NONLINEAR FEEDBACK CONTROL OF A COUPLED KINETIC MONTE CARLO-FINITE DIFFERENCE CODE

**Effendi Rusli, Timothy O. Drews, David L. Ma,
Richard C. Alkire, Richard D. Braatz**

*University of Illinois at Urbana-Champaign
Department of Chemical and Biomolecular Engineering
600 South Mathews Avenue, Urbana, IL 61801*

Abstract: Product quality variables for many electronics and materials processes are set at the nanoscale and smaller length scales. Although the control of these processes is of scientific and industrial interest, there is a shortage of feedback controller design methods based on the noncontinuum models that describe such nanoscopic phenomena. In this study, linear, gain-scheduled, and nonlinear feedback controllers are designed for a coupled kinetic Monte Carlo-finite difference code that simulates the manufacture of copper interconnects. The feedback controller designs incorporate a low order stochastic model constructed from the coupled continuum-noncontinuum code.

Keywords: Stochastic simulation, noncontinuum models, kinetic Monte Carlo simulation, nonlinear control, gain-scheduled control, stochastic control, Markov processes

1. INTRODUCTION

The vast majority of the literature on feedback controller design is based on continuum models, which are described by systems of algebraic, ordinary differential, and partial differential equations (Levine, 1995). The continuum modeling approach, however, is inadequate for modeling much of the molecular and mesoscale phenomena that occur in the complex chemical processes that constitute the attention of today's scientists and engineers (Maroudas, 2000). This is especially apparent in microelectronics processes, for which the critical phenomena occur at the nanometer and smaller length scales. Hence in recent years increasing efforts have been directed towards the development of noncontinuum models, such as kinetic Monte Carlo (KMC) simulation models, for which most existing controller design techniques are not directly applicable. The design of feedback controllers based on such noncontinuum models is an open research problem in the field of control (Murray, 2002).

Global competition has increased the importance of feedback control for the complex chemical processes that are best described by noncontinuum models. There is probably no place where this is more apparent than in the microelectronics industry, which has had an average annual growth of 20%, with sales of $200 billion in 2001. It is generally accepted that high performance feedback control will be required to achieve the small length scales required to provide high computational speed in future microelectronic devices (Sematech, 2001).

Here feedback controllers are designed for a coupled KMC-finite difference (FD) code that simulates the electrochemical deposition of copper into a trench, a key step in the manufacturing on-chip interconnects for microelectronic devices (Andricacos, *et al.*, 1998). The industrial need is to deposit copper uniformly into trenches and vias of small dimension (less than 100 nm) under galvanostatic (constant current) conditions. This industrial importance has motivated numerous experimental and simulation studies on the modeling of copper electrodeposition

in recent years (Alkire and Eliadis, 1999; Andricacos, *et al.*, 1998; Georgiadou, *et al.*, 2001; Gill, 2001; Harper, *et al.*, 1999; Merchant, *et al.*, 2000; Moffet, *et al.*, 2000, Moffet, *et al.*, 2001). The goal of the feedback controller is to maintain the current (or current density) at a constant specified value. This feedback controller allows the KMC simulations to operate under industrial operating conditions.

The paper is organized as follows. First, the coupled KMC-FD copper electrodeposition simulation code is described. This is followed by construction of a low order stochastic model that is used to design feedback controllers and associated filters to handle the non-Gaussian stochastic noise produced by the KMC code. Then the closed-loop responses of the controllers are compared in simulations of the low order stochastic model and the KMC-FD simulation code.

2. COUPLED KINETIC MONTE CARLO-FINITE DIFFERENCE SIMULATION CODE

Kinetic Monte Carlo (KMC) methods are used to simulate structural properties of matter that cannot be represented by a macroscopic continuum description, and are widely used for simulating dynamic chemical and materials processes. A KMC simulation is a realization of the Master equation (Fichthorn and Weinberg, 1991):

$$\frac{\partial P(\sigma,t)}{\partial t} = \sum_{\sigma'} W(\sigma',\sigma)P(\sigma,t) - \sum_{\sigma'} W(\sigma,\sigma')P(\sigma,t) \tag{1}$$

where σ and σ' are successive states of the system, $P(\sigma, t)$ is the probability that the system is in state σ at time t, and $W(\sigma',\sigma)$ is the probability per unit time that the system will undergo a transition from state σ' to σ. For a particular system being studied, the KMC code chooses randomly among the possible transitions of the system and accepts particular transitions with appropriate probabilities. After each accepted or attempted transition, the time variable is incremented by one Monte Carlo time step, and the process is repeated. If the probabilities satisfy certain conditions, the real time variable t corresponding to the number of Monte Carlo time steps can be computed.

