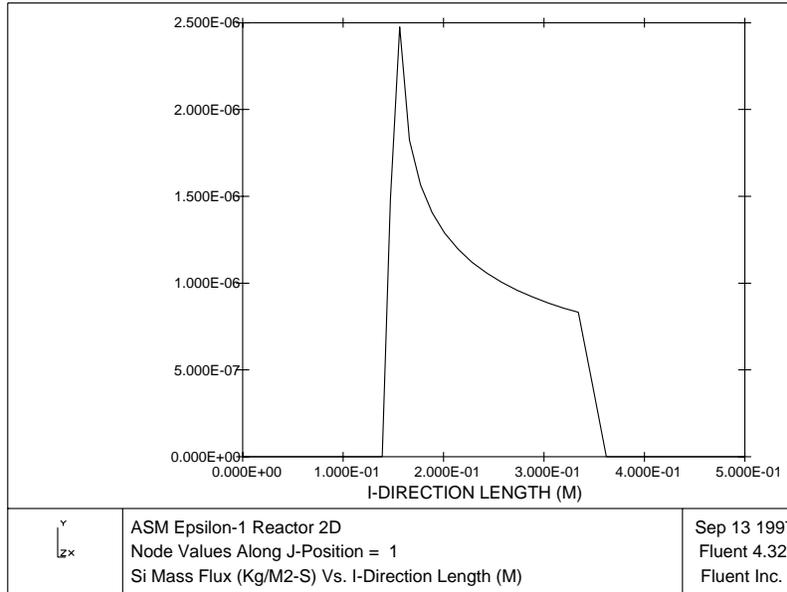


Figure 9: Plot of deposited silicon mass flux versus wafer radial position for ASM Epsilon-1.

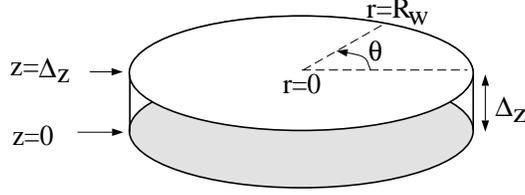


Figure 10: Wafer with coordinates and measurements used in the heat transfer model.

2.3.3 Wafer Heat Transfer

The model for wafer heat transfer is a modified version of models presented in [2, 8, 28, 41, 42]. It is based on an energy balance for a heat conducting solid which emits and absorbs heat radiation at its boundary surfaces. The model takes into account simplified effects of conductive, radiative (including lamp heating), and convective heat transfer. Both a continuum model and a discretized version are presented here, based on identical principles of energy balance. Although the wafer is a continuous solid body, a discretized model is required for purposes of numerical solution.

Since the wafer shape is assumed to be a perfect cylinder, the model is formulated in cylindrical coordinates with radial variable r , azimuthal variable θ , and axial variable z . The wafer radius is denoted R_w and the wafer thickness is denoted Δ_z , so that the top surface of the wafer has z -coordinate Δ_z and the bottom surface has z -coordinate 0. Figure 10 shows a diagram of the wafer with coordinates. Note that for purposes of the initial analysis, the wafer and susceptor have been combined into a single homogeneous solid body.

Continuum Model

The temperature field in the solid wafer is denoted $T_w = T_w(t, r, \theta, z)$ where t is the time variable. Time evolution of T_w is described by a partial differential equation (PDE) (usually referred to as the *heat equation*) which models heat conduction within the wafer, together with boundary conditions (BCs) which model net heat flow to and from the wafer boundary surfaces (top, bottom, and edge). The PDE is given in cylindrical coordinates by

$$\rho_w C_{p_w} \frac{\partial T_w}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left(k_w r \frac{\partial T_w}{\partial r} \right) + \frac{1}{r^2} \frac{\partial}{\partial \theta} \left(k_w \frac{\partial T_w}{\partial \theta} \right) + \frac{\partial}{\partial z} \left(k_w \frac{\partial T_w}{\partial z} \right) \quad (1)$$

for $t > 0$, $0 < r < R_w$, $0 \leq \theta < 2\pi$, and $0 < z < \Delta_z$ where ρ_w is the mass density of the wafer, C_{p_w} is the heat capacity of the wafer (the product $M_w = \rho_w C_{p_w}$ is often referred to as the wafer thermal mass), and k_w is the thermal conductivity of the wafer. The associated BCs are given by

$$\frac{\partial T_w}{\partial r} = 0; \quad r = 0 \quad (2)$$

$$k_w \frac{\partial T_w}{\partial r} = q_{edge}(\theta, z); \quad r = R_w \quad (3)$$

$$k_w \frac{\partial T_w}{\partial z} = -q_{bottom}(r, \theta); \quad z = 0 \quad (4)$$

$$k_w \frac{\partial T_w}{\partial z} = q_{top}(r, \theta); \quad z = \Delta_z \quad (5)$$

where the first BC results from symmetry about the wafer center, and q_{edge} , q_{bottom} , and q_{top} represent the net heat flow per unit surface area to and from the wafer edge, bottom, and top boundary surfaces, respectively, and will be described in more detail later.

For purposes of modeling film growth, we focus our attention on the top surface of the wafer where reactions take place. Invoking the assumption of azimuthal symmetry, so that no temperature gradients exist in the azimuthal direction (i.e., $\partial T_w / \partial \theta = 0$), time evolution of T_w at the wafer top surface is described by

$$\rho_w C_{p_w} \frac{\partial T_w}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left(k_w r \frac{\partial T_w}{\partial r} \right) + \frac{\partial}{\partial z} \left(k_w \frac{\partial T_w}{\partial z} \right) \quad (6)$$

for $t > 0$, $0 < r < R_w$, and $z = \Delta_z$ with BCs remaining the same as above except $q_{edge} = q_{edge}(z)$, $q_{bottom} = q_{bottom}(r)$, and $q_{top} = q_{top}(r)$. We also assume that the wafer thickness is sufficiently small so that no thermal gradients exist in the axial direction within the wafer interior. Therefore, we approximate the axial gradient term at the top surface by

$$\begin{aligned} \frac{\partial}{\partial z} \left(k_w \frac{\partial T_w}{\partial z} \right) &\simeq \frac{1}{\Delta_z} \left(k_w \frac{\partial T_w}{\partial z} \Big|_{z=\Delta_z} - k_w \frac{\partial T_w}{\partial z} \Big|_{z=0} \right) \\ &= \frac{1}{\Delta_z} (q_{top} + q_{bottom}) \end{aligned}$$

where we have made substitutions using the appropriate BCs. The resulting PDE describes the evolution of the wafer top surface temperature field as a function of time and radial position,

$$\rho_w C_{pw} \frac{\partial T_w}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left(k_w r \frac{\partial T_w}{\partial r} \right) + \frac{1}{\Delta_z} (q_{top} + q_{bottom}) \quad (7)$$

with BCs

$$\frac{\partial T_w}{\partial r} = 0; \quad r = 0 \quad (8)$$

$$k_w \frac{\partial T_w}{\partial r} = q_{edge}(\Delta_z); \quad r = R_w \quad (9)$$

$$\cdot \quad (10)$$

Now, we must find expressions for q_{top} and q_{bottom} , the net heat flow into the top and bottom surfaces of the wafer. For the initial analysis, we assume that the top and bottom surfaces are subject to identical heat transfer mechanisms, and let

$$q_{top} + q_{bottom} = q^{em} + q^{ab} + q^{conv} + q^{dist} \quad (11)$$

where the terms on the right hand side represent the flow of thermal energy to and from the wafer and are dependent on time, position, and wafer temperature. In particular, q^{em} is radiative energy emitted, q^{ab} is radiative energy absorbed, q^{conv} denotes energy losses due to convective heat transfer, and q^{dist} is energy transfer due to unmodeled effects such as heat generated by chemical reactions. In what follows we ignore q^{dist} .

The term q^{em} represents radiative losses from the wafer. We assume q^{ab} depends on radiant heat flux from a uniform ambient, in this case the chamber walls, and radiant heat flux from the lamps, but without reflections or other effects. The individual terms are given by

$$q^{em} = -2\epsilon_w \sigma_b T_w^4 \quad (12)$$

$$q^{ab} = 2\alpha_w \sigma_b T_c^4 + \alpha_w \sum_{i=1}^{10} Q_i u_i \quad (13)$$

where σ_b denotes the Boltzmann constant, ϵ_w denotes the wafer emissivity, α_w denotes the wafer absorptivity, T_c denotes the uniform ambient temperature of the chamber walls, $Q_i = Q_i(r)$ is a function of position describing the heat flux intensity incident on the wafer due to the i -th lamp group, and $u_i = u_i(t)$ is the time-varying actuated power level of the i -th lamp group.

The convective term is given by

$$q^{conv} = -h_v (T_w - T_g) \quad (14)$$

where h_v denotes the convective heat transfer coefficient and T_g denotes the temperature of the gas flowing past the wafer. Note that we have assumed a constant uniform gas temperature. In order to estimate h_v , we assume flow in the process chamber is a laminar flow along a flat plate. The mean heat transfer coefficient is given in [37] pp. 233–235 as

$$h_v = 2 [0.332 k_g Pr^{1/3} (Re^{1/2}/L)] \quad (15)$$

where k_g denotes the gas thermal conductivity, Pr denotes the gas Prandtl number, Re denotes the gas Reynolds number, and L denotes the length of the chamber. We have computed the Reynolds number Re to

be approximately 27 for the flow in the ASM Epsilon-1 during a typical deposition run, thus confirming the laminar assumption. The calculated value of h_v was then validated using flow and temperature data from a corresponding CFD simulation. See Appendix A for values of all of these physical constants.

It is sometimes convenient to assume that the wafer is a graybody, so that $\epsilon_w = \alpha_w$ for all relevant wavelengths of radiation and wafer temperatures. However, we do not make this assumption here, and use different values for emissivity and absorptivity. We also note that the parameters ρ_w , C_{p_w} , and k_w can be modeled as nonlinear functions of T_w , and the parameters ϵ_w and α_w can be modeled as nonlinear functions of T_w and deposition thickness. However, we invoke the assumption that mass density, heat capacity, thermal conductivity, emissivity, and absorptivity are constant, i.e., no variation with temperature, film thickness, position, or time. The PDE model specializes to

$$\begin{aligned} \frac{\partial T_w}{\partial t} = & \frac{k_w}{\rho_w C_{p_w}} \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T_w}{\partial r} \right) + \frac{h_v}{\rho_w C_{p_w} \Delta_z} (T_g - T_w) + \frac{2 \sigma_b \epsilon_w}{\rho_w C_{p_w} \Delta_z} (T_c^4 - T_w^4) \\ & + \frac{\alpha_w}{\rho_w C_{p_w} \Delta_z} \sum_{i=1}^{10} Q_i u_i \end{aligned} \quad (16)$$

where we recall that $T_w = T_w(t, r)$, $Q_i = Q_i(r)$, and $u_i = u_i(t)$.

Since the guard ring insulates the wafer from radiation directed at its edge boundary surface, we assume zero heat transfer at the wafer edge so that

$$q_{edge} = 0$$

giving the boundary conditions (BCs)

$$\frac{\partial T_w}{\partial r} = 0; \quad r = 0 \quad (17)$$

$$\frac{\partial T_w}{\partial r} = 0; \quad r = R_w. \quad (18)$$

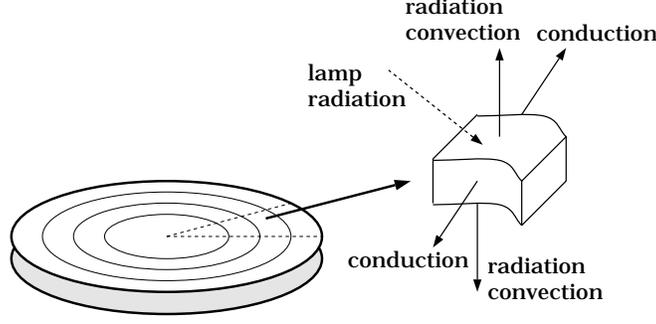


Figure 11: Wafer discretized into annular regions. Heat transfer in wafer described by energy balance for each discrete element.

Discretized Model

The continuum model as given by the above PDE and BCs can be discretized using a suitable scheme, e.g., finite differences or finite elements. However, for our simplified model it is easier to formulate a discretization by applying the energy balance principles directly to individual annular elements of the wafer. The general idea, which divides the wafer into annular regions, is illustrated in Figure 11. Annular regions will be numbered from 1 to n with element 1 being the innermost disk and element n being the outermost annular region. Element i has mean radius $r(i)$, and is bounded by an outer cylinder of radius r_{out} , inner cylinder of radius r_{in} , top surface at $z = \Delta_z$, and bottom surface at $z = 0$. The discretization is uniform so that

$$\Delta_r = r(i) - r(j)$$

is constant for all i, j .

The usual symmetry assumptions are invoked so that temperature is dependent upon radial position and time only. The discretized wafer temperature field is given by the n -vector $T_w(t)$, where the i -th entry of $T_w(t)$ represents the temperature at radial position $r(i)$ and time t .

The wafer heat transfer model is then given by the ODE

$$\dot{T}_w = A_c T_w + A_r T_w^4 + A_v T_w + \Gamma + B P \quad (19)$$

where A_c , A_r , and A_v are $n \times n$ matrices representing the effects of conductive, radiative, and convective heat transfer mechanisms, respectively, Γ is a constant n -vector which accounts for the gas and chamber wall temperature, B is a $n \times m$ matrix of discretized lamp zone radiant intensity profiles, and $P = P(t)$ is a m -vector of control inputs corresponding to lamp zone power levels. For the Epsilon-1 reactor, there are $m = 4$ independently actuated lamp zone control inputs. We present the details of the ODE model below.

The top surface area, volume, and mass of annular region i are given, respectively, by

$$\begin{aligned} S(i) &= \pi (r_{out}(i)^2 - r_{in}(i)^2) \\ V(i) &= S(i) \Delta_z \\ m(i) &= \rho_w V(i) \end{aligned} \quad (20)$$

The matrix representing conductive heat transfer is then represented by the tridiagonal matrix given by the entries

$$\begin{aligned} A_c(i, i) &= \frac{-2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{r_{out}(i) + r_{in}(i)}{r_{out}(i)^2 - r_{in}(i)^2} \\ A_c(i, i+1) &= \frac{2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{r_{out}(i)}{r_{out}(i)^2 - r_{in}(i)^2} \\ A_c(i, i-1) &= \frac{2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{r_{in}(i)}{r_{out}(i)^2 - r_{in}(i)^2} \end{aligned} \quad (21)$$

for $i = 2, \dots, n - 1$ and

$$\begin{aligned}
A_c(1, 1) &= \frac{-2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{1}{r_{out}(1)} \\
A_c(1, 2) &= \frac{2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{1}{r_{out}(1)} \\
A_c(n, n) &= \frac{-2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{r_{in}(n)}{r_{out}(n)^2 - r_{in}(n)^2} \\
A_c(n, n - 1) &= \frac{2 k_w}{\rho_w C_{p_w} \Delta_r} \frac{r_{in}(n)}{r_{out}(n)^2 - r_{in}(n)^2}
\end{aligned} \tag{22}$$

where we note that zero heat flux BCs have been incorporated into the model via boundary elements of matrix A_c .

The matrices representing radiative transfer from wafer surface to chamber walls and convective heat transfer from the process gases to wafer are given, respectively, by

$$A_r = \text{diag}\left[\frac{-\sigma_b \epsilon_w}{\rho_w C_{p_w} \Delta_z}\right] \tag{23}$$

and

$$A_v = \text{diag}\left[\frac{-h_v}{\rho_w C_{p_w} \Delta_z}\right] \tag{24}$$

where we note that A_r and A_v take diagonal form as a result of the simplifications made in our model.

The effect of radiation from chamber wall to wafer and convective transfer from gas to wafer are incorporated into the constant vectors Γ_r and Γ_v whose entries for $i = 1 \dots n$ are given by

$$\Gamma_r(i) = \frac{\epsilon_c \sigma_b \alpha_w}{\rho_w C_{p_w} \Delta_z} T_c^4 \tag{25}$$

$$\Gamma_v(i) = \frac{h_v}{\rho_w C_{p_w} \Delta_z} T_g \tag{26}$$

where ϵ_c is the emissivity of the quartz chamber walls. These effects are combined by summing into one constant vector

$$\Gamma = \Gamma_r + \Gamma_v. \tag{27}$$

Discretized lamp heat flux intensity profiles as given in Section 2.3.4 are arranged in a matrix Q and incorporated into the influence matrix B given by

$$B = \frac{\alpha_w}{\rho_w C_{p_w} \Delta_z} Q. \tag{28}$$

To avoid problems of scaling in computational work, we normalize variables and parameters so that all units cancel, i.e. write the model in dimensionless form. It is customary to adopt a notation for the dimensionless variables, e.g. T_w becomes \tilde{T}_w . Instead, we denote the dimensionless variables by the same symbol as their dimensional counterparts and caution the reader to keep this in mind. The conversions are

$$T_w \rightarrow \frac{T_w}{T_c}, \quad Q_i \rightarrow \frac{Q_i}{Q_{ref}}, \quad t \rightarrow \frac{t}{\tau}, \quad r \rightarrow \frac{r}{R_w}.$$

and are also given in Appendices B and C.

It has been observed in the ASM reactor that T_c , the chamber wall temperature, is approximately 300 K less than wafer temperature [29] during a typical processing run. As reference values we select a wafer temperature of 1000 K and chamber wall temperature of 700 K. The reference thickness h_{ref} of 1.0 micron was selected because it is on the order of the thickness of films we are interested in growing. The reference heat flux Q_{ref} of 29.24 W/cm² was computed using the lamp power specification of 6 kW radiating over one-half of a spherical surface area of radius 2.25 inches. All values are given in Appendix A.

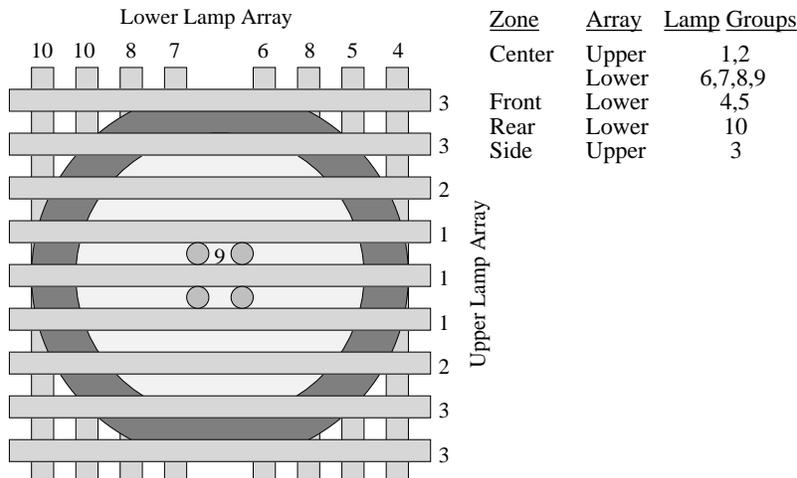


Figure 12: Organization (top view) of upper and lower lamp arrays and spot lamps: individual lamps are assigned to lamp groups and heat zones as shown.

Using the dimensionless variables as given above, the parameters (matrices and vectors) in equation (19) become

$$A_c \rightarrow \frac{\tau}{R_w^2} A_c, \quad A_r \rightarrow \tau T_c^3 A_r, \quad A_v \rightarrow \tau A_v, \quad \Gamma_r \rightarrow \frac{\tau}{T_c} \Gamma_r, \quad \Gamma_v \rightarrow \frac{\tau}{T_c} \Gamma_v, \quad B \rightarrow \frac{\tau}{T_c} B$$

to yield an ODE model equivalent to (19).