Electrochemical deposition of a copper film into a trench is simulated in this application. A KMC method was used since traditional continuum codes are not convenient for simulating the evolution of the roughness of the surface, which is an important characteristic of the produced copper film. The KMC code describes the mesoscale with a cubic lattice, where each subdomain in the simulation space represents a cluster of molecules (referred to as a mesoparticle) of a given species in the deposition bath (see Fig. 1). Each subdomain is cube of 12.5 nm on a side and is assumed to be homogeneous in both phase and composition. Similar mesoscale KMC methods have been applied by various researchers to a number of systems (Bird, 1994; Birdsall and, Langdon, 1985; Katsoulakis, *et al.*, 2002; Lu and Kushner, 2001). While molecular-scale simulations are of interest, this coarser mesoscale representation results in an efficient computational method that can simulate devices on the same scale as in the real system (Drews, *et al.*, 2003). The Monte Carlo simulation domain is a trench with aspect ratio 2:1, 40 subdomains wide, 80 subdomains high, and 6 subdomains deep.

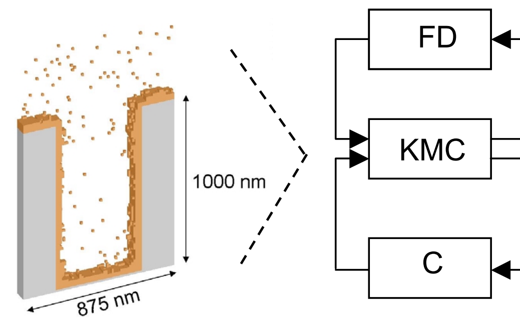


Fig. 1. Architecture of the KMC-FD simulation operating under feedback: FD denotes the finite difference code, KMC denotes the kinetic Monte Carlo code, and C denotes the controller. The KMC domain is on the left.

The kinetic Monte Carlo code simulates deposition phenomena by considering the likelihood of various actions that each mesoparticle can take at a given time step. These actions are bulk diffusion, surface diffusion, the reaction A→B, a combination reaction A+B→C, a splitting reaction A→B+C, and dissolution. All actions are computed as frequencies, with units of sec^{-1}. At a given Monte Carlo time step, a mesoparticle can make a maximum of one move. The possible moves that each species can make are a function of the location of the mesoparticle in the simulation space, as well as the number and type of the six nearest neighbors.

The Monte Carlo domain has periodic boundary conditions in the x and y directions, an impenetrable boundary at the electrode surface (in the z-direction), and a link to a continuum code at the top boundary in the z-direction. The continuum code is a one-dimensional FD code that provides diffusion fluxes of Cu^{2+} into the Monte Carlo domain by solving the diffusion equation. The KMC code provides the concentration of Cu^{2+} to the continuum code. The height of the continuum domain was set to 50 μm, which is close to the actual diffusion boundary layer

thickness that corresponds to typical processing conditions. In both the FD and KMC codes, an additive-free bath is simulated. The KMC code also produces a signal that is the charge passed during deposition, and reads as input the applied potential η. These signals serve as the input and output of the feedback controller (see Fig. 1).

Three time steps are tracked in the KMC simulation code: (1) the time step over which the continuum code is called for updated flux information, (2) the sampling interval for the feedback controller, and (3) the Monte Carlo (MC) time step. In order to capture the full dynamics of the system, the MC time step must be small enough to capture the action of the fastest species. For all the processes in this application, the Monte Carlo time step was computed to be ~2.8 μs. A complete KMC simulation run typically requires 1.08×10^8 MC time steps before the copper fills the trench. In this particular study, the linking time step and the sampling interval for the feedback controller are set to be 10^{-7} s and 10^{-2} s, respectively.

To carry out the galvanostatic (i.e., constant current) simulations associated with industrial operations, the feedback controller must manipulate the applied potential η to control the current i, based on the charge transferred as a function of time. There are two main performance requirements for the feedback controller. First, the feedback controller should have a tracking response as fast as possible. Second, 90% of the fluctuations in the applied potential should be within ±0.01 V. An additional requirement is for the controller to be low order, so that its computational cost is negligible compared to the cost of the KMC-FD calculations. The potential η enters the surface reaction frequencies in a nonlinear manner. This suggests that nonlinear control may give better performance than linear control. The next section describes how a low order stochastic model was constructed from input-output data collected from the KMC-FD code, and how this model was used to design feedback controllers.

3. IDENTIFICATION OF A LOW ORDER STOCHASTIC MODEL

The KMC-FD code is computationally expensive, highly stochastic, and nonlinear. To design low order feedback controllers, a low order stochastic model is constructed that is capable of capturing the most essential input-output behavior of the coupled KMC-FD code. This low order model is incorporated into model-based controller design and used for filter and controller tuning.

The output of the KMC-FD code is the cumulative charge passed up to current simulation time. To emulate the real physical system as closely as possible, the charge signal is converted to a current density signal. The current density was computed as the total charge passed in each 0.01 s, divided by 0.01 s and the surface area in cm^2. A larger time step interval could be used to compute the current from the charge, but this would lead to a more sluggish response, causing an inherent performance limitation in the feedback controller. On the other hand, decreasing the time step leads to more highly noise-corrupted signal. The manipulated variable is the applied potential, which affects the kinetics of the mechanisms simulated in the KMC-FD code and hence directly affects the current generation.

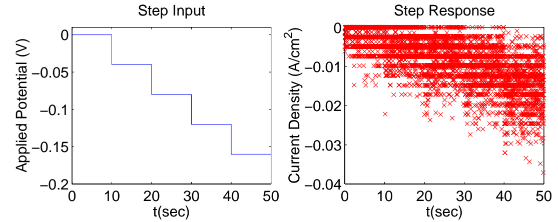


Fig. 2. Step input implemented on the KMC-FD code and the resulting step response.

The current density for a series of steps in the applied potential sent to the KMC-FD code is reported in Fig. 2. The applied potentials are selected to be within the normal operating condition of the KMC-FD simulation. Autocorrelations indicate that the current density reaches steady state within one sampling instance.

Upon reaching steady state, the output signal is bounded and its mean remains constant. These conditions justify the assumption that the signal is quasi-stationary (see Fig. 3). This assumption is verified by comparing the probability mass function of different time segments. The stochastic fluctuations are non-Gaussian and asymmetric, and can be modeled by a Poisson distribution for all normal operating conditions. To ensure consistency and accuracy, the identification procedure was repeated with different seed numbers. These sets of input-output data were used in the parameter estimation of a low order stochastic model:

$$P\big(i(k)=\kappa \,\big|\, \eta(k-1)\big)=\frac{\lambda^{-400\kappa}\exp(-\lambda)}{(-400\kappa)!} \qquad (2)$$

$$\lambda = 2.5285\exp(-6.5962\,\eta(k-1)) - 1.3622 \qquad (3)$$

where $\kappa \in \{-0.0025n, n \in Z\}$ and Z is a set of non-negative integers. The form of the nonlinearity was motivated by the expression for the surface kinetics. Figure 3 compares the stochastic current density produced by the low order model (2)-(3) and the KMC-FD code for a range of applied potential.

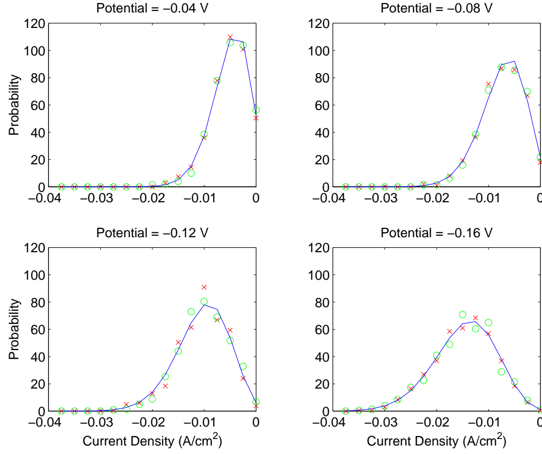


Fig. 3. Current density distributions for the low order model (solid line) and the KMC-FD code (× and o correspond to simulation data with different seed numbers).