The ODE (19) can be numerically integrated to approximate the PDE (16), given appropriate initial conditions, $T_{w_0} = T_w(0)$, to determine the temperature field on the wafer surface as a function of time and radial position. Typically, the initial condition is a constant temperature field set to ambient, i.e., $T_{w_0} = [700 \dots 700]^T$. For the work described in this report we used a fourth and fifth order Runge–Kutta integration scheme to perform the numerical integrations. The discretization resolution was typically set at $n = 101$.

2.3.4 Lamp Heating

Heat transfer in the RTCVD reactor is dominated by radiative effects such as radiative transfer between wafer and chamber and radiative transfer from lamps to wafer. The importance of a lamp heating model, as one component of the overall heat transfer model, is apparent, especially for describing a thermally activated deposition process.

Preliminaries

Figure 12 illustrates the layout and organization of the heat lamps in the ASM Epsilon–1 reactor. Upper and lower arrays are perpendicular to each other, with spot lamps in the center. Individual lamps are combined into groups, which are further combined into zones. It is the heat zones that are controlled independently in the ASM Epsilon–1. Thus, the Epsilon–1 has four control inputs for heating. The zone name roughly corresponds to the area of the wafer that receives the most intense illumination from the particular zone, e.g. center, front (upstream), rear (downstream), and side. The number of lamps in each zone varies as indicated.

Lamp heating appears in the dynamic model for wafer heat transfer as the control input term, and is described by a set of influence functions which represent the heating effect, or incident heat flux intensity, of each lamp heat zone on the wafer surface. Each influence function is multiplied by a corresponding dimensionless scalar control input which represents the proportion of full power applied to the particular lamp zone actuator. Each control input is typically a function of time and/or sensor measurements, e.g. thermocouple temperature readings. In the heat transfer model (19) the lamp influence functions are appropriately

discretized and arranged into a matrix, denoted B . Similarly, the control inputs are arranged into a vector of time-varying functions, denoted P .

The influence functions physically correspond to lamp heat flux intensity spatial profiles. A heat flux intensity spatial profile is a function of position on the wafer surface whose value at a spatial point is the heat flux intensity (measured in Watts per unit area) irradiated on the given point. In the case where we have azimuthal symmetry, such as when the wafer rotates at a uniform rate, azimuthal variations are averaged over 360 degrees, so that each profile is a function of radial position only. Such a profile is determined for each individual lamp in the ASM Epsilon-1 reactor. Then, individual profiles are combined using superposition into profiles for the ten lamp groups and four lamp zones of the Epsilon-1.

The flux intensity profiles are determined analytically and verified experimentally. One possible experimental approach would be to infer flux intensity profiles from temperature measurement data. For example, an instrumented wafer with nine attached thermocouples is available at Northrop Grumman to provide temperature data at nine points on the wafer surface. However, based on past experience, there is some doubt that the instrumented wafer can provide data of sufficient accuracy for this purpose, especially when used under typical flow conditions in the reactor. An alternative method for temperature measurement is inferring temperature from deposition thickness data. We used this method for validation purposes and it is described later. The main problem with using experimental temperature data to compute lamp heat flux intensity profiles is achieving the desired spatial resolution. With the instrumented wafer, we get readings at only nine spatial points. Using the deposition thickness approach, a very large number of thickness measurements must be taken, and it is difficult or impossible to determine the exact spatial location of the measurement on the wafer. Therefore, the analytically determined profiles are used in the model, since they can be computed using an arbitrarily fine discretization at exactly those points we choose. Experimental results are used only for comparative validation.

In what follows we describe our methodology for determining the profiles, present some preliminary results, and discuss improvements which are necessary for the lamp heating models to accurately predict the effect of lamp heating in the ASM Epsilon-1 reactor for use in high fidelity models of epitaxial growth.

Methodology

The analytical approach we take to determine the heat flux spatial profiles is based on the concept of view factor [37, 45] which describes the radiation exchange between two or more surfaces separated by a nonparticipating medium that does not absorb, emit, or scatter radiation. The view factor between two surfaces represents the fraction of radiative energy leaving one surface that strikes the other surface directly.

In this method, the geometry of the chamber, including location and shape of lamps, susceptor, reflectors, and possibly other apparatus, is what mainly determines the form of the resulting flux profiles. This geometric approach was adopted in [16], where the authors consider only a two-dimensional slice of the chamber geometry, and includes the effect of reflectors behind the lamp banks. There, the two-dimensional approach was reasonable, perhaps, since the lamp arrangement in the reactor under consideration was axisymmetric about the wafer center. This situation is, however, not the case in the ASM Epsilon-1 reactor. Hence, our analysis is similar to that used in [10], where the authors consider the chamber geometry from a three-dimensional point of view. However, in that paper, as in this paper, the effect of reflectors is not included.

Scope

In this report, radiant flux profiles for both linear and spot lamps are determined. In the actual reactor, the internal surface of the chamber lid is gold plated to reflect infrared rays from the linear lamps, and the spot lamps are placed in gold plated parabolic reflectors. However, in this report, we do not consider the effect of reflections on the lamp heating of the wafer.

Finally, the literature indicates that “virtual images”, or radiation from the heated wafer to the reflectors and chamber walls which is reflected back to the wafer, will cause additional radiative effects. These effects are not included in the analysis here.

Assumptions

We shall consider all surfaces to be diffuse reflectors and diffuse emitters. Radiant intensity from lamps is assumed to be independent of direction and constant across the length of the lamp. We assume that the quartz walls and the process gases transmit heat radiation from the lamps perfectly at the wavelengths of

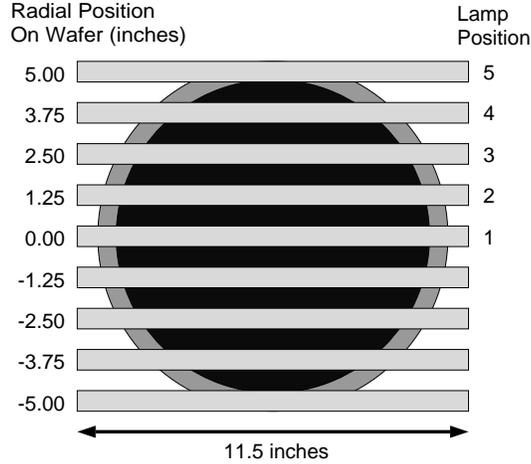


Figure 13: Geometry (top view) of upper lamp array: radial position of each lamp is given in inches from center; lamps are identified by five uniquely distinguishable positions, numbered 1 through 5.

interest. Furthermore, we assume that the path from lamps to wafer (or lamps to susceptor) is completely free of any other obstacles.

Relevant Geometry

Figure 13 shows a schematic of the upper lamp array superimposed over the susceptor and wafer, which is based on a description and diagram provided in [5]. For computational purposes, we consider each linear lamp to be a straight line segment of length 11.5 inches with the array consisting of parallel equally spaced lamps. The array begins directly above the susceptor edge, 5.0 inches (horizontally) from the susceptor center. The distance between neighboring parallel lamps in the array is 1.25 inches. The vertical distance between wafer and upper lamp array is 2.25 inches, and the vertical distance between wafer and lower lamp array is 3.50 inches. We note that the distances given are estimates based on crude measurements taken on the reactor itself.

There are five lamp positions for the linear lamps that can be uniquely differentiated from the others. This is due to the wafer rotation. For example, two linear lamps equally distant from the center linear lamp have an identical irradiating effect on the wafer surface. The five lamp positions are numbered 1 through 5. The spot lamps have their own unique geometry and are analyzed separately later.

The source of radiation for each lamp is a tungsten filament, which we assume to be a straight line segment stretching the length of the lamp. Figure 14 shows the geometry used to perform the analysis. We assume that for each filament the radiant intensity is independent of direction and constant across the length of the filament.

For each point on the wafer, w , there is an irradiance contribution from each point on the filament, f , depending upon the distance between them, $d = \|w - f\|$, the angle θ_w formed by the vector $w - f$ and the vector n_w normal to the wafer surface, and the vertical distance, h , from wafer surface to filament. Note that on the filament diagram the endpoint values are $f_1 = -5.75$ inches and $f_2 = 5.75$ inches, and the vertical distance h is either 2.25 inches for the upper array or 3.50 inches for the lower array.

Heat Flux Calculation

For the derivation of the expression for heat flux radiant power per unit area on the wafer surface, we adopt the notation used in [37]. The rate of radiative energy dQ_f leaving a differential surface area dA_f (containing the point f) on the filament that strikes a differential surface area dA_w (containing the point w) on the wafer surface is given by

$$dQ_f = dA_f I_f \cos(\theta_f) d\omega_{fw} \quad (29)$$

where I_f is the intensity of radiative energy leaving dA_f in all directions in hemispherical space (in dimensions of Watts per unit area per steradian), θ_f is the angle formed by the vector $w - f$ and a vector normal to

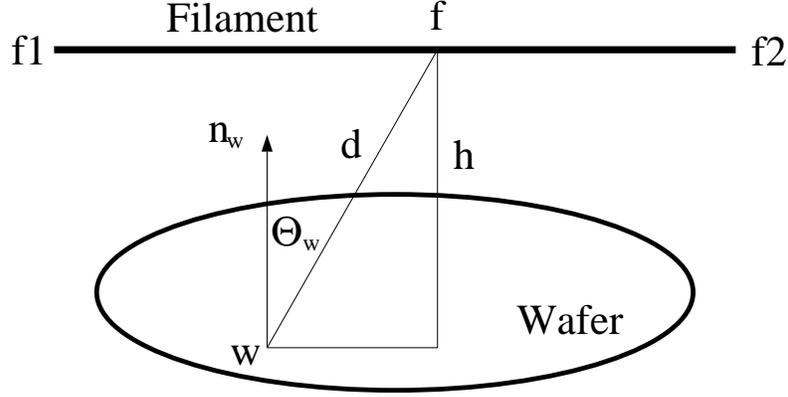


Figure 14: Geometry for view factor analysis used to calculate heat flux intensity profiles for linear lamps.

dA_f , and $d\omega_{fw}$ is the solid angle subtended by dA_w from f given by

$$d\omega_{fw} = \frac{dA_w \cos(\theta_w)}{d^2}. \quad (30)$$

Substituting (30) into (29) yields

$$dQ_f = dA_f I_f \frac{\cos(\theta_f) \cos(\theta_w) dA_w}{d^2}. \quad (31)$$

Now, the rate of radiation energy Q_f leaving the surface element dA_f on the filament in all directions over hemispherical space is [37]

$$Q_f = \pi I_f dA_f. \quad (32)$$

The elemental view factor $dF_{dA_f-dA_w}$ is defined as the ratio of the radiative energy leaving dA_f that strikes dA_w directly to the radiative energy leaving dA_f in all directions into the hemispherical space. Thus, we divide (31) by (32) to give the view factor

$$dF_{dA_f-dA_w} = \frac{dQ_f}{Q_f} = \frac{\cos(\theta_f) \cos(\theta_w) dA_w}{\pi d^2}. \quad (33)$$

Since we are assuming that filament radiant intensity is independent of direction, we take $\theta_f = 0$ independent of filament position f so that $\cos(\theta_f) = 1$ and

$$dF_{dA_f-dA_w} = \frac{\cos(\theta_w) dA_w}{\pi d^2}. \quad (34)$$

We are interested in the radiative energy illuminating a differential area on the wafer due to the entire filament. To compute the appropriate view factor, F_{f-dA_w} , we average (34) across the length of the filament

$$F_{f-dA_w} = \frac{dA_w}{|f_1 - f_2|} \int_{f_1}^{f_2} \frac{\cos(\theta_w)}{\pi d^2} df. \quad (35)$$

Finally, we observe that

$$\cos(\theta_w) = \frac{h}{d}$$

for the given geometry, so that

$$F_{f-dA_w} = \frac{dA_w}{|f_1 - f_2|} \int_{f_1}^{f_2} \frac{h}{\pi d^3} df \quad (36)$$

where we recall that $d = \|w - f\|$.

To determine the radiant heat flux profile for a given lamp, the view factor F_{f-dA_w} must be computed for each differential area dA_w on the wafer surface. For practical purposes, we discretize the wafer surface by choosing a cylindrical grid of wafer points $w = (r, \phi)$. We then assume that the differential area, dA_w , is constant for all wafer points w . Thus, (36) yields the view factor function $F_{f-dA_w}(r, \phi)$ which gives the fraction of radiative energy leaving the given lamp filament that strikes the given wafer point $w = (r, \phi)$ directly.

Now, we let P_f denote the radiant power supplied by the filament, so that P_f/dA_w gives the radiant heat flux intensity striking the differential area dA_w . The radiant heat flux intensity profile of the illumination due to the lamp filament is then given by

$$q_f(r, \phi) = F_{f-dA_w}(r, \phi) \frac{P_f}{dA_w} \quad (37)$$

$$= \frac{P_f h}{\pi |f_1 - f_2|} \int_{f_1}^{f_2} \frac{1}{d(r, \phi)^3} df \quad (38)$$

where the value we use for P_f is provided by the manufacturer. In the case of the ASM Epsilon-1 reactor, the linear lamps supply 6000 Watts and the spot lamps supply 1000 Watts.

Since the wafer rotates at a uniform rate, this function is averaged over the circle (i.e., $0 \leq \phi < 2\pi$) at each radial position r on the wafer top surface

$$q_f(r) = \frac{P_f}{dA_w} \frac{1}{2\pi} \int_0^{2\pi} F_{fw}(r, \phi) d\phi \quad (39)$$

to give the heat flux profile

$$q_f(r) = \frac{P_f h}{2\pi^2 |f_1 - f_2|} \int_0^{2\pi} \int_{f_1}^{f_2} \frac{1}{d(r, \phi)^3} df d\phi. \quad (40)$$

A similar analysis was performed for the spot lamps, except that each spot lamp was considered to be a point source of radiant energy, thus simplifying the analysis significantly.

The computational procedure was performed for each of the five different linear lamp positions for both upper and lower arrays, and for the spot lamps. Using the resulting heat flux intensity spatial profiles, we can then compute the desired profiles for each of the ten lamp groups, and the four heat zones of the Epsilon-1 reactor by appropriately combining the profiles determined from the individual lamps.

Results

Here we discuss some results of the analysis. Note that in what follows, the term “wafer surface” may represent the top surface of the wafer and exposed susceptor, or the bottom surface of the susceptor, depending upon the lamp group being considered.

Individual Lamps

Figures 15, 16, and 17 show the heat flux irradiated on the wafer surface by lamps in positions 1, 2, 3, and 4 of the upper and lower array, respectively, and the spot lamp position. As expected, points on the wafer surface directly under (or over) the lamp filament receive the most intense illumination, i.e. the maximum flux value. Intensities are greater in magnitude for lamps in the upper array since it is physically closer to the wafer than the lower array and spot lamps. Spot lamps have lower flux intensities than linear lamps due to the smaller supplied power.

To account for wafer rotation, the flux intensity profiles are averaged around 360 degrees resulting in profiles that are a function of radial position only. Figures 18 and 19 show the heat flux profiles, after averaging, for each of the individual lamp positions. Observe that as expected the lamp position directly over (or under) the wafer center irradiates the wafer center with greater intensity than the other positions. Lamp positions closer to the edge irradiate the edge with greater intensity than they irradiate the center.

Lamp Groups

Figure 20 shows the heat flux profiles for each of the ten lamp groups.

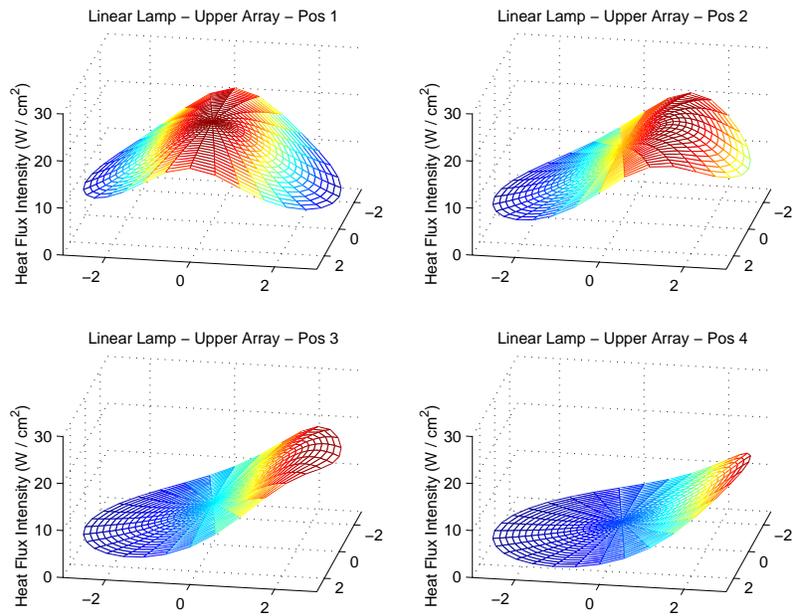


Figure 15: Heat flux intensity profiles for linear lamps in upper array: flux intensity (W/cm²) versus position in two dimensions. Upper left: Position 1; Upper right: Position 2; Lower Left: Position 3; Lower right: Position 4.

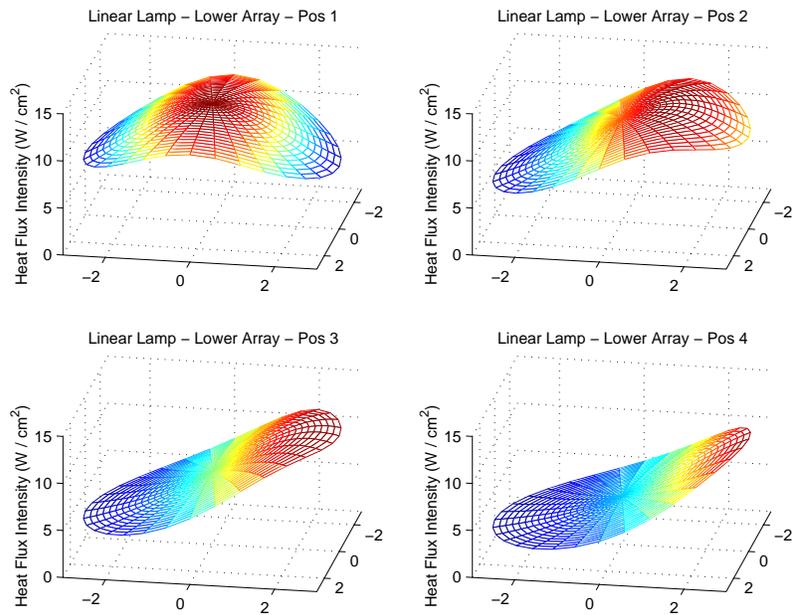


Figure 16: Heat flux intensity profiles for linear lamps in lower array: flux intensity (W/cm²) versus position in two dimensions. Upper left: Position 1; Upper right: Position 2; Lower Left: Position 3; Lower right: Position 4.

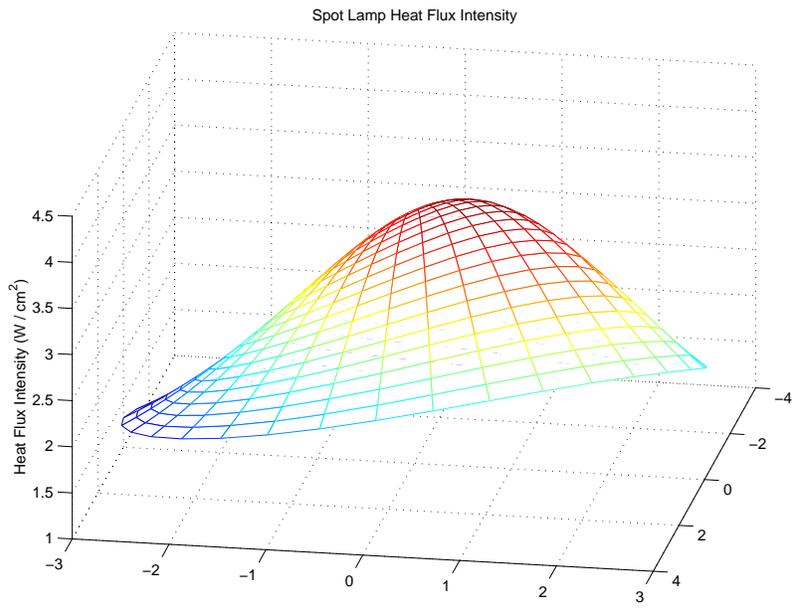


Figure 17: Heat flux intensity profile for spot lamp: flux intensity (W/cm²) versus position in two dimensions.

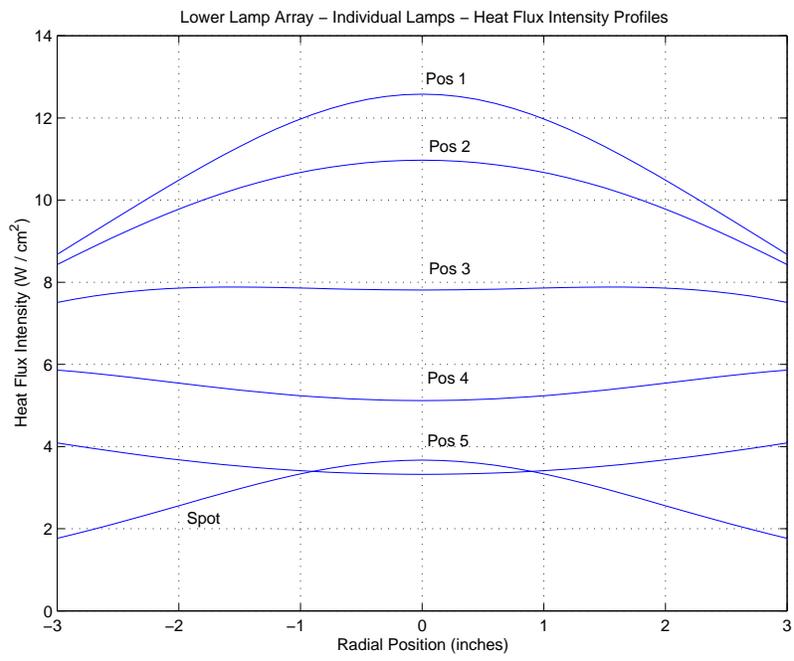


Figure 18: Heat flux intensity profiles for individual lamps in the lower array: flux intensity (W/cm²) versus radial position for the five uniquely distinguishable linear lamp positions and the spot lamp position.

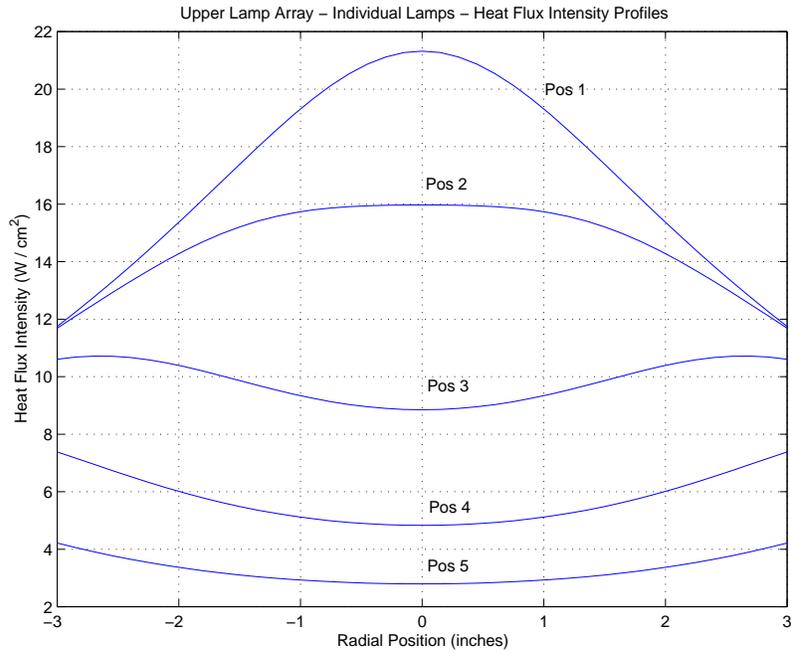


Figure 19: Heat flux intensity profiles for individual lamps in the upper array: flux intensity (W/cm^2) versus radial position for the five uniquely distinguishable linear lamp positions.

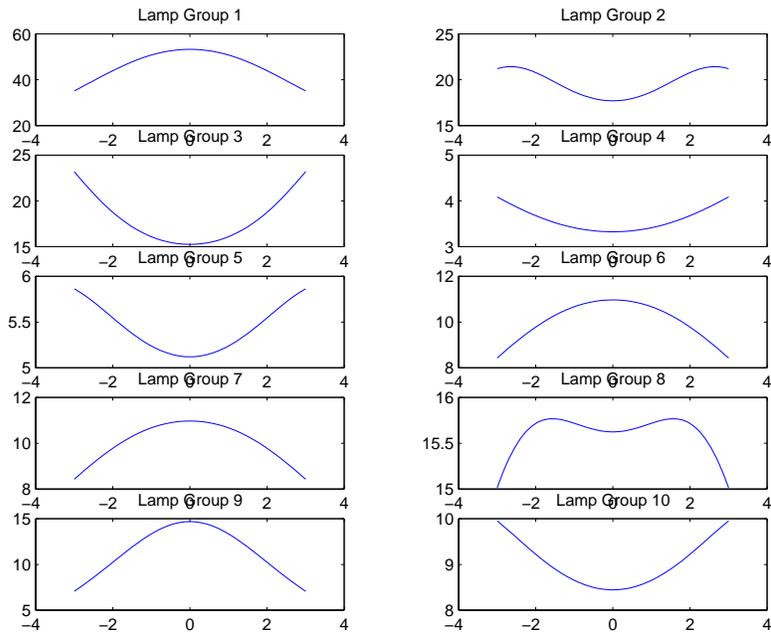


Figure 20: Heat flux intensity profiles for ASM Epsilon-1 lamp groups: flux intensity (W/cm^2) versus radial position for the ten lamp groups.

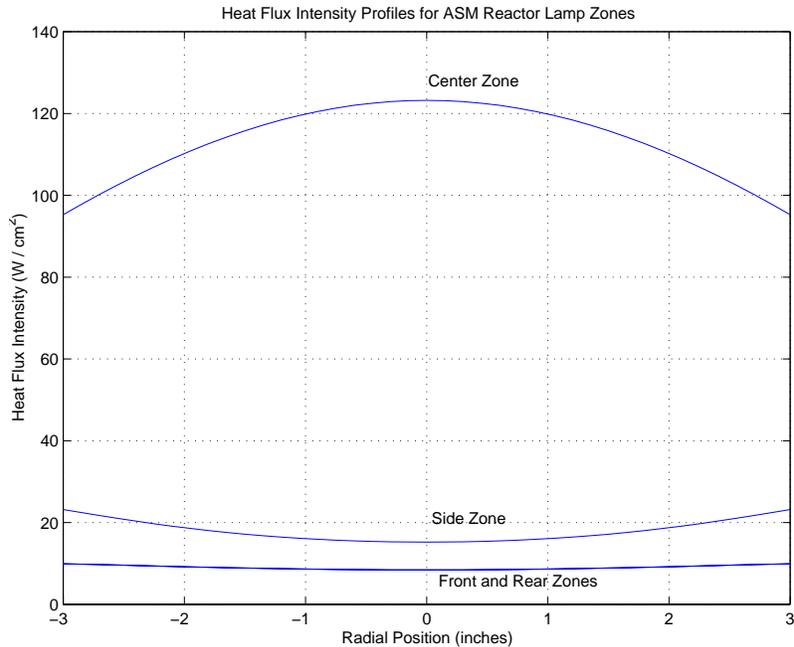


Figure 21: Heat flux intensity profiles for ASM Epsilon-1 heat zones: flux intensity (W/cm^2) versus radial position for the four heat zones.

Heat Zones

Figure 21 shows the heat flux profiles for the four heating zones - center, front, rear, and side. The flux intensity for the center zone is significantly greater than for the others, indicating that it will have the greatest heating effect. Observe that profiles for front and rear zones are identical due to the symmetry assumptions and the way in which the individual lamps are organized to form the zones.

Deficiencies and Improvements

There are several deficiencies in the lamp heating model as it currently stands. Improvements depend on gathering of more accurate chamber geometry data, determination of lamp and reflector properties, and inclusion of more complicated effects.

Reflectors

The internal surface of the chamber lid is gold plated to reflect infrared rays from the lamps. In addition, the spot lamps are placed in gold plated parabolic reflectors. The effects of reflections must be included in the model. In order to compute these effects, the properties and geometry of the reflectors must be known.

Accurate Geometry

The analysis has proceeded based on reactor manual diagrams which do not provide exact measurements for the chamber geometry. Accurate dimensions must be provided to improve the model.

Lamp Intensity

The lamp heat flux intensity profiles have been determined using the power ratings listed for the lamps in the reactor manual. These ratings are 6 kW for the linear lamps and 1 kW for the spot lamps. However, we do not know the efficiency of the lamps in terms of proportion of power consumed to power supplied for heating the wafer. Thus, the profiles may be off by a scale factor, the magnitude of which cannot be determined exactly without more information.

Physical Obstacles and Apparatus

Apparatus or other physical obstacles in the reactor may have an effect on the lamp heating by blocking radiation from certain directions. Currently these are not considered. If such obstacles exist, the analysis will be more complicated.

Virtual Images

The wafer itself will irradiate the reflector on the chamber lid, creating virtual images of the wafer which will reflect heat flux back onto the wafer surface. These effects may be significant.

2.3.5 Chemical Reaction Kinetics

The model we use for deposition kinetics is based on the Arrhenius relationship for reaction rate described in [20, 39, 42, 51]. We assume equilibrium of silane and hydrogen adsorption and desorption, and that the rate of silicon deposition is limited by the rate at which surface adsorbed silane decomposes to produce silicon and hydrogen in the reaction



The reaction rate at each position on the wafer surface is dependent on the mole fraction of silane, X_{SiH_4} , and the wafer temperature, T_w , at that position. It is given by

$$R_{\text{Si}}(T_w, X_{\text{SiH}_4}) = k_s(T_w) X_{\text{SiH}_4} \quad (42)$$

where $k_s(T_w)$ is the reaction rate parameter given by

$$k_s(T_w) = k_0 \exp\left(\frac{-E_a}{R_g T_w}\right) \quad (43)$$

and k_0 is the Arrhenius coefficient, E_a is the activation energy, and R_g is the gas constant.

Equivalently, we can replace the silane mole fraction dependence in (42) with dependence on silane concentration, C_{SiH_4} , or silane partial pressure, P_{SiH_4} :

$$R_{\text{Si}}(T_w, C_{\text{SiH}_4}) = k'_s(T_w) C_{\text{SiH}_4} \quad (44)$$

$$R_{\text{Si}}(T_w, P_{\text{SiH}_4}) = k''_s(T_w) P_{\text{SiH}_4} \quad (45)$$

where in each case the value and units of the Arrhenius coefficient k_0 in (43) is appropriately converted to account for the change of units. However, for our purposes, it is most convenient to use the rate expression (42) since the mole fraction X_{SiH_4} is dimensionless and hence all units can be accounted for in k_0 .

Values for the activation energy E_a and the Arrhenius coefficient k_0 depend on the process operating conditions, the source gases used, and the reactor itself. The molecular weight and mass density of the deposited material, silicon in this case, is also incorporated into k_0 . We experimentally determined the values for k_0 and E_a as will be described in Section 2.4.1. The values are given in Appendix A. Note that the value for E_a is in the range of 1.5–1.7 eV and the ratio E_a/R_g is in the range of 18000–20000, both of which lie within the range of values typically found for a variety of operating conditions (see, e.g., [39, 51]).

Reaction rate as a function of time and radial position on the wafer surface is given by

$$R_{\text{Si}}(t, r) = k_0 \exp\left(\frac{-E_a}{R_g T_w(t, r)}\right) X_{\text{SiH}_4}(t, r) \quad (46)$$

so that

$$\frac{\partial h}{\partial t}(t, r) = R_{\text{Si}}(t, r) = k_0 \exp\left(\frac{-E_a}{R_g T_w(t, r)}\right) X_{\text{SiH}_4}(t, r) \quad (47)$$

describes the time evolution of the deposition thickness profile $h(t, r)$. Equation (47) is numerically integrated in conjunction with equation (19), the evolution equation for the wafer temperature field $T_w(t, r)$, to determine the thickness profile of deposited silicon on the wafer surface. The mole fraction profile $X_{\text{SiH}_4}(t, r)$ can be determined from the models for gas flow and transport of chemical species in the reactor. We can also simplify the situation and assume a pre-determined constant silane mole fraction profile $X_{\text{SiH}_4}(r)$, e.g., a uniform value equal to the average silane mole fraction found during a particular deposition run.

When dimensionless variables are used to avoid scaling problems in computational work, we use the conversions

$$h \rightarrow \frac{h}{h_{ref}}, \quad t \rightarrow \frac{t}{\tau}, \quad r \rightarrow \frac{r}{R_w}$$

so that the time evolution of film thickness is given by

$$\dot{h} = \beta_r \exp\left(\frac{-\beta_e}{T_w}\right) X_{\text{SiH}_4}$$

where

$$\beta_r = \frac{\tau k_0}{h_{ref}}, \quad \beta_e = \frac{E_a}{R_g T_c}$$

are dimensionless constants, values of which are given in Appendix A.

2.3.6 Preliminary Results

Several simulations were run to demonstrate the predictive capabilities of the wafer heat transfer and chemical kinetics models. These simulations are tests to determine if the models yield physically reasonable results. In order to validate time evolution of the wafer temperature field, further experimentation will be necessary. For example, we could record the wafer temperature profile as a function of time using an array of thermocouples, and compare with simulation data. Time evolution of deposition thickness is validated by comparing with the experimental results presented in Section 2.4.1.

Wafer Temperature Ramp Up

In this simulation we start the wafer at a uniform room temperature and apply 25% power to all lamp zones for one minute. Time evolution of wafer temperature is shown in Figure 22. The wafer temperature increases nonuniformly across the surface due to the greater intensity of the center lamp zone.

During ramp-up, lamp heating dominates the heat transfer mechanisms until radiative losses become more apparent as the temperature increases beyond ambient chamber wall temperature. Effects of conduction do not appear since the temperature profiles are relatively uniform throughout and the intensity of radiative transfer is relatively high. The average temperature levels off at approximately 900 C.

Wafer Temperature Ramp Down

In this simulation we start the wafer at a uniform processing temperature of 725 C and turn off all lamp groups for 5 minutes. Time evolution of wafer temperature is shown in Figure 23. The wafer temperature drifts down to approximately 230 C due to radiative and convective losses from the wafer. Note that the temperature profile has a stable equilibrium, settling to a spatially uniform profile $T_w(t, r) \rightarrow T_{const}$ for all r as $t \rightarrow \infty$. Here, T_{const} is equal to the ambient chamber wall temperature T_c when convective losses are ignored, and lower than T_c when convective losses are included. Also, note that the temperature changes more rapidly than an exponential decrease due to the nonlinearity of the radiative effect. Heat diffusion in the solid has no effect in this simulation since the temperature remains uniform throughout.

Temperature Behavior In Response To Disturbances

This simulation demonstrates what happens when the temperature profile is locally perturbed from a uniform profile. Lamp heating is turned off in this simulation so that we can observe the effects of heat diffusion. We inject a temperature spike half-way between the wafer center and edge and record temperature profiles at intervals of 1.2 seconds over 12 seconds. The results are shown in Figure 24. Heat conduction in the solid wafer causes the expected smoothing effect. This behavior is consistent with what we expect to happen under our modeling assumptions, i.e., the wafer thermal dynamics are globally asymptotically stable about a uniform spatial profile slightly below chamber wall temperature (slightly below due to convective losses).

Film Growth Under Uniform Temperature Conditions

We demonstrate film growth under uniform wafer surface temperature conditions by simulating the film growth under uniform temperature profiles of 700 C, 725 C, 750 C, and 775 C over a 5 minute deposition

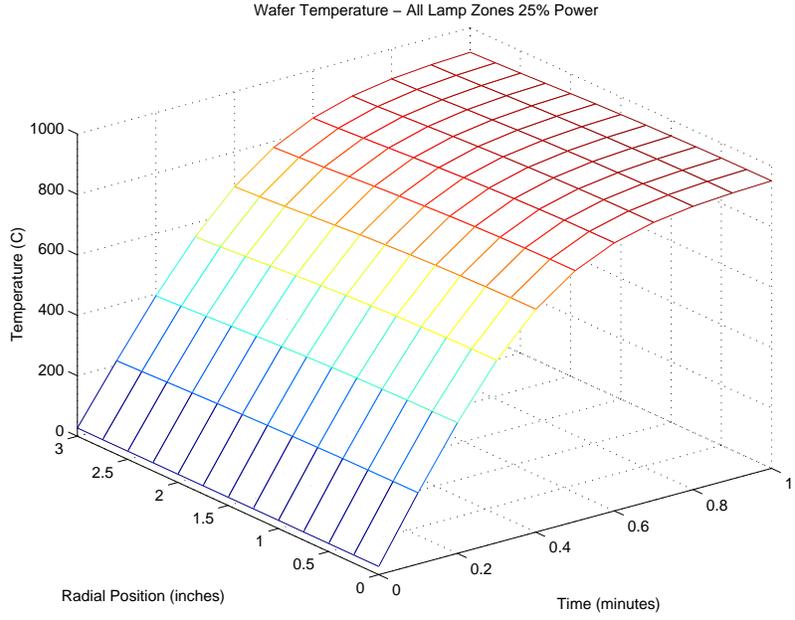


Figure 22: Time evolution of temperature profile across wafer surface: all lamp zones at 25% power over 1 minute.

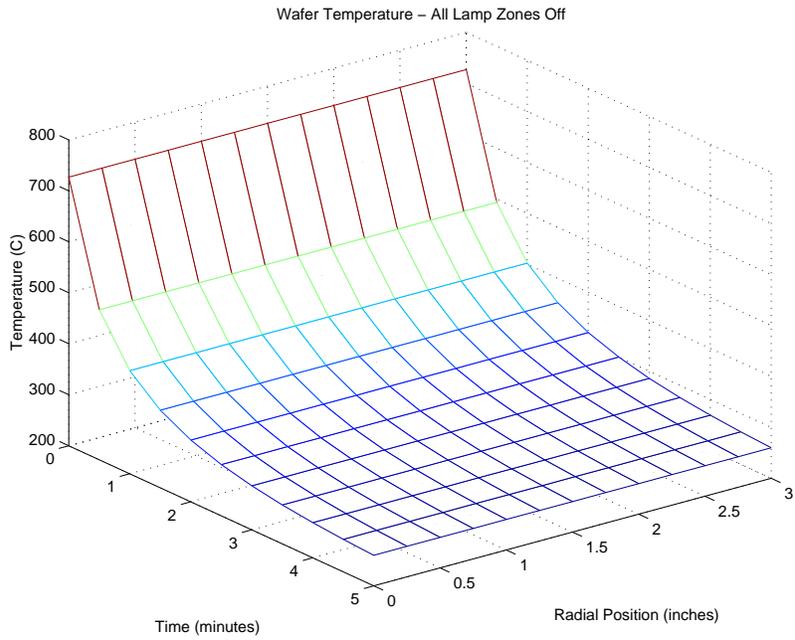


Figure 23: Time evolution of temperature profile across wafer surface: all lamp zones at 0% power over 5 minutes. Note: For visual clarity axes are reversed from that in Figure 22.