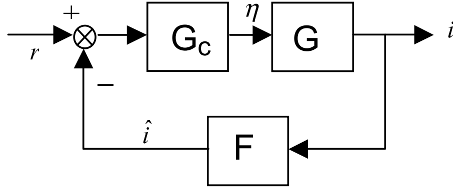


Fig. 4. Block diagram for the closed-loop system

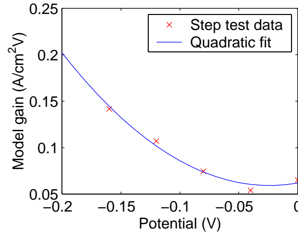


Fig. 5. Model gain of the KMC-FD code computed from step data.

4. FEEDBACK CONTROLLER DESIGN

Linear, gain-scheduled, and nonlinear inversion-based controllers were designed based on the low order stochastic model (2)-(3). Each controller incorporates a first-order filter (see Fig. 4):

$$F(z) = \frac{\alpha}{1-(1-\alpha)z^{-1}} \quad (4)$$

with filter constant α. This filter is used to reduce fluctuations in the manipulated variable without filtering the reference signal. The linear and nonlinear feedback controllers incorporate the deterministic part of the low order model (2)-(3):

$$i(k) = -6.3213\times10^{-3}\exp(-6.5962\,\eta(k-1)) + 3.4055\times10^{-3} \quad (5)$$

An alternative deterministic model used by the gain-schedule controller is to directly compute the model gain as a function of the manipulated variable (see Fig. 5). The model gain is computed based on the initial steady state condition at zero applied potential. The best least-squares quadratic fit to the model gain is:

$$K = 4.6058\eta^2 + 2.2074\times10^{-1}\eta + 6.1912\times10^{-2} \quad (6)$$

The two plant descriptions give almost the same output prediction.

4.1 Linear Controller

The range in system gain is given by

$$\{K \mid K\in\Re,\ 0.05\le K\le 0.1417\} \quad (7)$$

where the upper bound was selected to exceed slightly the steady-state value for regulating the current density at –0.015 A/cm^2.

The linear feedback controller was designed using internal model control. Many other controller design techniques such as generic model control, direct synthesis, and geometric control give the same or similar control structures. The desired closed-loop response is first-order-plus-time-delay:

$$\frac{i}{r} = GG_c(1+FGG_c)^{-1} = \frac{(1-\exp(-\Delta t/\tau))z^{-1}}{1-\exp(-\Delta t/\tau)z^{-1}} \quad (8)$$

where τ is the desired closed loop time constant. This equation is rearranged to give the feedback controller

$$G_c = \frac{(1-\phi)-(1-\alpha)(1-\phi)z^{-1}}{1-(1+\phi(1-\alpha))z^{-1}+\phi(1-\alpha)z^{-2}}\cdot\frac{1}{K} \quad (9)$$

where $\phi = \exp(-\Delta t/\tau)$. Applying the small gain theorem to systems with time varying perturbations (Braatz and Morari, 1997) shows that choosing K = 0.1417 in the linear controller provides robustness for the full range of time-varying model gains in (7). The value τ = 10^{-5} s ensures fast response yet not faster than the dynamics of the KMC-FD simulation which is on the order of 10^{-6} s. The tuning of the filter constant α is discussed in Section 4.4.

4.2 Gain-scheduled Controller

The structure of the gain-scheduled controller is identical to the linear feedback controller. The only difference is that the gain K in (9) is updated at every time step using (6).

4.3 Nonlinear Controller

The nonlinear controller inserts an inverter derived from (5):

$$\eta(k) = -\frac{1}{6.5962}\ln\left(\frac{i(k) - 3.4055\times10^{-3}}{-6.3213\times10^{-3}}\right) \quad (10)$$

before the plant in the block diagram in Fig. 4. The plant combined with the inverter is a simple one-delay system that is controlled using the linear feedback controller with $K = 1$ in (9).

4.4 Filter Design

The filter constant α is tuned to ensure that at least 90% of the fluctuations in the applied potential are within ±0.01 V over the entire operating regime, while avoiding too much filtering which leads to unnecessarily sluggish response. The filter constant is designed based on the probability density distribution of the applied potential at the final time, that is, the time required to fill up the trench with copper. The reason for using the final time to design the filter coefficient is that the applied potential is the most negative at the final time, and the stochastic fluctuations are largest when the applied potential is the most negative. A filter coefficient that adequately filters the stochastic fluctuations at the final time also provides adequate filtering at earlier times.