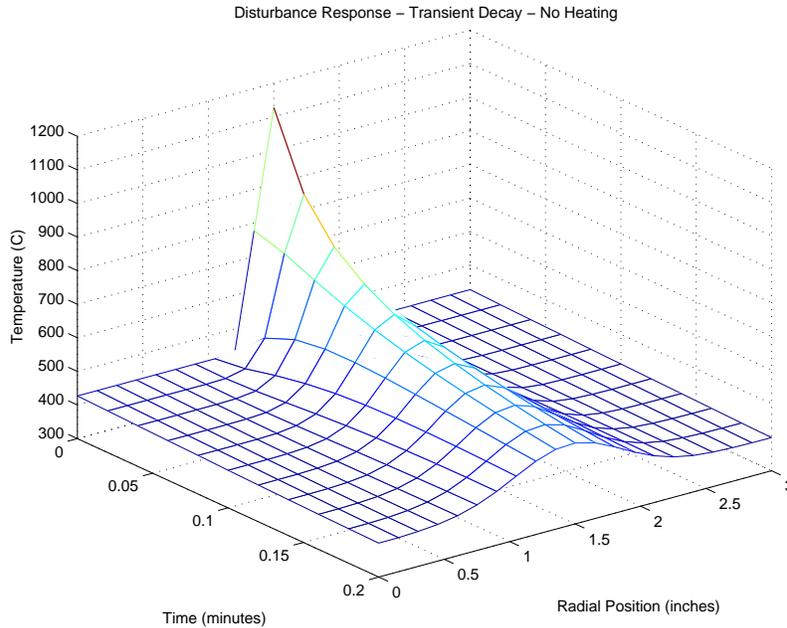


Figure 24: Response to temperature disturbance with all lamp groups at 0% power over 12 seconds.

period. Results are shown in Figure 25. The strong temperature dependence of the growth dynamics are apparent. These simulation results match the experimental results from growth experiments described in Section 2.4.1.

Film Growth Under Nonuniform Temperature Conditions

Here we simulate growth under a spatial temperature profile that linearly increases from 725.0 C at the center to 725.7 C at the edge. This 0.7 degree C difference amounts to a 1.2% variation in wafer temperature from center to edge. Growth is once again simulated over a 5 minute deposition period. Results are shown in Figure 26. The variation in temperature is amplified into a 19 Angstrom or 26% variation in thickness from center to edge at the end of the 5 minute period. This clearly demonstrates the importance of temperature control for reduced deposition nonuniformity.

2.4 Experimental Validation

Andrew Newman of ISR and Paul Brabant of Northrop Grumman ESSD conducted thin film deposition and lamp heating experiments using the ASM Epsilon-1 RTCVD reactor on site at Northrop Grumman ESSD from May 13, 1997 through May 15, 1997. Additional measurements were taken on July 27, 1997.

The objectives of these experiments were to

- validate an assumed Arrhenius relationship between wafer temperature and deposition rate, and determine the unknown parameters of this relationship under typical operating conditions;
- validate the correctness of analytically determined lamp heating models described earlier in this report;
- measure physical dimensions of various parts of the reactor in order to determine model parameters; and
- familiarize the ISR student participant with the operation and capabilities of the reactor and various other tools used at the Northrop Grumman ESSD facility.

The ASM Epsilon-1 reactor at Northrop Grumman ESSD is a production tool in regular use. The authors are grateful for the time and resources that were made available for experimentation.

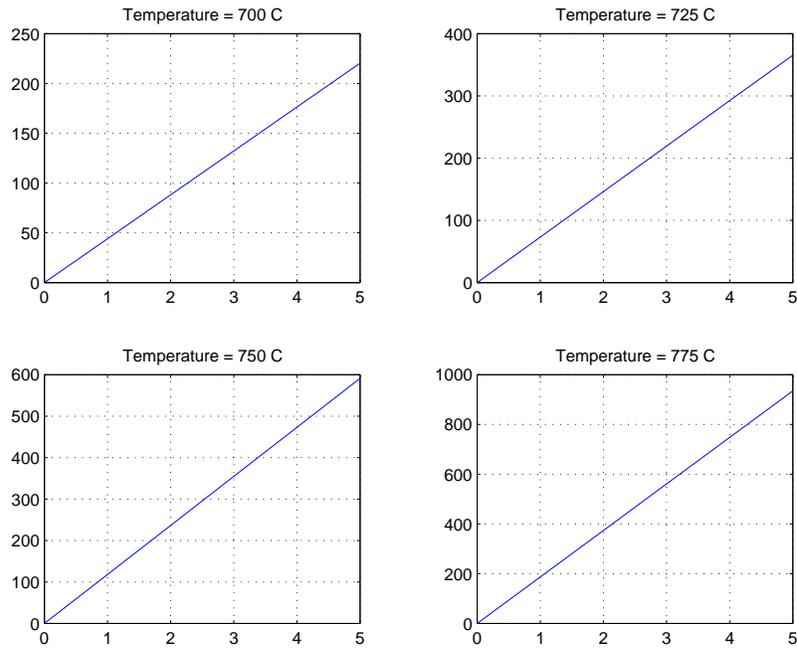


Figure 25: Film thickness (Angstroms) versus time (minutes) for a range of wafer temperatures in the thermally activated regime.

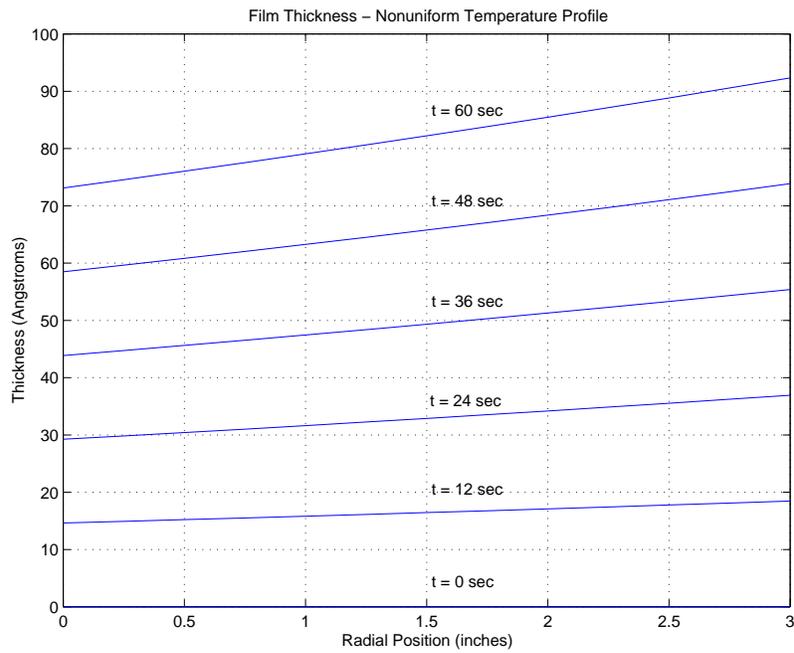


Figure 26: Film thickness profiles at various instants of time under nonuniform wafer temperature conditions.

2.4.1 Deposition Kinetics

One objective of our experiments was to validate an assumed Arrhenius relationship between wafer temperature and deposition rate, and determine the unknown parameters of this relationship under typical operating conditions. Once a relationship between wafer temperature and deposition rate was established, it could then be used in the model for deposition kinetics. Furthermore, it has also been used in the validation analysis for lamp heating models, where heat flux intensity profiles have been estimated using wafer temperature data which in turn has been inferred from film thickness data. This analysis is described later in Section 2.4.2.

Although the overall modeling project focuses on epitaxial growth, polycrystalline silicon was deposited instead of epitaxial silicon. This is because under conditions for thermally activated growth, deposition rates are relatively slow, making impractical the growth of films with thicknesses greater than one micron. For epitaxial silicon, available instrumentation at Northrop Grumman (FTIR) does not allow for the measurement of these relatively thin (less than one micron) film thicknesses. On the other hand, polycrystalline silicon film thicknesses between 5 Angstroms and 1 micron can be and were measured by the available ellipsometer and nanospec. Previous work by Brabant has shown growth rates for epitaxial and polycrystalline silicon to be approximately the same [6].

Procedure

Thin films of polycrystalline silicon were deposited from a silane precursor over a five minute period at reduced pressure (20 Torr). Deposition was performed under a combination of operating conditions consisting of four different wafer temperatures and three different silane flow rates (and hence three different silane mole fractions).

Temperatures were set in the surface–reaction controlled regime so that growth would be thermally activated. This regime is roughly from 600 C to 800 C for deposition of silicon from silane gas. We chose the following wafer temperatures at which to deposit silicon: 650 C, 700 C, 725 C, and 750 C.

The precursor gas mixture was silane (SiH_4) diluted at 2% in hydrogen (H_2). Three different flow rates were used for this 2% silane in hydrogen precursor: 1.5 slm, 2.5 slm, and 3.5 slm. Considering the 2% dilution, these three flow rates correspond to 30 sccm, 50 sccm, and 70 sccm of silane, respectively. The silane–hydrogen precursor was again diluted in 20 slm of the carrier hydrogen (H_2) gas. Thus, the three flow rates correspond to three mole fractions 1.4×10^{-3} , 2.2×10^{-3} , and 3.0×10^{-3} , respectively.

The reactor was operated in its usual, automatic mode (i.e., using PID control loops for temperature regulation and wafer rotation for uniformity), using pre–programmed recipes. Recipes were programmed to set chamber pressure at 20 Torr and to deposit silicon from silane precursor for five minutes onto the bare silicon wafers. Film thicknesses were measured later using the nanospec.

Results

We attempted twelve deposition experiments – one for each combination of the four wafer temperatures (650, 700, 725, 750 C) and three silane flow rates (30, 50, 70 sccm). Each of the twelve depositions was performed on a different wafer. At 650 C, there was no appreciable deposition for any of the flow rates. Thus, these three wafers provided no data for analysis. At 700 C and above, enough silicon was deposited so that measurements could be taken. Thickness was measured using the nanospec at five different points on the wafer surface. These five points are shown in Figure 27. In the case where temperature was 700 C and silane flow rate was 30 sccm, film thickness was less than 100 Angstroms, the minimum readable by the nanospec, and hence is recorded as less than 100 Angstroms. The data is presented in Table 1.

Deposition kinetics are modeled using the Arrhenius relationship

$$R_{\text{Si}} = k_0 \exp\left(\frac{-E_a}{R_g T_w}\right) X_{\text{SiH}_4} \quad (48)$$

where R_{Si} denotes deposition rate, k_0 denotes the pre–exponential constant, E_a denotes the activation energy, R_g denotes the gas constant, T_w denotes the wafer temperature, and X_{SiH_4} denotes the silane mole fraction. We call a plot of the logarithm of deposition rate versus inverse temperature an Arrhenius plot. The Arrhenius plots associated with the data we collected are shown in Figure 28. According to equation (48), the slope of the Arrhenius plot gives the activation energy E_a while the intercept (along with knowledge of

**Silicon Film Thickness and Deposition Rate
As Function Of Temperature and Silane Flow Rate**

<i>Operating Conditions</i>	
Pressure	20 Torr

Temperature = 700 C						
Wafer Position	Silane Flow Rate					
	30 sccm		50 sccm		70 sccm	
	h (A)	R (A/min)	h (A)	R (A/min)	h (A)	R (A/min)
1	< 100	< 20.0	314	62.8	392	78.4
2	< 100	< 20.0	337	67.4	393	78.6
3	< 100	< 20.0	332	66.4	414	82.8
4	< 100	< 20.0	332	66.4	391	78.2
5	< 100	< 20.0	327	65.4	412	82.4
Avg	< 100	< 20.00	328.4	65.68	400.4	80.08

Temperature = 725 C						
Wafer Position	Silane Flow Rate					
	30 sccm		50 sccm		70 sccm	
	h (A)	R (A/min)	h (A)	R (A/min)	h (A)	R (A/min)
1	365	73.0	526	105.2	684	136.8
2	368	73.6	553	110.6	716	143.2
3	365	73.0	539	107.8	693	138.6
4	365	73.0	521	104.2	686	137.2
5	365	73.0	526	105.2	689	137.8
Avg	365.6	73.12	533.0	106.60	693.6	138.72

Temperature = 750 C						
Wafer Position	Silane Flow Rate					
	30 sccm		50 sccm		70 sccm	
	h (A)	R (A/min)	h (A)	R (A/min)	h (A)	R (A/min)
1	584	116.8	851	170.2	1072	214.4
2	602	120.4	883	176.6	1116	223.2
3	594	118.8	859	171.8	1082	216.4
4	590	118.0	842	168.4	1070	214.0
5	587	117.4	848	169.6	1077	215.4
Avg	591.4	118.28	856.6	171.32	1083.4	216.68

Table 1: Measured silicon film thickness, h (Angstroms) and deposition rate, R (Angstroms per minute): Five minute deposition; three wafer temperatures; three silane flow rates.

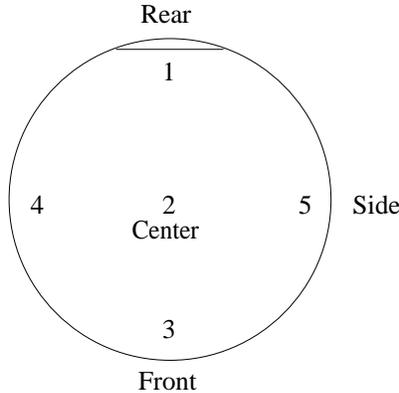


Figure 27: Points on wafers where thickness was measured. Rear refers to side of wafer closest to outlet (downstream); front closest to inlet (upstream).

the silane mole fraction) gives the pre-exponential constant k_0 . Computed parameters are given in Table 2. These values are then used in the implementation of the models for chemical reaction kinetics.

The activation energies range from 1.57 eV to 1.69 eV depending on silane mole fraction. This range is very close to the activation energy of 1.82 eV determined experimentally by the manufacturer, ASM, Inc. [29] In addition, the pre-exponential constants range from 3.8×10^8 to 1.85×10^9 , a range which includes the value of 7.9×10^8 given by the manufacturer.

The plots of deposition rate as a function of silane flow rate and silane mole fraction are shown in Figures 29 and 30, respectively. As expected, the relationship is nearly linear, with slope proportional to the exponential of inverse temperature, which confirms the assumption used in the Arrhenius model. The relationship between deposition rate and silane mole fraction can be used to determine the effect of reactant depletion across the wafer surface during deposition.

2.4.2 Lamp Heating

The objective of lamp heating experiments was to provide a basis for comparison to validate the analytically determined heat flux intensity profiles described in Section 2.3.4. In particular, the goal here was to isolate individual lamp groups in order to determine their heating effect on the wafer surface. Recall that lamp heating enters the wafer heat transfer ODE (19) via the lamp influence matrix B in the control input term, which was determined using analytical methods. It can be compared with experimentally determined lamp influence functions for purposes of validation.

Recall that for each individual lamp we computed a spatial profile that gives a value of heat flux intensity (in W/cm^2) irradiating each point on the wafer surface, where points are defined by their radial and azimuthal coordinates. Then, spatial profiles were averaged around all azimuthal positions to simulate wafer rotation and provide 1-dimensional profiles that are functions of radial position only. Finally, profiles for individual lamps were combined appropriately to give heat flux intensity profiles for the ten lamp groups and four lamp zones of the ASM Epsilon-1 reactor.

Analytically determined influence functions can be computed using an arbitrarily fine spatial resolution. The only cost is computing time, and the computations require relatively little time on a high performance workstation. For example, computation of heat flux intensity for one individual lamp on a cylindrical grid with 3200 points (100 radial, 32 azimuthal) required less than 3 minutes. As we shall see, the procedure that we used for experimental determination requires a physical measurement of film thickness at each spatial point. It is apparent that doing this for a grid with a large number of points (say more than 100) becomes impractical. Therefore, the experimentally determined profiles are used only for validation.

Methodology

The following methodology was used to empirically determine the lamp influence functions. Details of the experimental procedure are described later.

1. Polysilicon is deposited on a non-rotating wafer for a fixed period of time τ with lamp group i manually

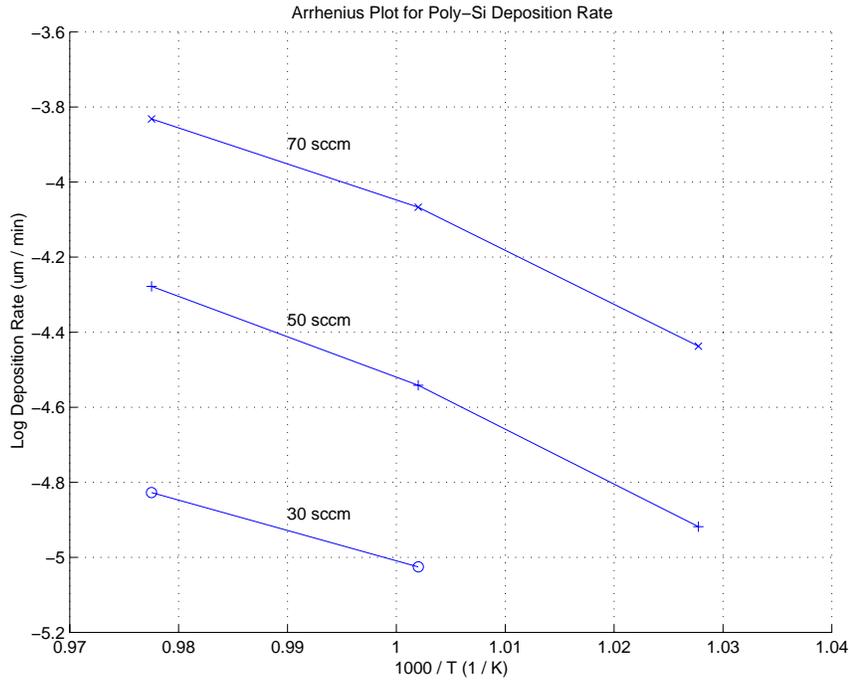


Figure 28: Arrhenius plots for silicon deposition from silane gas: each plot represents log of deposition rate (microns per minute) versus inverse absolute temperature for one of the three silane flow rates used.

Parameters For Arrhenius Relationship Describing Silicon Deposition Kinetics

$$R_{Si} = k_0 \exp\left(\frac{-E_a}{R_g T_w}\right) X_{SiH_4}$$

Symbol	Description	Data		
V_{mix}	Silane/Hydrogen Mixture Flow Rate (slm)	1.5	2.5	3.5
V_{SiH_4}	Silane Flow Rate (sccm)	30	50	70
X_{SiH_4}	Silane Mole Fraction ($\times 10^{-3}$)	1.4	2.2	3.0
E_a	Activation Energy (eV)	1.69	1.67	1.57
	Activation Energy (J/mol) ($\times 10^5$)	1.63	1.61	1.51
E_a/R_g	Ratio (K) ($\times 10^4$)	1.96	1.94	1.82
k_0	Pre-exponential Constant (um/min) ($\times 10^9$)	1.85	1.30	0.38
k_0	Pre-exponential Constant (cm/sec) ($\times 10^3$)	3.08	2.16	6.54

Table 2: Parameters for Arrhenius relationship describing silicon deposition kinetics: activation energy (eV) and pre-exponential constant (microns per minute) for each silane flow rate used.

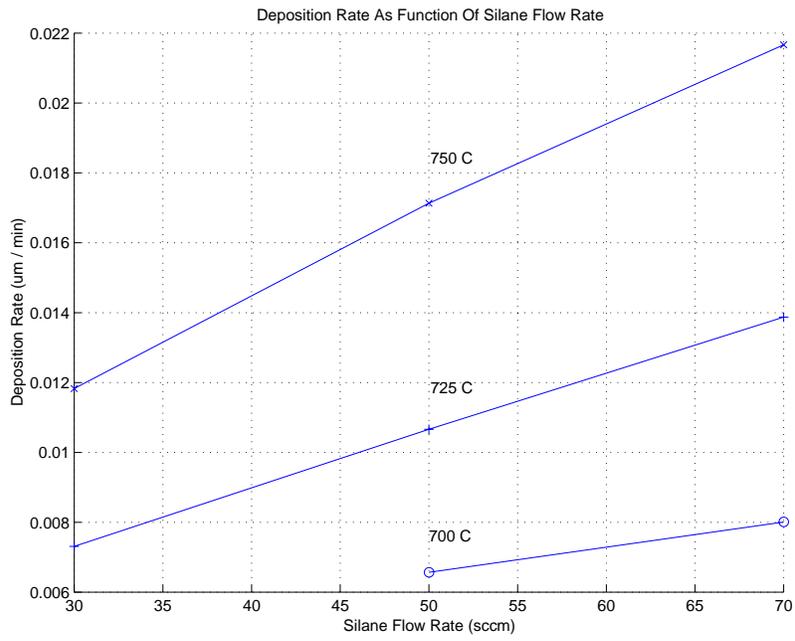


Figure 29: Growth rate as function of silane flow rate for each of three temperatures used.

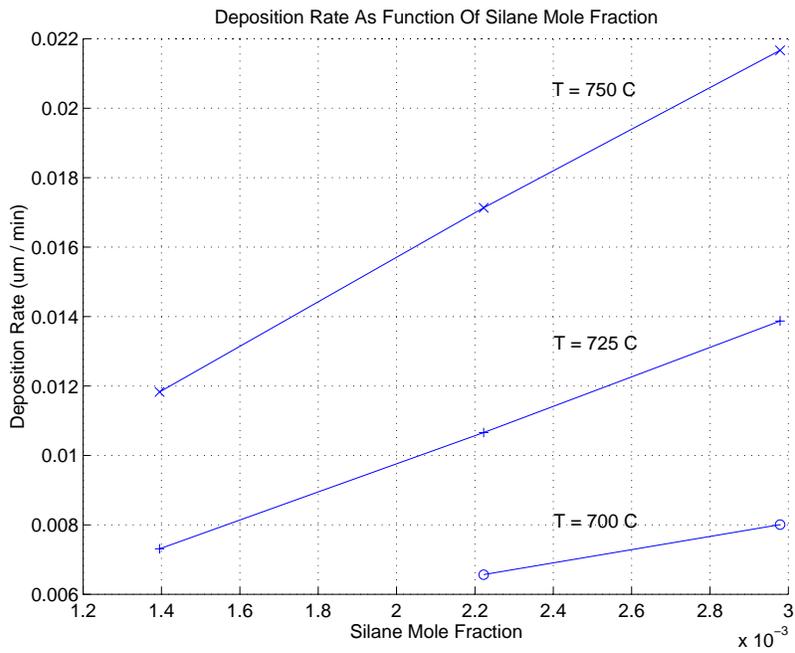


Figure 30: Growth rate as function of silane mole fraction for each of three temperatures used.

set to power setting P . Measuring thickness yields a thickness function $h_{(i,P,\tau)}(r, \theta)$. Growth rate is computed as $R_{(i,P)}(r, \theta) = h/\tau$. Averaging azimuthally gives growth rate in terms of radial position $R_{(i,P)}(r)$.

2. The Arrhenius law (46) is inverted to determine temperature as a function of radial position

$$T_{(i,P)}(r) = (E_a/R_g) (\ln(k_0 X_{\text{SiH}_4}) - \ln(R_{(i,P)}(r)))^{-1} \quad (49)$$

3. The temperature field $T_{(i,P)}(r)$ is substituted into the evolution equation for temperature (19) in the wafer heat transfer model. By applying the steady-state condition $\dot{T} = 0$ we can solve

$$0 = A_c T + A_r T^4 + A_v T + \Gamma + B_i P \quad (50)$$

for the discretized influence function $B_i(r)$.

Experimental Procedure and Results

Isolation of the individual lamp groups was achieved by operating the reactor in manual mode, i.e., with the automatic control loops for temperature regulation turned off. In manual mode, the lamp groups are no longer organized into four zones for the purpose of temperature control. Instead, each of the ten groups can be toggled on and off individually, and the power setting of each (between 0% and 100%) can be set manually. To isolate a particular lamp group, all others were turned off, while the power setting for the lamp group being tested is set manually to an appropriate level.

The wafer was heated with an individual lamp group, whose power setting was adjusted manually, until at least one of the thermocouple readings reached the range where deposition would occur, approximately 700 C. The exact temperature readings were not important because in the next step temperature would be inferred from thickness data. Then, flow of silane in hydrogen carrier was started. Silicon was deposited for five minutes. Wafer rotation was turned off so that effects of asymmetry would appear in the resulting deposition.

This procedure was followed to test four of the lamp groups: 1, 8, 9, and 10. Lamp group 1 is in the upper lamp array and radiates directly toward the top center of the wafer. Using lamp group 1 alone, we were able to heat the wafer to a temperature sufficiently high for deposition to occur and to record sufficient data for analysis. Lamp groups 8, 9, and 10 are in the lower lamp array and radiate toward the bottom of the susceptor. Due to conduction and losses throughout the susceptor, it was more difficult to heat the wafer using each of these lamp groups alone. Of the lamp groups we isolated in the lower array, only lamp group 8 provided enough radiant energy to heat the wafer to a temperature sufficiently high for deposition to occur. However, wafer temperature oscillated and was highly nonuniform in this case, causing the data to be unreliable. We focus now on the experiment that tested lamp group 1 from which reliable data was obtained.

Lamp group 1 was isolated and set to 45% of full power which brought the center thermocouple reading to 740 C, sufficiently high for deposition to occur. Silane flow rate was set at 30 sccm. After a five minute deposition period, the wafer was removed and thickness measurements were taken at 100 points on the wafer surface as shown in Figure 31.

Figure 32 shows two views of the resulting polysilicon film thickness profile. Thermally activated growth using the isolated lamp group 1 produces a “hill” of polysilicon. The deposition pattern reaches a maximum in a line across the wafer center parallel to the lamps in lamp group 1, and decreases toward the wafer edges. Qualitatively, this result is what we would expect given the geometry of this particular experimental setup.

The thickness data is then used to compute an empirically determined heat flux intensity profile for lamp group 1 as outlined previously. Figure 33 shows the result along with the analytically determined profile for purposes of comparison. The result indicates a reasonable agreement between the analytical model and experimental data.

2.5 Feature Scale Models

In order to study deposition on patterned wafers, the macroscopic level equipment and process model will be integrated with a feature scale model. This model should describe epitaxial growth of thin films at the

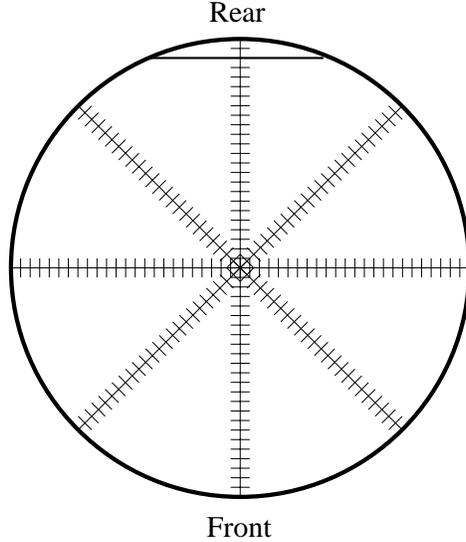


Figure 31: Points on wafer surface where thickness measurements were taken for testing of lamp group 1.

microscopic level, i.e., based on the physics of atom motion on surfaces. In this way, we can study how the oxide pattern on a wafer affects the deposition process, and in particular how pattern pitch affects uniformity.

The motion of atoms “raining down” on the surface of a wafer during the deposition process largely determines the arrangement of these atoms in terms of crystal structure and ultimately the quality of the film for its intended purpose [27]. One framework for studying atomic mechanisms of crystal growth is the terrace–ledge–kink model introduced by Burton, Cabrera, and Frank [7]. The general idea is illustrated in Figure 34. In this framework, arriving atoms land on the surface which contains terraces, steps, vacancies, and kinks. Atoms can evaporate, move on the terrace, cross over steps, nucleate new islands on any terrace, or become incorporated into steps. Transport of atoms over and along steps is the controlling factor in epitaxial growth technology, because of its influence on the film morphology, i.e., roughness [27].

One useful model for our purposes here is the “Eden model” for growth of interfaces relaxing by surface diffusion. In [13], an evolution equation for film growth is given under the assumption that the growth rate depends only on the local surface morphology through rotational invariants (such as curvature),

$$\frac{\partial h}{\partial t} = -K(\nabla^2)^2 h + \nu \nabla^2 h + \frac{1}{2} \lambda (\nabla h)^2 + \eta \quad (51)$$

where $h = h(r, t)$ is the height (thickness) profile, $\eta = \eta(r, t)$ is the statistical noise of incoming particle flux, and K and ν are identified from physical considerations.

We are currently investigating the established body of work in this area in the hope of developing an adequate microscopic model for purposes of modeling epitaxial growth of silicon and Si–Ge on wafers with oxide patterns.

3 Model Reduction

Dynamical system models have been presented in this report that describe the time evolution of various physical phenomena involved in RTCVD. Model reduction deals with methods for reducing the dimensionality of dynamical system models. The motivation is that models of lower dimension are less complex and easier to work with for various purposes such as simulation, optimization, and control. In this section, we describe our recent work in model reduction, mainly as applied to the models for wafer heat transfer described in this report. We also briefly discuss ongoing work to extend these ideas to the overall reactor model including evolution of gas temperature and chemical species transport.

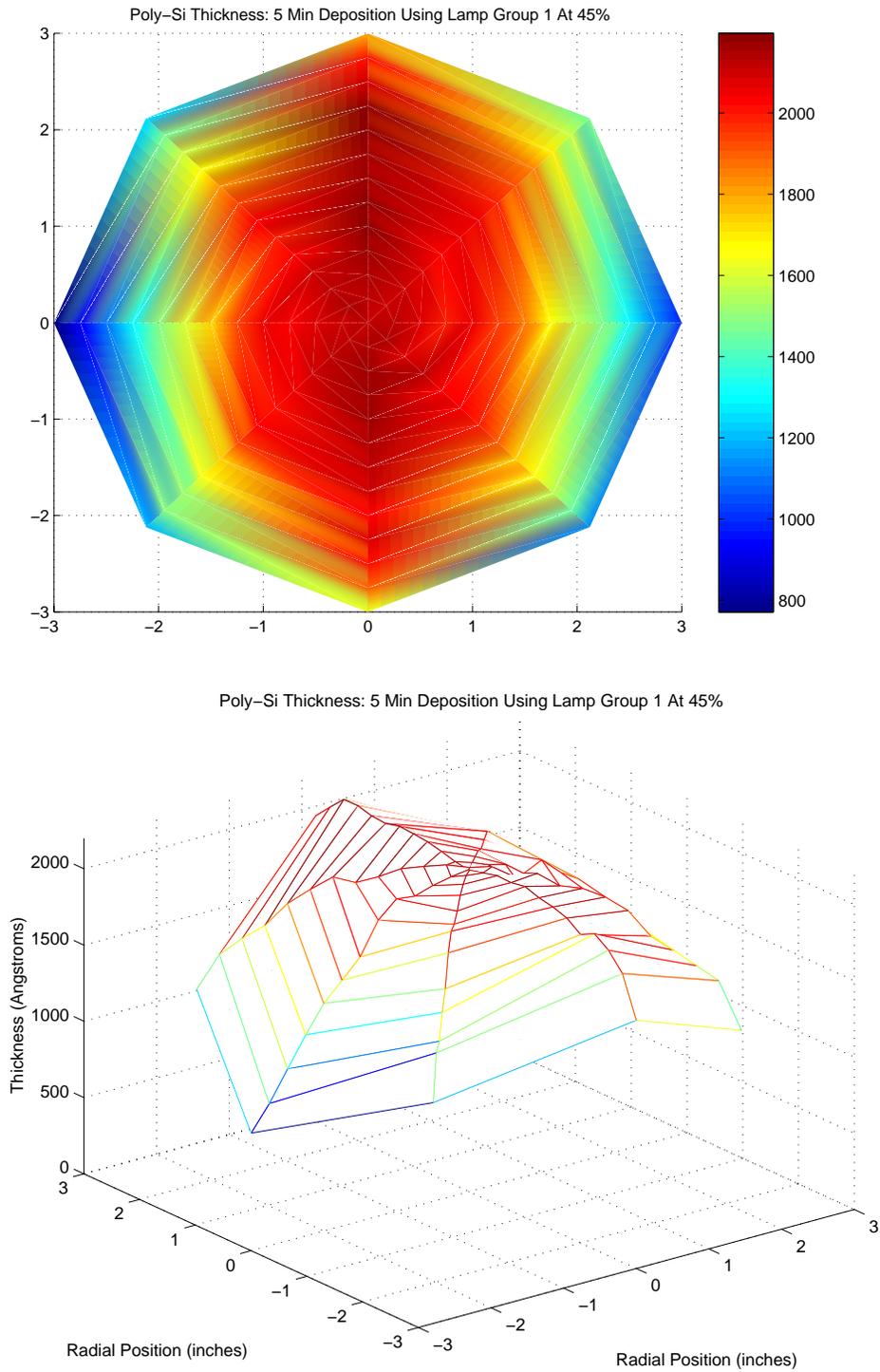


Figure 32: Two views of polysilicon film thickness profile resulting from 5 minute deposition using lamp group 1 at 45% power and silane flow rate of 30 sccm. Top figure shows contour map where colors/shades represent thicknesses. Bottom figure shows 3-dimensional view (“hill”).

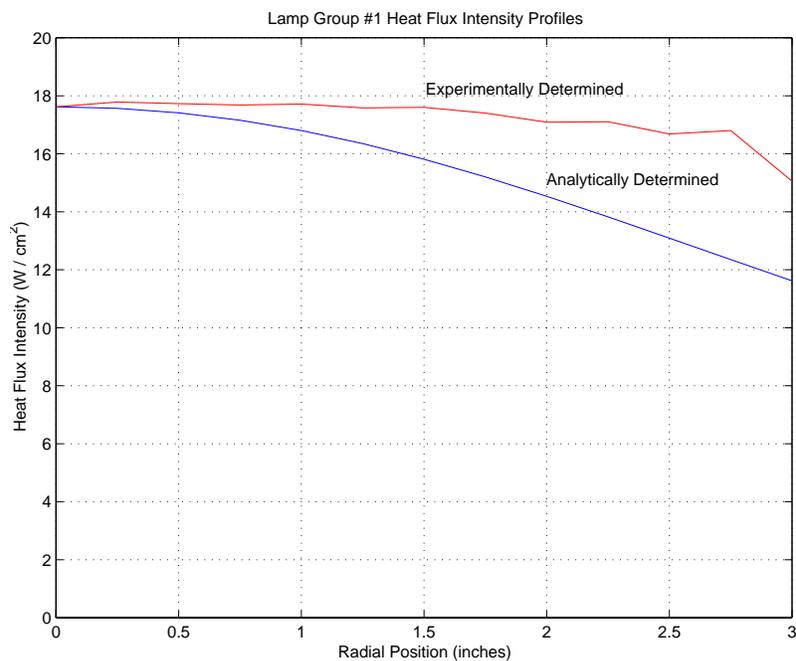
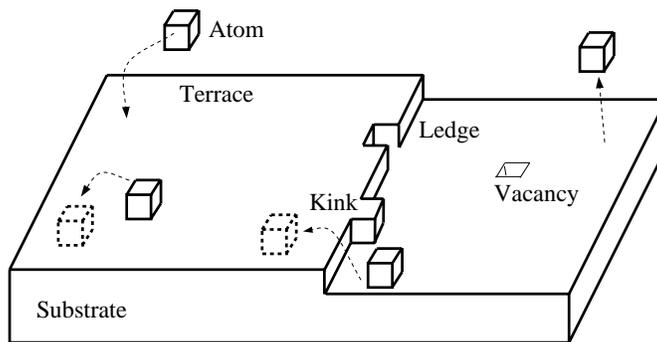


Figure 33: Experimentally determined heat flux intensity profile for lamp group 1 along with analytically determined profile for purposes of comparison.



c.f. Lagally, 1993

Figure 34: Terrace–ledge–kink framework for describing atom motion on surfaces.

3.1 Motivation and Overview

As described in Section 2 of this report, process and equipment models for RTCVD consist mainly of balance equations for conservation of energy, momentum, and mass, along with equations that describe the relevant chemical mechanisms. In their continuum form, the balance equations for RTCVD yield a system of nonlinear coupled PDEs with associated BCs and ICs. Lumped versions of the equations can be obtained using an appropriate discretization method, e.g., finite-elements, to yield a system of coupled nonlinear ODEs. The system of ODEs can be decoupled by invoking certain simplifying assumptions. Even the resulting simplified nonlinear system is typically of relatively high-order, so that not only is the model computationally demanding for simulation, but moreover it is computationally prohibitive for real-time control. Thus, the motivation for reducing the model order is apparent.

A general approach to model reduction is to find a coordinate transformation of the state space under which the state components can be ranked in a meaningful way in terms of their influence on the system behavior. Then, state components of the transformed system with relatively small influence can be truncated without substantially degrading the correctness, i.e., predictive capability, of the model. We note that for systems evolving on \mathbb{R}^n , each coordinate transformation can be identified with a corresponding set of basis n -vectors.

To illustrate ideas in model reduction, we focus on applying model reduction methods to the dynamical system ODE model describing wafer heat transfer, given by (19). We examine two model reduction approaches in this regard: the proper orthogonal decomposition (POD) and the method of balancing. A comparison of the effectiveness of the two approaches is presented via numerical studies using the wafer heat transfer model.

3.2 Simplified Wafer Heat Transfer Model

Recall that in Section 2.3.3, the evolution of the temperature field on the wafer surface was given by the ODE

$$\dot{T}_w = A_c T_w + A_r T_w^4 + A_v T_w + \Gamma + B P \quad (52)$$

with the initial condition

$$T_w(0) = T_{w_0} \quad (53)$$

where T_{w_0} represents the discretization of a typical initial temperature profile, e.g., uniform ambient (ambient temperature across entire surface).

To model the measurement of temperature at discrete points on the wafer surface via thermocouples, we augment the nonlinear state equation (52) with the linear output equation

$$T_{tc} = C T_w \quad (54)$$

where T_{tc} is a p -vector of thermocouple measurements, and C is a $p \times n$ matrix with entries corresponding to thermocouple locations.

As described in Section 2.3 of this report, the various parameters in (52) and (54) are derived from a detailed analysis of the ASM Epsilon-1 reactor. This includes the lamp heat flux intensity profiles, thermal parameters, and optical parameters used in the model. Under our modeling assumptions, the reactor has three independent lamp actuators, called lamp zones, which provide $m = 3$ independent influence functions. During a deposition process, there are no thermocouples on the wafer surface. However, for purposes of this study, we assume that there are thermocouples placed at $p = 3$ locations on the wafer surface: center, edge, and midpoint between center and edge.

Later in this report we use a linearized version of (52). To linearize, first observe that

$$\dot{x} = A_c x + A_r (x + \Gamma)^4 - A_r \Gamma^4 + A_v x + B u \quad (55)$$

has an equilibrium point at $x = 0$ and is equivalent to (52) using the changes of variable $x = T_w - \Gamma$ and $u = P - P_{ss}$, where P_{ss} is the control input that results in a steady state temperature field of $T_w = \Gamma$. Linearizing (55) about the origin gives

$$\dot{x} = A x + B u \quad (56)$$

with

$$A = A_c + A_v + 4F \quad (57)$$

where

$$[F]_{ij} = [A_r]_{ij} \Gamma_j^3$$

and x and u are translations of T_w and P , respectively.