A primary goal of this study was to create a filter and controller design procedure that can be quickly repeated when physicochemical parameters in the KMC-FD code are changed. Due to the high computational cost of running the KMC-FD code, its use in filter and controller design is limited to the creation of data for constructing the low order stochastic model (2)-(3). The low order model is then used to design the filter and controller. The probability density distribution of the applied potential at the final time was obtained by running the closed-loop simulation of the low order stochastic model 10,000 times at several α values. From this probability density distribution, the mean and the deviation corresponding to the 90% confidence level were estimated. Figure 6 shows how the deviation varies with the filter constant α. Table 1 reports the filter constants that result in 90% of the applied potential being within ±0.01 V at the final time.

Table 1. Filter constants for the three controllers

Controller type	α
Linear	0.03806
Gain-scheduled	0.03951
Nonlinear	0.03260

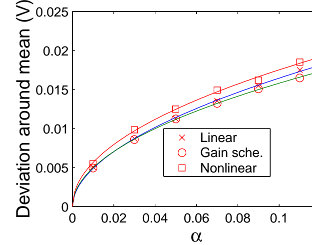


Fig. 6. The relationship between the filter constant α and the deviation around the mean at the final time corresponding to the 90% confidence level

5 RESULTS AND DISCUSSION

The controllers were implemented in the KMC-FD code and the low order stochastic model (2)-(3). Figure 7 shows agreement between the closed-loop predictions of the original and low order models. As specified, the applied potential is within ±0.01 V of its steady-state value 90% of the time, except for the initial transient.

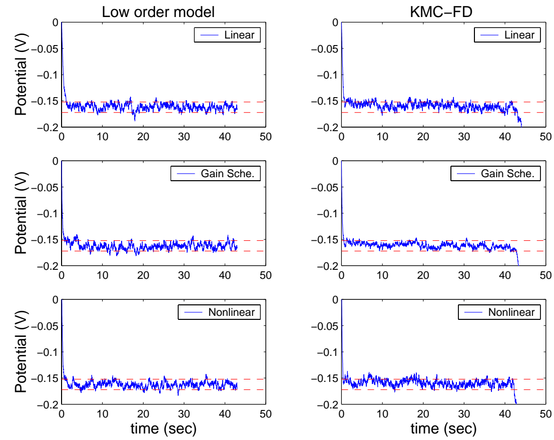


Fig. 7. Closed-loop responses for the linear, gain-scheduled, and nonlinear inversion-based controllers implemented on the low order model (2)-(3) and the KMC-FD code

Figures 8 and 9 focus on the initial time responses. The closed-loop performance is similar for the controllers, with the gain-scheduled controller slightly better than the others. Differences between the closed-loop simulations obtained with the low order stochastic model (2)-(3) and the KMC-FD code are within the stochastic variation in the responses. This is further support that use of the low order model for filter and controller design was justified. The applied potential in Fig. 8 reaches a quasi-steady-state value in ~0.5 s. Since the process dynamics are very fast, the unfiltered current density (not shown due to extremely large stochastic noise) reaches a quasi-state-value in ~0.5 s. The filtered current density, which includes the filter lag, reaches a quasi-steady-state value in 1 s.

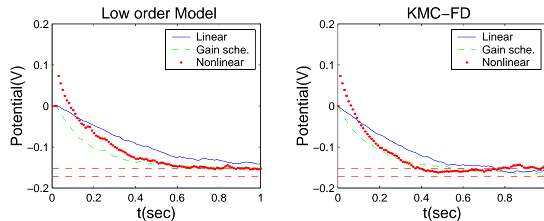


Fig. 8. The applied potentials for the three controllers implemented on the low order model (2)-(3) and the CED code

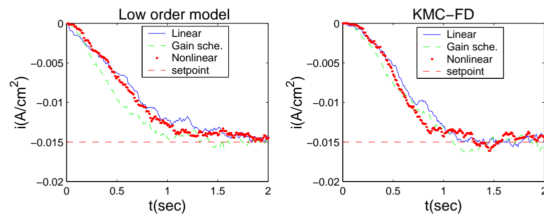


Fig. 9. The filtered current density for the three controllers implemented on the low order model (2)-(3) and the CED code