As before, all evolution equations are numerically integrated using a fourth and fifth order Runge–Kutta algorithm, and the number of states, i.e., discretization points, for the original model and the linearized model is set at $n = 101$.

3.3 Proper Orthogonal Decomposition

One approach to finding a basis for the desired coordinate transformation is application of the POD, which is based on the classical Karhunen–Loeve decomposition of a stochastic process (see, e.g., [14, 18, 35, 47]). The POD is also known as the method of empirical eigenfunctions, or principal components analysis (PCA). The POD is a statistical pattern analysis technique for finding the dominant structures in an ensemble of spatially distributed data. These structures can be used as an orthogonal basis for efficient representation of the ensemble. In particular, the POD basis elements are the orthogonal eigenfunctions of the two–point spatial covariance of the data ensemble. When the data ensemble consists of vectors in \mathbb{R}^n , the POD basis vectors are just the columns of the matrix U in the singular value decomposition

$$X = U\Sigma V^T \quad (58)$$

where X is a matrix whose columns are the members of the data ensemble.

For the case of a dynamical system describing a flow, e.g., a heat flow, the data ensemble is typically time series data, i.e., “snapshots” of the flow captured at equally spaced intervals in time. The time series data is typically generated by simulating the original evolution equations, driven by an ensemble of representative control inputs, and using an ensemble of representative ICs. The POD coordinate transformation diagonalizes, or decouples, the covariance of the time series data. The basis elements of the coordinate transformation are the principal axes of the flow which generated the time series empirical data. Each has a corresponding eigenvalue (given by the entries on the diagonal of Σ in (58)), which provides a measure of the relative “energy”, i.e., mean square fluctuation, associated with that particular direction in the state space. This measure can also be interpreted as the relative amount of time that the flow spends along the corresponding principal axis. Thus, it serves as a well–defined measure of the influence of a state component on the system behavior.

The POD basis is optimal from the points of view of data compression and error minimization. Specifically, a truncated series representation of the data has a smaller mean square error than a representation by any other basis of the same dimension (see, e.g., [19]).

Application

The attractive properties of the POD have led to success in applying it to areas such as turbulence modeling (e.g., [19]) and pattern recognition (e.g., [46]). Recently, much work has been done to study its use for RTCVD model reduction (e.g., [1, 2, 3, 4, 36, 50]). We have performed a similar study toward finding a low–order approximation to (52).

To generate empirical time series data, i.e., snapshots of the wafer temperature field, the system (52–54) was simulated using two different types of control input recipes. (A recipe refers to a function giving the lamp power setting for each of the three lamp zones at each instance of time.) They are referred to as Ramp–Soak–Cool (RSC) and Perturbation–Of–Constant (POC).

Control Input Recipe Type 1 - Ramp–Soak–Cool

The RSC recipe mimics a typical processing recipe in which a lamp zone power setting is gradually ramped up from zero to full power, maintained at full power for a specified period of time, and then gradually ramped down from full to zero power, as shown in Figure 35. This recipe is applied to one of the lamp zones, while the other two zones are held at zero power. The simulation is then repeated using RSC individually for each

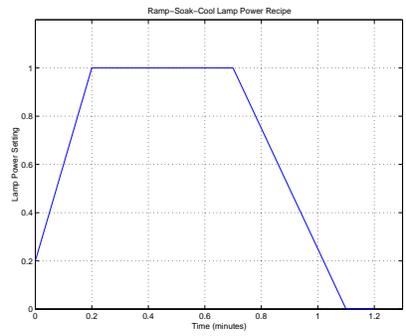


Figure 35: Lamp power settings for RSC recipe.

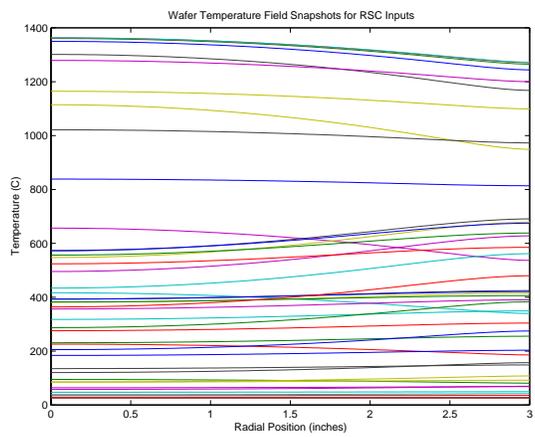


Figure 36: Snapshots of wafer temperature field with RSC input and uniform IC.

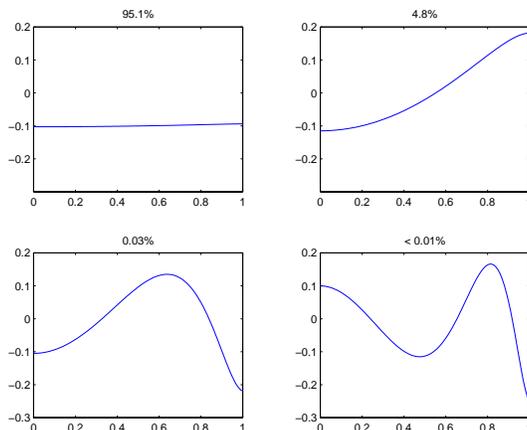


Figure 37: Basis elements computed using POD from RSC empirical data.

Lamp Power Settings for POC Recipe

Note: Settings are constant for all time.

Recipe	Zone 1	Zone 2	Zone 3
P_{unif}	0.0798	0.4265	0.1965
$P^{(1)}$	0.0718	0.4265	0.1965
$P^{(2)}$	0.0878	0.4265	0.1965
$P^{(3)}$	0.0798	0.3838	0.1965
$P^{(4)}$	0.0798	0.4691	0.1965
$P^{(5)}$	0.0798	0.4265	0.1768
$P^{(6)}$	0.0798	0.4265	0.2161

Table 3: Lamp power settings for POC recipe.

of the other two lamp zones. In this manner, the system response to excitation from an RSC recipe for each of the three lamp zones will appear in the time series data. The time series data is shown in Figure 36.

The entire ensemble (three sets) of time series data is combined and arranged into a data matrix, each column of which represents one “snapshot” of the wafer temperature field. The POD basis elements and associated eigenvalues, or relative energy values, are then computed via SVD and ranked according to magnitude of relative energy. The basis elements with the four largest eigenvalues are shown in Figure 37. Corresponding relative energy values are contained in Table 4.

Control Input Recipe Type 2 - Perturbation Of Constant

The POC recipe applies small perturbations of a predetermined set of constant power settings which, if left unperturbed, would result in a uniform steady state temperature field of 1000K. The perturbations are achieved by adjusting the power setting of each lamp zone, one at a time, first to 110% and then to 90%, of the original setting. This results in 6 different control recipes, as shown in Table 3. Note that if the nominal constant power settings are used, then the wafer temperature field will evolve as a uniform field for all time. Thus, the perturbations are used to elicit a response that would be characteristic of the system behavior in response to certain types of disturbances.

The system response to excitation from each of the six POC recipes is sampled and combined as the time series data for computing POD basis elements. Time series snapshots are shown in Figure 38. Once again, POD basis elements are computed and ranked by corresponding relative energy value. The basis elements with the four largest eigenvalues are shown in Figure 37. Corresponding relative energy values are contained

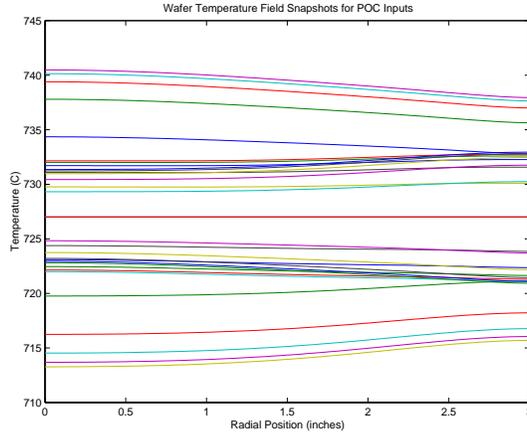


Figure 38: Snapshots of wafer temperature field with POC input and uniform IC.

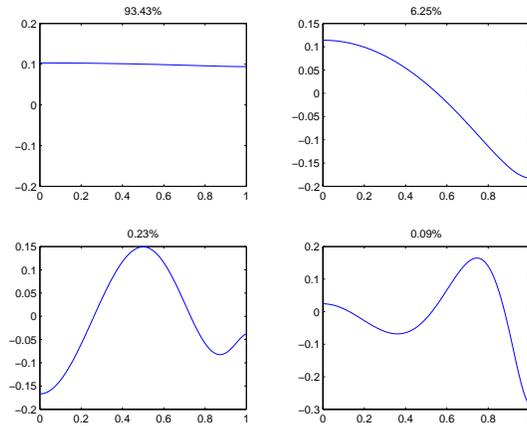


Figure 39: Basis elements computed using POD from POC empirical data.

in Table 4.

Shortcomings

It is clear that the efficiency of a basis determined via the POD method depends strongly on how well the data ensemble captures the relevant system behavior. This leads to serious shortcomings for model reduction of control systems with inputs and outputs. The POD basis elements will be sensitive to the choice of control inputs, ICs, and BCs used to generate the empirical time-series data. For nonlinear systems, small perturbations in these parameters may produce qualitatively different system responses. In addition, the data may fail to capture dynamical effects occurring at widely differing time scales. Usually, these issues are ignored, because the model is only being used for a particular purpose, e.g., to simulate tracking of one particular reference trajectory. In that case, a representative set of control inputs, BCs, and ICs is selected for generating the data ensemble. But the optimality property of the POD basis, and the predictive capability of the resulting reduced order model, may be localized to a relatively small region of the space of allowable inputs, BCs, and ICs. In addition, the POD approach fails to consider the influence of state components on the system measurements, or outputs. It would be desirable to truncate state components that have small influence on the outputs, i.e. that do not appear in our measurements. The POD method does not do this, thus its efficiency may be diminished when constructing “black-box” input-output models of the system.

3.4 Balancing

In response to concerns regarding the POD method, we considered an alternative approach based on the method of balanced realizations (see, e.g., [12, 33]). In this method, a coordinate transformation is computed which allows state components to be ranked according to their influence on the input–output behavior of the system as measured by the Hankel norm of the system, i.e., the gain from past inputs to future outputs. The resulting basis for the linear transformation makes the transformed realization “equally controllable and observable,” i.e. balanced, and depends only upon intrinsic properties of the original model, specifically controllability and observability, embodied in its evolution equation and output equation. In the linear case, explicit bounds can be computed for the error between the original and reduced order models. These error bounds are independent of any particular set of control input signals, BCs, or ICs. Although explicit error bounds have not yet been found for the nonlinear case, we still wish to exploit the property that the correctness of an approximation using a truncated balanced realization does not depend upon generating a representative data ensemble. Furthermore, the balancing method emphasizes state components that are both strongly controllable and strongly observable, so that state components which are least likely to influence the measurements are truncated.

The theory of balancing for linear systems is well established. See Appendix D for an exposition of the main ideas and results. Computation of a basis for the balanced realization is readily accomplished by use of matrix operations and decompositions. A general theory and procedure for balancing of a class of nonlinear systems has been presented [43], but in contrast to the linear case, algorithms for performing the required computations do not currently exist. We note that development of such algorithms is a topic of current research being undertaken by the authors. However, in the meantime, application of the balancing method is restricted to linear systems. Therefore, we apply the balancing method to the linearized version (56) of the nonlinear wafer heat transfer ODE (19).

Numerical Difficulties for Nonminimal Systems

Numerical difficulties may arise in the use of the balancing model reduction procedure. The problem of calculating the balancing transformation T_{bal} will be badly conditioned when the matrix $W_c W_o$ has a high condition number, i.e., when some modes are nearly uncontrollable or unobservable. To see this, consider Glover’s algorithm for computing the balancing transformation. Given W_c and W_o , let the Cholesky factorization of W_o be given by

$$W_o = R^T R.$$

Difficulties immediately appear when W_o is singular, or nearly singular. If R can be computed, then $R W_c R^T$ is diagonalized as

$$R W_c R^T = U \Sigma^2 U^T$$

with $U^T U = \mathbb{I}$ and $\Sigma = \text{diag}[\sigma_1, \dots, \sigma_n]$. The balancing transformation is given by

$$T_{bal} = \Sigma^{-1/2} U^T R.$$

Once again, difficulties appear when W_c is singular, or nearly singular.

It turns out that we encounter precisely these difficulties when attempting to calculate a balancing transformation for the linearized wafer temperature field evolution system. The condition numbers of the objects of interest are

$$\begin{aligned} \text{cond}(W_c) &= 3.2 \times 10^{18} \\ \text{cond}(W_o) &= 3.8 \times 10^{18} \\ \text{cond}(W_c W_o) &= 9.4 \times 10^{18} \end{aligned}$$

so that the system is nearly uncontrollable and unobservable, at least to an order of numerical precision that makes it impractical to compute a balancing transformation using the standard technique. This result is expected, since lamp influence functions and initial wafer temperature profiles are always smooth and relatively flat. Hence, there will be non–smooth or spatially fluctuating temperature profiles that are almost impossible to produce using the available control inputs.

Schur Method For Balancing

In order to alleviate the numerical difficulties, we use a procedure presented by Safanov and Chiang [40] for calculating a realization of a k th-order reduced model. The reduced model is not necessarily balanced, but has transfer function $\widehat{G}(s)$ which is exactly the same as that obtainable by applying Moore's (or Glover's) balanced realization model reduction procedure to any minimal realization of $G(s)$, the transfer function of the original n th-order model. Since the reduced transfer function is the same, it enjoys the same attractive error bound that was shown by Glover.

The procedure presented in [40] is referred to by Safanov and Chiang as a ‘‘Schur method’’ for balanced-truncation model reduction because the resulting transformation is computed using the Schur decomposition of the matrix $W_c W_o$. The idea is that balancing can be avoided altogether. Instead, we only need to compute bases for the left and right eigenspaces associated with the ‘‘big’’ eigenvalues of $W_c W_o$. The ordered Schur form of $W_c W_o$ is used to compute orthonormal bases enabling numerically robust computation of a nonbalanced realization of the k th-order reduced model $\widehat{G}(s)$. The algorithm is stable and effective without regard to nearness to unobservability or uncontrollability.

General Procedure

The general procedure is summarized as follows. Let (A, B, C, D) be a realization of $G(s)$ and let k denote the order of the reduced model.

1. Compute matrices $V_{r,big}, V_{l,big} \in \mathbb{R}^{n \times k}$ whose columns form basis for the respective right and left eigenspaces of $W_c W_o$ associated with the ‘‘big’’ eigenvalues $\sigma_1^2, \dots, \sigma_k^2$. This is done using the ordered Schur decomposition of $W_c W_o$ and is described in the next paragraph.

2. Let

$$E_{big} = V_{l,big}^T V_{r,big}$$

and compute the SVD

$$U_{E,big} \Sigma_{E,big} V_{E,big}^T.$$

3. The not necessarily balancing transformations are

$$S_{l,big} = V_{l,big} U_{E,big} \Sigma_{E,big}^{-1/2} \in \mathbb{R}^{n \times k}$$

$$S_{r,big} = V_{r,big} V_{E,big} \Sigma_{E,big}^{-1/2} \in \mathbb{R}^{n \times k}$$

so that the reduced model is given by

$$(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}) = (S_{l,big}^T A S_{r,big}, S_{l,big}^T B, C S_{r,big}, D)$$

Specific Procedure for Eigenspaces of $W_c W_o$

1. Compute an orthogonal real matrix V such that $V W_c W_o V^T$ is upper triangular, i.e., put $W_c W_o$ in Schur form. The fact that W_c and W_o are real and symmetric ensures the existence of a real Schur transformation matrix V .
2. Compute orthogonal real transformations V_a and V_d which order the Schur forms in ascending and descending order, respectively,

$$V_a^T W_c W_o V_a = \text{diag}(\lambda_{a,n}, \dots, \lambda_{a,1}) + T_a \tag{59}$$

$$V_d^T W_c W_o V_d = \text{diag}(\lambda_{d,1}, \dots, \lambda_{d,n}) + T_d \tag{60}$$

$$\tag{61}$$

where T_a and T_d are strictly upper triangular and such that

$$\{\lambda_{a,1}, \dots, \lambda_{a,k}\} = \{\lambda_{d,1}, \dots, \lambda_{d,k}\} = \{\sigma_1^2, \dots, \sigma_k^2\} \tag{62}$$

$$\{\lambda_{a,k+1}, \dots, \lambda_{a,n}\} = \{\lambda_{d,k+1}, \dots, \lambda_{d,n}\} = \{\sigma_{k+1}^2, \dots, \sigma_n^2\} \tag{63}$$

$$\tag{64}$$

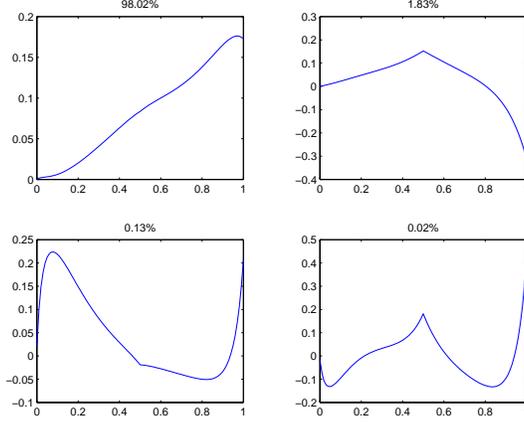


Figure 40: Left basis elements for balancing transformation.

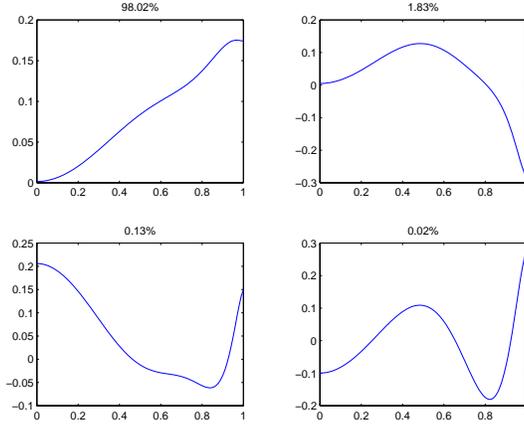


Figure 41: Right basis elements for balancing transformation.

3. Partition V_a and V_d as

$$V_a = \left[\overbrace{V_{r,small}}^{n-k} \mid \overbrace{V_{l,big}}^k \right] \quad (65)$$

$$V_d = \left[\overbrace{V_{r,big}}^k \mid \overbrace{V_{l,small}}^{n-k} \right] \quad (66)$$

Balancing Results

The Schur method for balancing is applied to the linearized model (56) and (54), i.e., the realization $(A, B, C, 0)$. The condition numbers for all of the matrices used in the procedure are less than 1000. The left and right basis elements for the balancing transformation are shown in Figures 40 and 41. The corresponding relative energy values are given in Table 4.

3.5 Comparison and Remarks

Validation of the predictive capability of the reduced models is accomplished by comparing simulation results using the original n th-order model with simulation results using reduced k th-order approximations for various values of the reduced model order k . In particular, the maximum deviation of the output signals, i.e., the thermocouple readings, between the original and reduced models are computed for each of the model reduction approaches we have previously described.

Procedure

Model reduction is achieved by projecting the evolution equations of our wafer thermal model onto a k -dimensional subspace ($k \ll n$) of the original n -dimensional state space. This is done via standard Galerkin projection. The k -dimensional subspace is chosen as the span of a set of basis vectors $\{\phi_1, \dots, \phi_k\}$ generated using one of the previously described POD or balancing methods. In particular, they are the k basis elements with the largest corresponding energy values. Thus, the $n - k$ state components with the smallest energy values are deleted by this transformation. Each projection results in a k th-order ODE model whose evolution in time yields an approximation to the evolution of the original n th-order model. Finally, an approximation to the wafer temperature field T_w is reconstructed by transforming back to the original state space.

If we arrange the basis vectors in a matrix

$$\Phi = [\phi_1 \dots \phi_k]$$

then the approximation to the wafer temperature field, \hat{T}_w , is given by

$$\hat{T}_w = \Phi z \tag{67}$$

where $z \in \text{span}\{\phi_1, \dots, \phi_k\}$ is the state variable in the k th-order ODE model

$$\dot{z} = \Phi^T [A_c \Phi z + A_r (\Phi z)^4 + A_v \Phi z + \Gamma + BP] \tag{68}$$

resulting from Galerkin projection. Observe that Φ^T maps elements of the original n -dimensional state space to the reduced order k -dimensional state space, while Φ maps elements of the reduced order space back to the original space. Also note that only in the case when $n = k$ does Φ represent a complete orthonormal basis, i.e., $\Phi \Phi^T = \Phi^T \Phi = \mathbb{I}$. Once we have selected Φ , using, e.g., POD or balancing, (68) is numerically integrated and an approximation to T_w is reconstructed using (67).

Validation Tests

The model reduction methods are tested by numerically integrating the original model and reduced models using a uniform initial temperature field and two test recipes as lamp control inputs. The test recipes are different from those used to generate the RSC and POC data ensembles.

$$\text{Test Recipe 1} \quad P = [0.5 \ 0.5 \ 0.5] \quad t \in [0, 1]$$

$$\begin{aligned} \text{Test Recipe 2} \quad P &= [1.0 \ 0.0 \ 0.0] & t \in [0, 0.4) \\ P &= [0.0 \ 1.0 \ 0.0] & t \in [0.4, 0.7) \\ P &= [0.0 \ 0.0 \ 1.0] & t \in [0.7, 1.0] \end{aligned}$$

The reduced models for $k = 1, 2, 3, 4$, and 5 are numerically integrated using the same control recipes and ICs as for the original $n = 101$ order model. Simulated thermocouple readings are recorded for each simulation. The error between the original and reduced models is computed as

$$e(k) = \|T_{tc} - \hat{T}_{tc}\|_{max}, \quad k = 1, 2, 3, 4, 5 \tag{69}$$

where we define the norm $\|y\|_{max}$ for time-dependent p -vector $y(t)$ as

$$\|y\|_{max} = \max \{y_i(t) : 0 \leq t < \infty, 1 \leq i \leq p\} \tag{70}$$

where $y_i(t)$ corresponds to the temperature reading of thermocouple i at time t . Thus, (69) gives the maximum deviation between actual and estimated thermocouple readings over the entire simulated time sequence and over all three thermocouples, i.e., a “worst case” error.

Results

Due to the shape of the lamp heat flux intensity profiles and the smoothing effect of the diffusion operator, the evolution of the wafer temperature field does not produce especially interesting behavior, e.g., spatial profiles whose fluctuations from the mean vary substantially in the mean square sense from the initial profile,

Percent Energy Associated With Transformation Basis Elements

Method	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
POD RSC	95.06	4.77	0.14	0.03	0.00
POD POC	93.43	6.25	0.23	0.09	0.01
Balancing	98.02	1.83	0.13	0.02	0.00

Table 4: Normalized eigenvalues, i.e., percent energy, corresponding to basis elements used in model reduction for POD method with RSC data, POD method with POC data, and balancing approach.

Maximum deviation (degrees C) between outputs of original and reduced models

Simulation	Reduction Method	Reduced Model Order				
		1	2	3	4	5
Test 1	POD RSC	27.23	2.68	0.58	0.11	0.01
	POD POC	26.85	1.26	1.13	0.10	0.05
	Balancing	50.68	7.03	0.44	0.08	0.02
Test 2	POD RSC	72.33	5.22	1.48	0.18	0.05
	POD POC	72.60	4.79	4.35	0.43	0.10
	Balancing	80.81	14.28	1.70	0.12	0.04

Table 5: Maximum deviation (degrees C) between outputs of original and reduced models for POD method with RSC data, POD method with POC data, and balancing approach.

assuming the initial profile is relatively smooth. Thus, we expect little difficulty in capturing the essence of the input–output behavior of the system in a low dimensional model. Our results show that this is indeed the case.

Tables 4 and 5 give the relative energy values for basis elements, and the maximum thermocouple temperature deviations for the original and reduced order models. Figures 42, 43, and 44 show simulated thermocouple readings resulting from test recipe 1, for the original $n = 101$ order model, and reduced models of order $k = 1, 2,$ and 3 . Figures 45, 46, and 47 show simulated thermocouple readings resulting from test recipe 2. From the figures, it is clear that the measured response for the reduced models is close to that of the original. We further analyze the data as follows.

For the inputs used in the validation tests, the input–output behavior of the wafer heat transfer system can be reconstructed using reduced models of order 4 so that thermocouple readings are within 1 degree C of the readings using the original model. This holds whether the POD or balancing method is used, and for whichever set of empirical data was used for computing the POD transformation. Even reduced models of order 2 produce a reasonable approximation with “worst case” errors less than 15 degrees C.

The POD method appears to have performed slightly better than the balancing method in this study. One reason for this result is that the balancing transformation was computed for the linearized system, while the validation tests were performed for the reduced order nonlinear system. Another reason is the simple input–output behavior of this particular system. The principal components of the flow are relatively insensitive to the choice of inputs, and hence, any set of principal components derived from empirical data for this system will likely be efficient for model reduction purposes.

3.6 Reduction from CFD Models

Figure 48 shows time series data points, or snapshots, of the temperature field in the process chamber of the ASM Epsilon–1. This is a sampled data representation of the gas temperature evolving in time over a

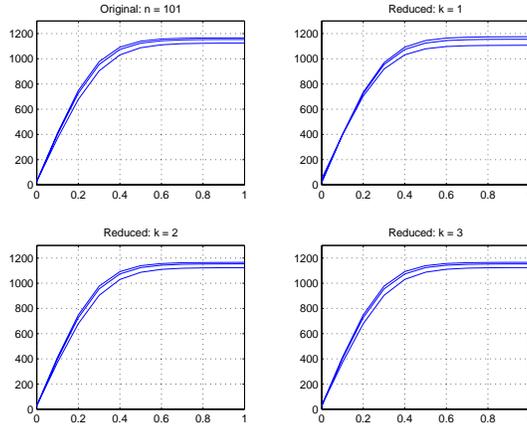


Figure 42: Thermocouple readings for original and reduced models with Test Recipe 1 using transformation from POD RSC.

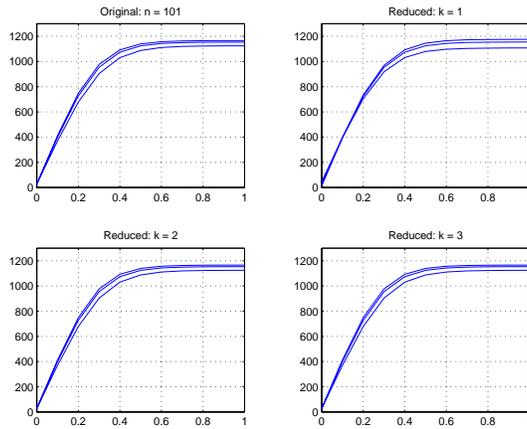


Figure 43: Thermocouple readings for original and reduced models with Test Recipe 1 using transformation from POD POC.

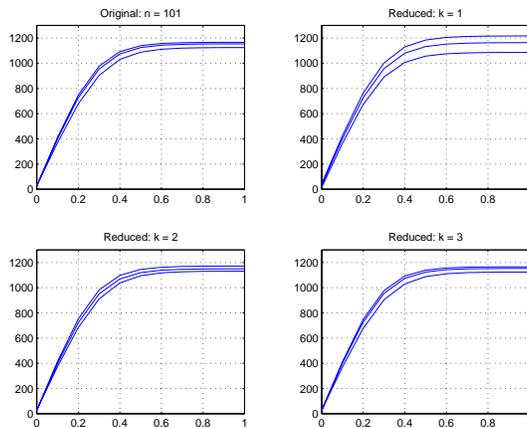


Figure 44: Thermocouple readings for original and reduced models with Test Recipe 1 using balancing transformation.

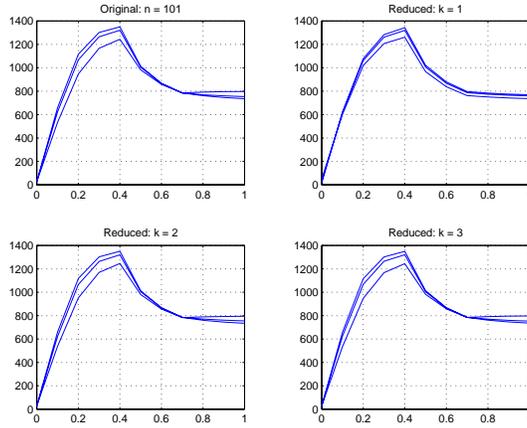


Figure 45: Thermocouple readings for original and reduced models with Test Recipe 2 using transformation from POD RSC.

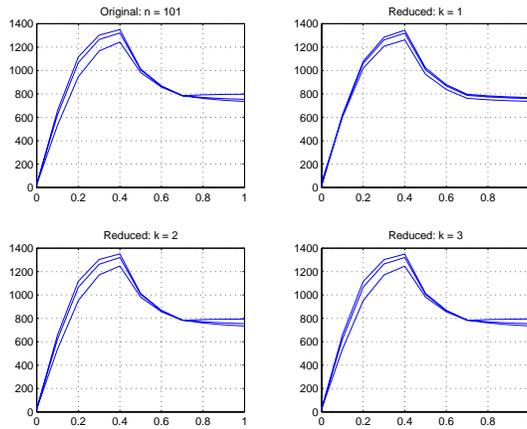


Figure 46: Thermocouple readings for original and reduced models with Test Recipe 2 using transformation from POD POC.

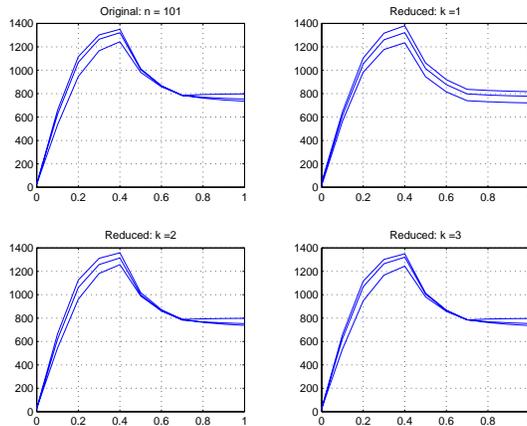


Figure 47: Thermocouple readings for original and reduced models with Test Recipe 2 using balancing transformation.

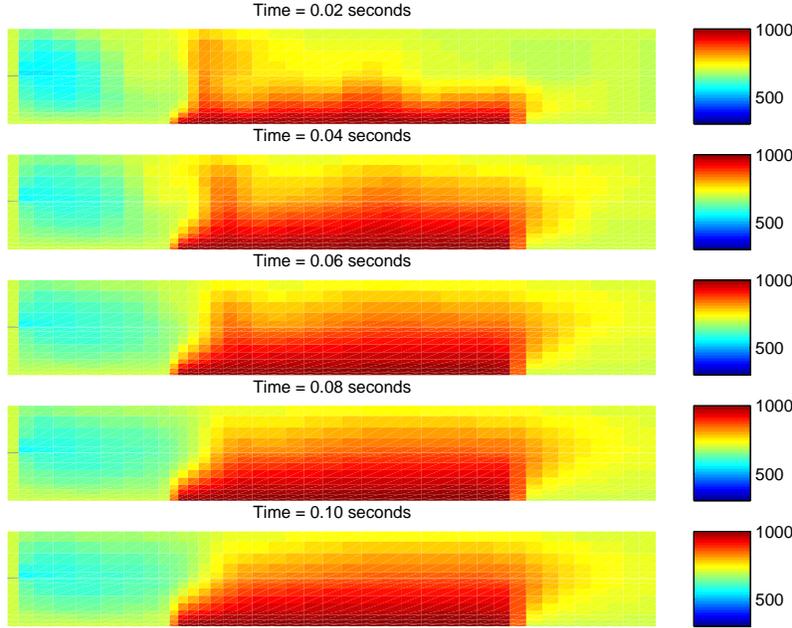


Figure 48: Snapshots of temperature field in process chamber: $t = 0.02, 0.04, 0.06, 0.08, 0.10$ seconds. Flowing gases are heated as they pass the wafer.

period of 0.1 seconds under typical operating conditions.

In order to generate this simulation data, the fully coupled set of nonlinear transport equations were solved on a grid with 720 nodes using Fluent. Even if the energy balance equations for gas temperature were uncoupled from the fluid flow and species transport equations, the simulator is numerically integrating a model with 720 states, one for each grid point. Coupling among transport equations makes the order of the system much higher than that.

This simulation required approximately 15 seconds of computing time to converge to a solution for each simulated time step, or a factor of 750 slower than the actual time period being simulated. This means that not only is the CFD model computationally demanding for simulation, but also it is computationally prohibitive for purposes of real-time model-based control. In addition, controller design and optimization for a system with hundreds of states will be excessively complicated.

We apply our methodology for model reduction, in which we extract spatial structures, or patterns, that dominate the flow dynamics as given by the snapshots of the heat flow. These “coherent structures”, or principal components, are shown in Figure 49. We note that these are the empirically determined eigenfunctions of the two-point spatial covariance of the fluctuations from the mean heat flow. They contain 92.7%, 6.9%, and 0.5%, respectively, of the energy in this flow. This preliminary analysis suggests that we can model the evolution of gas temperature with a model of dimension 2 to a high degree of correctness.

A similar analysis has been done with time evolution of silane mole fraction. Figure 50 shows the snapshots recorded over a simulated period of 0.2 seconds (silane mole fraction reaches steady-state in approximately double the time of gas temperature). Figure 51 shows the corresponding principal components. Again, 99% of the energy is contained in 2 basis functions.

4 Conclusion

Development of high-fidelity and low-order models is the first step in achieving the objectives of this project. The task now is to use the results to improve manufacturing effectiveness. To this end, we outline here some directions for further work, and summarize what we have accomplished so far.

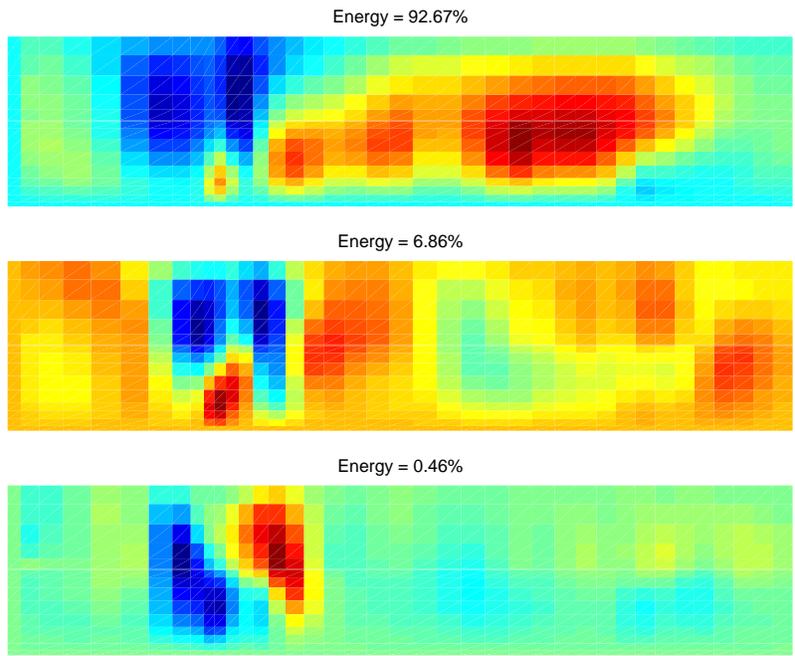


Figure 49: Principal components (empirically determined eigenfunctions) of heat flow in process chamber, corresponding to energies 92.7%, 6.9%, and 0.5% (top to bottom).

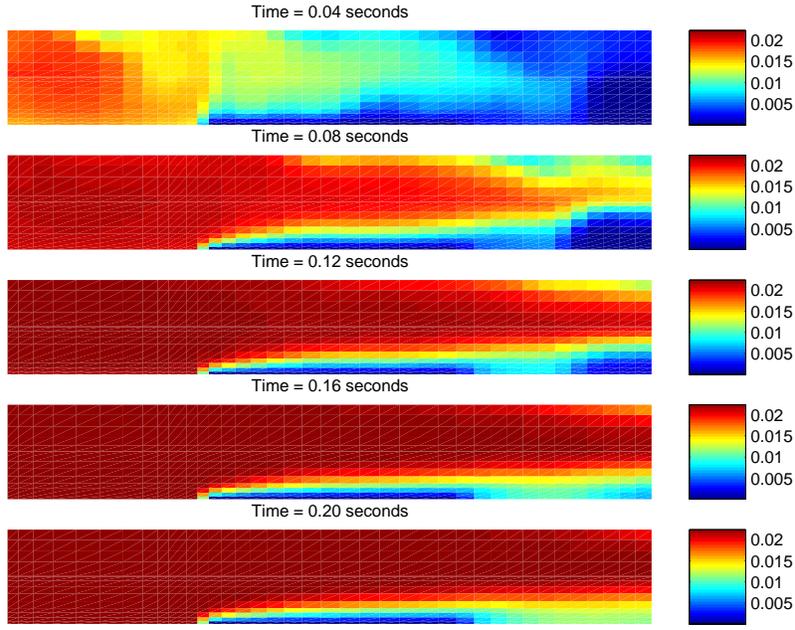


Figure 50: Snapshots of silane mole fraction in process chamber: $t = 0.04, 0.08, 0.12, 0.16, 0.20$ seconds. Silane is depleted as surface reactions take place.

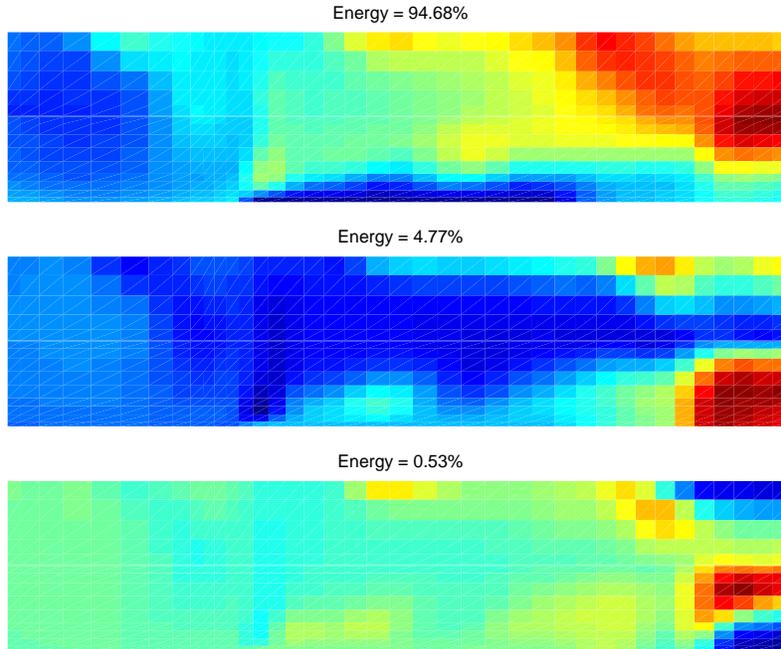


Figure 51: Principal components (empirically determined eigenfunctions) of silane species transport in process chamber, corresponding to energies 94.7%, 4.8%, and 0.5% (top to bottom).