6. CONCLUSIONS

This paper demonstrates the design of low order linear, nonlinear, and gain-scheduled feedback controllers for a coupled kinetic Monte Carlo-finite difference code that simulates infill of a trench during copper electrodeposition. The feedback controllers and associated filters were constructed from a low order stochastic model constructed from data collected from the KMC-FD code. The controllers enable the KMC-FD code to operate with nearly constant current, which is the industrial operating condition.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from the National Science Foundation (CTS-0135621, NRAC-MCA01S022). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

Alkire, R.C., and E.D. Eliadis (1999). Electro-deposition of copper: The effect of various organic compounds. *Z. Phys. Chem.*, **208**, 1-15.

Andricacos, P.C., C. Uzoh, J.O. Dukovic, J. Horkins, and H. Deligianni (1998). Damascene copper electroplating for chip interconnections. *IBM J. Res. Dev.*, **42**, 567-574.

Bird, G.A. (1994). *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Clarendon Press, Oxford.

Birdsall, C.K. and A.B. Langdon (1985). *Plasma Physics via Computer Simulation*. McGraw-Hill, New York.

Braatz, R.D. and M. Morari (1997). A multivariable stability margin for systems with mixed time-varying parameters. *Int. J. of Robust and Nonlinear Control*, **7**, 105-112.

Drews, T.O., R.D. Braatz, and R.C. Alkire (2003). Parameter sensitivity analysis of Monte Carlo simulations of copper electrodeposition with multiple additives. *J. Electrochem. Soc.* accepted.

Fichthorn, K.A. and W.H. Weinberg (1991). Theoretical foundations of dynamic Monte-Carlo simulations. *J. Chem. Phys.*, **95**, 1090-1096.

Georgiadou, M., D. Veyret, R.L. Sani, and R.C. Alkire (2001). Simulation of shape evolution during electrodeposition of copper in the presence of additive. *J. Electrochem. Soc.*, **148**, C54-C58.

Gill, W.N., D.J. Duquette, and D. Varadarajan (2001). Mass transfer models for the electrodeposition of copper with a buffering agent. *J. Electrochem. Soc.*, **148**, C289-C296.

Harper, J.M.E., C. Cabral, P.C. Andricacos, L. Gignac, I.C. Noyan, K.P. Rodbell, and C.K. Hu (1999). Mechanisms for microstructure evolution in electroplated copper thin films near room temperature. *J. Appl. Phys.*, **86**, 2516-2525.

Katsoulakis, M.A., A.J. Majda, and D.G. Vlachos (2002). Course-grained stochastic processes and Monte Carlo simulations in lattice systems. Univ. of Delaware, Newark. Technical report.

Levine, W.S., ed. (1995). *The Control Handbook*. CRC Press, Boca Raton, FL.

Lu, J., and M.J. Kushner. Trench filling by ionized metal physical vapor deposition (2001). *J. Vac. Sci. Technol. A*, **19**, 2652-2663.

Maroudos, D. (2000). Multiscale modeling of hard materials: Challenges and opportunities for chemical engineering. *AIChE J.*, **46**, 878.

Merchant, T.P., M.K. Gobbert, T.S. Cale, and L.J. Borucki (2000). Multiple scale integrated modeling of deposition processes. *Thin Solid Films*, **365**, 368-375.

Moffet, T.P., J.E. Bonevich, W.H. Huber, A. Stanishevsky, D.R. Kelly, G.R. Stafford, and D. Josell (2000). Superconformal electrodepositon of copper in 500-90 nm features. *J. Electrochem. Soc.*, **147**, 4524-4535.

Moffet, T.P., D. Wheeler, W.H. Huber, and D. Josell (2001). Superconformal electrodeposition of copper. *Electrochem. Solid State*, **4**, C26-C29.

Murray, R. M., ed. (2003). *Control in an Information Rich World*, SIAM Press, Philadelphia, PA.

Sematech (2001). International Technology Roadmap for Semiconductors, Semiconductor Industry Association, http://public.itrs.net/.

OPTIMAL CONTROL OF TRANSIENT ENHANCED DIFFUSION

R. Gunawan, M. Y. L. Jung, E. G. Seebauer, and R. D. Braatz

Department of Chemical and Biomolecular Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

Abstract: Transient enhanced diffusion of boron inhibits the formation of ultrashallow junctions needed in the next-generation of microelectronic devices. Reducing the junction depth using rapid thermal annealing with high heating rates comes at a cost of increasing sheet resistance. The focus of this study is to design the optimal annealing temperature program that gives the minimum junction depth while maintaining satisfactory sheet resistance. Comparison of different parameterizations of the optimal trajectories shows that linear profiles gave the best combination of minimizing junction depth and sheet resistance. Worst-case robustness analysis of the optimal control trajectory motivates improvements in feedback control implementations for these processes. *Copyright © 2003 IFAC*

Keywords: optimal control, model-based control, semiconductors

1. INTRODUCTION

Moore's law requires a continued shinkage of feature sizes in microelectronic devices. For example, advanced CMOS devices will require junction depths between 13 to 22 nm in the source and drain extension region by the year 2005 according to the 2001 International Technology Roadmap for Semiconductors. The current technology for the formation of such ultrashallow junctions depends on ion implantation of dopant, such as boron, into silicon. Although the junction depth can be made shallower by reducing the implant energy, the effectiveness of this approach is limited by the need to anneal out the point and/or extended defects generated by ion implantation. Silicon self-interstitial defects can mediate the diffusion of dopants during the annealing process, which leads to a significant increase of the junction depth. This phenomenon is known as "transient enhanced diffusion" (TED). For this reason, considerable efforts have been put forth in the modeling of the TED for designing appropriate post-implant annealing programs to produce the desired junction depth (see (Jain*, et al.*, 2002) and references therein).

The state-of-the-art in post-implant annealing employs a lamp-based rapid thermal annealing (RTA). Figure 1 shows a typical RTA "spike" anneal program, which consists of a stabilization step at constant temperature (~650 °C), followed by a linear heating step at a constant rate (~100 °C/s) reaching a maximum temperature (~1000 °C), and finally a radiative cooling step at a initial rate of several tens of degrees per second. In the literature, there exists conflicting experimental evidence on the efficacy of using high heating rates (up to 400 ºC/s) in the spike anneal profile to reduce TED (Downey*, et al.*, 1999; Shishiguchi*, et al.*, 1997). Recent results (Gelpey*, et al.*, 2002; Mannino*, et al.*, 2001) tend to confirm the benefit of using high heating rates. The results also suggest that the reduction in the junction depth comes at the expense of an undesired increase in the sheet resistance. The tradeoff in reducing the junction depth without sacrificing the sheet resistance motivates a careful optimization of the post-implant annealing temperature program.

A recently developed comprehensive TED model consists of a set of reaction-diffusion equations combined with Poisson's equation to account for the electric field effects on charged species (Jung, *et al.*,

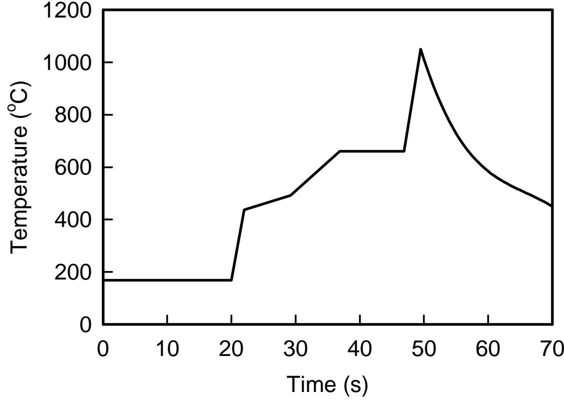


Fig. 1. A typical RTA temperature program.

1999). The activation energies arising from the reaction rate constants and diffusivities in the TED model are not exactly known, but there exist extensive experimental and computational estimates of these parameters. To resolve problems with regard to conflicting estimates in the literature, maximum likelihood (ML) estimation was applied to give the most likely values and the standard deviations for the parameters from the published parameter estimates. Furthermore, maximum *a posteriori* (MAP) estimation was applied to produce improved parameters from the ML estimates and experimental Boron profile data collected at International Sematech (Gunawan*, et al.*, 2003).