4.1 Future Directions

There are several ways in which we can attempt to improve manufacturing effectiveness. One is by improving product quality as measured by deposition uniformity. Another is by providing the process engineer with tools to increase efficiency and provide insight and predictions.

Of particular interest is the improvement of deposition uniformity in the presence of microfeatures on the wafer surface. One way in which we hope to attack this problem is by quantifying the tradeoffs between pattern pitch and uniformity. Then, for a given uniformity requirement, we can find the minimum allowable pattern pitch, or for a given pattern pitch, we can find the maximum possible uniformity.

Currently, the mechanisms which affect deposition uniformity in the presence of oxide patterns on the wafer surface are not well understood. Development of the feature scale model, which will work in conjunction with the reactor and process model, will provide insight toward understanding these mechanisms and important cause and effect relationships.

A process recipe consists of a sequence of processing steps, each of which performs a given task over a fixed time period, and is characterized by a set of operating conditions selected by the process engineer. Such operating conditions include choice of process gases, inlet flow rates, chamber pressure, and wafer temperature. The ASM Epsilon-1 reactor system has the capability of controlling these conditions automatically, according to the pre-programmed recipe, during the processing run. In addition, there are several settings over which the process engineer has control and which remain constant throughout the process recipe. These are manually set by the engineer, and include inlet injector slit widths, thermocouple offsets, and PID control gains in temperature control loops.

A given process recipe and set of reactor settings will produce a corresponding thin film deposition on the wafer surface. We are interested in certain characteristics of this result including deposition thickness, uniformity, and morphology. Therefore, we want our models to ultimately have the capability to predict these characteristics given a process recipe and set of reactor settings. Once this capability is achieved, and experimentally verified, the models can be used for development of tools to aid the process engineer in determining

- optimal thermocouple offsets for a given uniformity requirement and a given set of process conditions;

- optimal inlet injector slit widths for a given uniformity requirement and a given set of process conditions;
- final thickness, deposition rate, film composition, uniformity, and morphology for a given process recipe and given thermocouple offsets and inlet injector widths, all relevant recipe steps included; input to this tool would be similar to how the recipe is input to the ASM Epsilon-1 itself; output would be a set of predicted results;
- optimal pressure, flow rate, and temperature recipes;
- deleterious effects such as deposition on walls and mechanical stresses on equipment and product for a given recipe;
- optimal PID gains for Foxboro temperature controllers.

The tools would be useful in "what if" analysis of various production runs. The recipes could be tested using simulation without actually performing the run, avoiding waste of time, energy, and materials, and allowing for rapid tuning of adjustable parameters and testing of ideas before production. The process engineer could verify that the thermocouple offsets and inlet slit widths being used are good or find ones that provide better uniformity. The tools should also provide some insight as to why certain settings are better than others.

Successful implementation of low-order process and equipment models will also be useful in design of real-time model-based control strategies for tracking and regulation of desired temperature profiles. Using sophisticated control systems simulation software packages such as Simulink⁴ and SystemBuild⁵, "system blocks" can be constructed from the low-order models. These system blocks interface to other system blocks via user-defined inputs and outputs. In this manner a controller block can be designed and a feedback control strategy can be simulated. The power of these models can be extended further via commercially available rapid prototyping tools, which work in conjunction with the simulation software, and provide the capability for system blocks to connect physically with the actual inputs and outputs of the systems they are simulating.

4.2 Summary

We have described work done to date on a joint project between Northrop Grumman ESSD and the ISR to investigate the epitaxial growth of Si-Ge heterostructures in the ASM Epsilon-1 RTCVD reactor. An important goal of this research is to improve manufacturing effectiveness by controlling and optimizing deposition characteristics such as uniformity in the presence of microfeatures on the wafer surface.

High-fidelity models have been developed and implemented using sophisticated general purpose and specialized CFD software as well as stand-alone hard-coded programs. The models describe transport mechanisms and chemical reactions in the reactor including heat transfer, fluid flow, species transport, and deposition kinetics. Experimental work has been done to determine growth parameters and validate lamp heating models.

A comparison of methods for model reduction via linear transformation and truncation has been performed. Low order models derived using both the POD and balancing methods appear to provide an accurate approximation to the original model. The input-output behavior of the heat transfer system can be reconstructed to a high degree of correctness with models of dimension 4 and higher. The wafer heat transfer model's lack of "interesting" behavior facilitates the model reduction, and causes there to be little difference between approximation properties of reduced models derived from the POD and balancing approaches in this case. Preliminary results in applying model reduction techniques to other mechanisms such as transport of energy in the gas phase and transport of silane species suggest that low-order models can be developed to accurately predict their time evolution.

Efforts toward incorporation of more complicated mechanisms, integration of the various models, and model reduction will lead to the development of useful tools for improvement of manufacturing effectiveness and product quality.

⁴The Mathworks, Inc., Natick, MA

⁵Integrated Systems, Inc., Sunnyvale, CA

Appendices

A Physical Constants

Listed here are the physical constants used in the models. The units have been selected for convenience and consistency. Properties of the wafer are those of pure silicon. Chamber wall properties are those of quartz. Properties of the process gases are those of hydrogen at 1000 K and 1 ATM. Chemical kinetics parameters are those experimentally determined from reactions involving thermally activated deposition of polysilicon from 30 sccm of 2% silane in hydrogen.

Constant	Description	Value	Units
k_0	Arrhenius Coefficient	3.0787×10^3	cm sec ⁻¹
E_a	Activation Energy	1.6330×10^5	J mol ⁻¹
R_g	Gas Constant	8.314	J mol ⁻¹ K ⁻¹
h_{ref}	Reference Thickness	1.0×10^{-4}	cm
β_r	Rate Pre-Exponential Constant	1.8472×10^9	dimensionless
β_e	Rate Exponential Constant	2.8059×10^1	dimensionless
k_w	Thermal Conductivity of Wafer	0.22	W cm ⁻¹ K ⁻¹
ρ_w	Mass Density of Wafer	2.3	g cm ⁻³
C_{pw}	Heat Capacity of Wafer	2.3	J g ⁻¹ K ⁻¹
σ_b	Boltzmann Constant	5.677×10^{-12}	W cm ⁻² K ⁻⁴
ϵ_w	Emissivity of Wafer	0.7	dimensionless
α_w	Absorptivity of Wafer	0.5	dimensionless
R_w	Radius of Wafer	7.62	cm
Δ_z	Thickness of Wafer	0.05	cm
h_v	Convective Heat Transfer Coefficient	2.6474×10^{-4}	W cm ⁻² K ⁻¹
Re	Reynolds Number of Gas Flow	27.2	dimensionless
k_g	Thermal Conductivity of Gas	4.40×10^{-3}	W cm ⁻¹ K ⁻¹
Pr	Prandtl Number of Gas Flow	0.686	dimensionless
L	Chamber Length	50.8	cm
T_c	Chamber Wall (Ambient) Temperature	700	K
ϵ_c	Emissivity of Chamber Wall	0.37	dimensionless
T_g	Gas Temperature	300	K
τ	Reference Time	60	seconds
Q_{ref}	Reference Heat Flux	29.24	W cm ⁻²

B Dependent Variables

Variable	Parameters	Description	Units	Dimensionless Conversion
T_w	t, r	Wafer Temperature Field	K	T_w / T_{amb}
X_{SiH_4}	r	Silane Mole Fraction	dimensionless	X_{SiH_4}
C_{SiH_4}	r	Silane Concentration	mol cm ⁻³	C_{SiH_4} / C_{tot}
P_{SiH_4}	r	Silane Partial Pressure	torr	P_{SiH_4} / P_{tot}
R_{Si}	t, r	Silicon Deposition Rate	cm sec ⁻¹	
h	t, r	Film Thickness	cm	h / h_{ref}
Q_i	r	Lamp i Heat Flux Profile	W cm ⁻²	Q_i / Q_{ref}
u_i	t	Lamp i Power Setting	dimensionless	

C Independent Parameters

Parameter	Description	Units	Domain	Dimensionless Conversion
t	Time	seconds	$[0, \infty)$	t / τ
r	Radial Position	cm	$[0, R_w]$	r / R_w
θ	Azimuthal Position (Angle)	radians	$[0, 2\pi]$	$\theta / 2\pi$
z	Axial Position (Height)	cm	$[0, \Delta_z]$	z / Δ_z

D Balancing For Linear Systems

Consider the n th-order stable linear system with minimal state space realization (A, B, C, D) of transfer function $G(s) = D + C(s\mathbf{I} - A)^{-1}B$ and associated controllability and observability Gramian matrices, W_c and W_o respectively, determined by the Lyapunov equations

$$W_c A^T + A W_c + B B^T = 0 \quad (71)$$

and

$$W_o A + A^T W_o + C^T C = 0. \quad (72)$$

It is known [33, 12] that there exists an invertible linear transformation of the state space $T_{bal} \in \mathbb{R}^{n \times n}$ such that the transformed realization

$$(A_{bal}, B_{bal}, C_{bal}, D_{bal}) = (T_{bal}^{-1} A T_{bal}, T_{bal}^{-1} B, C T_{bal}, D)$$

has controllability and observability Gramian matrices of the form

$$W_{c-bal} = T_{bal}^{-1} W_c (T_{bal}^{-1})^T = \text{diag}(\Sigma_1, \Sigma_2, 0, 0) \quad (73)$$

$$W_{o-bal} = T_{bal}^T W_o T_{bal} = \text{diag}(\Sigma_1, 0, \Sigma_3, 0) \quad (74)$$

where $\Sigma_1, \Sigma_2, \Sigma_3$ are positive definite diagonal matrices given by

$$\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_r)$$

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$$

and the σ_i are determined by

$$\sigma_i = (\lambda_i(W_c W_o))^{1/2}.$$

Such a transformation is called a *balancing transformation* and the resulting state space realization is called a *balanced realization*. When transformed to the balanced realization, the input-output influence of each state component relative to the others is indicated by the corresponding singular value σ_i . Thus, the realization can be partitioned, and the $n - k$ least reachable-observable states truncated, yielding the k th-order transfer function $\hat{G}(s)$. It has been shown that the reduced model in stable, minimal, and balanced, i.e., both controllability and observability Gramian matrices are equal to

$$\hat{\Sigma}_{bal} = \text{diag}(\sigma_1, \dots, \sigma_k).$$

It has also been shown that the reduced model enjoys the error bound

$$\bar{\sigma}(G(j\omega) - \hat{G}(j\omega)) \leq 2 \sum_{i=k+1}^r \sigma_i \quad \forall \omega. \quad (75)$$

References

- [1] R. A. Adomaitis. Rapid thermal process model reduction via empirical eigenfunctions: A collocation approach. In *AIChE Annual Meeting, Miami FL*, 1995.
- [2] R. A. Adomaitis. RTCVD model reduction: A collocation on empirical eigenfunctions approach. Technical Report T.R. 95-64, Institute for Systems Research, 1995.
- [3] H. Aling, Suman Banerjee, Anil K. Bangia, Vernon Cole, Jon Ebert, Abbas Emami-Naeini, Klavs F. Jensen, Ioannis G. Kevrekidis, and Stanislav Shvartsman. Nonlinear model reduction for simulation and control of rapid thermal processing. In *Proceedings of the American Control Conference*, pages 2233–2238, 1997.
- [4] H. Aling, J.L. Ebert, A. Emami-Naeini, and R.L. Kosut. Application of a nonlinear model reduction method to rapid thermal processing reactors. In *Proceedings of the IFAC World Congress*, 1996.
- [5] ASM, Inc., Phoenix, AZ. *ASM Epitaxy Epsilon-One Reactor Manual*, 1996.
- [6] Paul Brabant. Private communication. Northrop Grumman ESSD, Linthicum, MD, May 1997.
- [7] Burton, Cabrera, and Frank. The growth of crystals and the equilibrium structure of their surfaces. *Philosophical Transactions of the Royal Society Series A*, 243:299–358, 1951.
- [8] Y.M. Cho and T. Kailath. Model identification in RTP systems. *IEEE Transactions on Semiconductor Manufacturing*, 6(3):233–245, August 1993.
- [9] J.D. Cressler. Re-engineering silicon: Si-Ge heterojunction bipolar technology. *IEEE Spectrum*, 32(3):49–55, March 1995.
- [10] Jean-Marie Dilhac, Nicolas Nolhier, Christian Ganibal, and Christine Zanchi. Thermal modeling of a wafer in a rapid thermal processor. *IEEE Transactions on Semiconductor Manufacturing*, 8(4):432–439, November 1995.
- [11] Fluent, Inc., Lebanon, NH. *Fluent User's Guide*, 4.3 edition, 1995.
- [12] Keith Glover. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds. *International Journal of Control*, 39(6):1115–1193, 1984.
- [13] Leonardo Golubovic and Robijn Bruinsma. Surface diffusion and fluctuations of growing interfaces. *Physical Review Letters*, 66(3):321–324, January 1991.
- [14] Michael D. Graham and Ioannis G. Kevrekidis. Alternative approaches to the Karhunen-Loeve decomposition for model reduction and data analysis. *Comp. Chem. Engr.*, 20(5):495–506, 1996.
- [15] D.W. Greve. Growth of epitaxial germanium-silicon heterostructures by chemical vapour deposition. *Materials Science and Engineering B*, 18(1):22–51, February 1993.
- [16] R.S. Gyurcsik, T.J. Riley, and F. Y. Sorrell. A model for rapid thermal processing: Achieving uniformity through lamp control. *IEEE Transactions on Semiconductor Manufacturing*, 4(1):9–13, February 1991.
- [17] D. Hara and et. al. SiGe HBT technology: Device and application issues. In *Proceedings of the 1995 IEDM, Washington, DC*, pages 731–734, 1995.
- [18] Philip Holmes, John L. Lumley, and Gal Berkooz. *Turbulence, Coherent Structures, Dynamical Systems, and Symmetry*. Cambridge University Press, 1996.
- [19] Philip Holmes, John L. Lumley, and Gal Berkooz. *Turbulence, Coherent Structures, Dynamical Systems, and Symmetry*. Cambridge University Press, 1996.
- [20] Richard C. Jaeger. *Introduction to Microelectronic Fabrication*, chapter 6. Addison-Wesley, 1993.

- [21] T.I. Kamins. Pattern sensitivity of selective $Si_{1-x}Ge_x$ chemical vapor deposition: Pressure dependence. *J. Appl. Phys.*, 74(9):5799–5802, November 1993.
- [22] T.I. Kamins and D.J. Meyer. Effect of silicon source gas on silicon-germanium chemical vapor deposition kinetics at atmospheric pressure. *Applied Physics Letters*, 61(1):90, July 1992.
- [23] T.I. Kamins, D.W. Vook, P.K. Yu, and J.E. Turner. Kinetics of selective epitaxial deposition of $Si_{1-x}Ge_x$. *Appl. Phys. Lett.*, 61(6):669–671, August 1992.
- [24] W. J. Kiether, M. J. Fordham, Seungil Yu, A. J Silva Neto, K.A. Conrad, J. R. Hauser, F.Y. Sorrell, and J.J. Wortman. Three-zone rapid thermal processor system. In *Proceedings of the 2nd International Rapid Thermal Processing Conference, RTP94, Monterey, CA*, pages 96–101, 1994.
- [25] Jin-Won Kim, Myung-Kwan Ryu, Ki-Bum Kim, and Sang-Joo Kim. Low pressure chemical vapor deposition of $Si_{1-x}Ge_x$ films using Si_2H_6 and GeH_4 source gases. *Journal of the Electrochemical Society*, 143(1):363–367, January 1996.
- [26] C.R. Kleijn. Chemical vapor deposition processes. In M. Meyyappan, editor, *Computational Modeling in Semiconductor Processing*, chapter 4, pages 97–229. Artech House, 1995.
- [27] Max G. Lagally. Atom motion on surfaces. *Physics Today*, 46(11):24–31, November 1993.
- [28] H. A. Lord. Thermal and stress analysis of semiconductor wafers in a rapid thermal processing oven. *IEEE Transactions on Semiconductor Manufacturing*, 1(3):105–114, August 1988.
- [29] Doug Meyer. Private communication. ASM, Inc., March 1997.
- [30] Bernard S. Meyerson, Franz J. Himpsel, and Kevin J. Uram. Bistable conditions for low-temperature silicon epitaxy. *Applied Physics Letters*, 57(10):1034–1036, September 1990.
- [31] Bernard S. Meyerson, Kevin J. Uram, and Francoise K. LeGoues. Cooperative growth phenomena in silicon-germanium low temperature epitaxy. *Applied Physics Letters*, 53(25):2555–2557, December 1988.
- [32] B.S. Meyerson. High speed silicon germanium electronics. *Scientific American*, pages 62–67, March 1994.
- [33] Bruce C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, February 1981.
- [34] Andrew Newman, P.S. Krishnaprasad, Sam Ponczak, and Paul Brabant. Modeling and model reduction for epitaxial growth. In *Proceedings of the SEMATECH AEC/APC Workshop IX. SEMATECH*, September 1997.
- [35] Andrew J. Newman. Model reduction via the Karhunen-Loeve expansion Part I: An exposition. Technical Report T.R. 96-32, Inst. Systems Research, April 1996.
- [36] Andrew J. Newman. Model reduction via the Karhunen-Loeve expansion Part II: Some elementary examples. Technical Report T.R. 96-33, Institute for Systems Research, April 1996.
- [37] M. Necati Ozisik. *Heat Transer: A Basic Approach*. McGraw-Hill, 1985.
- [38] Sam Ponczak, Michael O’Loughlin, and Paul Brabant. Private communication. Northrop Grumman ESSD, Linthicum, MD, December 1996.
- [39] W. R. Runyan and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology*, chapter 7, pages 294–360. Addison-Wesley, 1990.
- [40] M. G. Safonov and R. Y. Chiang. A Schur method for balanced truncation model reduction. *IEEE Trans. Automatic Control*, 34(7):729–733, July 1989.
- [41] C.D. Schaper, Y.M. Cho, and T. Kailath. Low-order modeling and dynamic characterization of rapid thermal processing. *Applied Physics A*, 54:317–326, 1992.

- [42] Charles Schaper and Thomas Kailath. Thermal model validation for RTCVD of polysilicon. *Journal of the Electrochemical Society*, 143(1):374–381, January 1996.
- [43] J.M.A. Scherpen. Balancing for nonlinear systems. *Sys. Cont. Lett.*, 21:143–153, 1993.
- [44] T.O. Sedgwick and D.A. Grutzmacher. Low temperature atmospheric pressure chemical vapor deposition for epitaxial growth of Si–Ge bipolar transistors. *J. Electrochemical Society*, 142(7):2458–2463, July 1995.
- [45] Robert Siegel and John R. Howell. *Thermal Radiation Heat Transfer*. Hemisphere, 1992.
- [46] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, March 1987.
- [47] Lawrence Sirovich. Turbulence and the dynamics of coherent structures; Part I: Coherent structures. *Quarterly Appl. Math.*, 45(3):561–571, October 1987.
- [48] D. Brian Spalding, editor. *The PHOENICS Journal of Computational Fluid Dynamics and Its Applications*, volume 8. CHAM Ltd, December 1995.
- [49] James C. Sturm. RTCVD growth and applications of epitaxial $Si_{1-x}Ge_x$ alloys. *Journal of Metals*, pages 44–47, 1991.
- [50] Artemis Theodoropoulou, Raymond A. Adomaitis, and Evangelos Zafiriou. Model reduction for optimization of RTCVD systems. Technical Report T.R. 96-64r1, Institute for Systems Research, November 1996.
- [51] Pei-Jih Wang. Epitaxy. In C. Y. Chang and S. M. Sze, editors, *ULSI Technology*, chapter 3, pages 105–143. McGraw-Hill, 1996.