This paper focuses on the design of the spike anneal program that optimizes the junction depth subject to a constraint on the sheet resistance. The TED model is implemented using the process simulator FLOOPS (Law and Tasch, 2000). Different parameterizations of the optimal trajectory are used to elucidate the true optimal annealing program. Worst-case analysis of the resulting optimal trajectory quantifies the performance degradation with respect to control implementation inaccuracies and model uncertainties.

2. TRANSIENT ENHANCED DIFFUSION MODEL

Transient enhanced diffusion arises from reaction-diffusion processes consisting of Fickian diffusion, electrical drift motion, and reaction networks including boron activation and interstitial clustering. The model comprises of coupled continuity equations (i.e., mass balances) for each species and Poisson's equation to include the electrical field effect on the charged species. The general continuity equation is

$$\frac{\partial N_i}{\partial t} = -\frac{\partial J_i}{\partial x} + G_i \tag{1}$$

where N_i denotes the concentration, J_i is the flux, and G_i is the net generation rate of species *i*. The flux J_i includes terms from the Fickian diffusion and the electric field drift motion:

$$J_i = -D_i \frac{\partial^2 N_i}{\partial x^2} + \gamma_i \mu_i N_i E(x), \tag{2}$$

where D_i denotes the diffusivity and $E(x)$ is the electric field. The mobility μ_i follows the Einstein relation

$$\mu_i = \frac{qD_i}{kT}, \tag{3}$$

where q is the electron charge, k is the Boltzmann constant, and T is the temperature. The term γ_i describes the average charge of species according to

$$\gamma_i = \sum_j z_j \gamma_{z_j}, \tag{4}$$

where z_j are the possible charge states (*i.e.,* +2, 0, -1, etc.) and γ_{z_j} is the fraction of species *i* having charge z_j according to the Fermi-Dirac statistics.

Poisson's equation describes the electric field induced by the spatial charge imbalance:

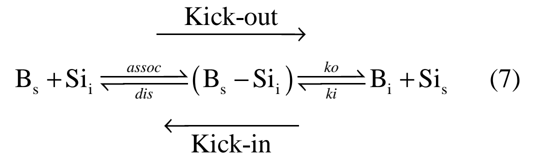
$$\frac{\partial^2 \psi}{\partial x^2} = \frac{Q(x)}{\varepsilon} \tag{5}$$

where ε denotes the dielectric constant and the charge density $Q(x)$ is given by

$$Q(x) = p - n + \sum_i \gamma_i N_i \tag{6}$$

with p and n denoting the hole and electron concentrations, respectively. The concentrations p and n are assumed to be in thermal equilibrium.

The generation term G_i includes the formation and annihilation rates due to the boron activation reaction and/or clusters formation and dissociation. The boron activation reaction provides a pathway between mobile interstitial boron B_i to and from immobile activated boron (*i.e.,* substitutional boron, B_s):



$$B_s + Si_i \underset{dis}{\overset{assoc}{\rightleftharpoons}} (B_s - Si_i) \underset{ki}{\overset{ko}{\rightleftharpoons}} B_i + Si_s \tag{7}$$

(Kick-out: forward; Kick-in: reverse)

In addition, the intermediate (B_s–Si_i) acts as nucleation centers for mixed boron-silicon clusters. The rates of reactions follow reactant-limited rate expressions with the reaction rate constants adhering to the Arrhenius law.

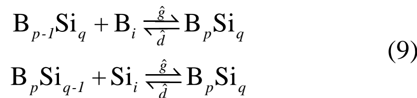
Clusters of interstitial atoms have been shown to form during TED (Collart*, et al.*, 2000; Stolk*, et al.*,

1997). There is evidence supporting the formation of clusters consisting of pure boron (Collart*, et al.*, 2000), pure silicon (Stolk*, et al.*, 1997), and mixed boron-silicon (Haynes*, et al.*, 1996). During thermal annealing, the clusters can act as reservoir (during the stabilization step) and source (during ramp up) for mobile interstitial boron and silicon. The formation and dissolution of pure interstitial clusters follow the reactions

$$I + I \underset{d}{\overset{f}{\rightleftharpoons}} I_2$$
$$I_{m-1} + I \underset{d}{\overset{g}{\rightleftharpoons}} I_m \quad (8)$$

where I denotes the interstitials (boron and silicon) and the indices m denote the sizes of the clusters. The cluster formation rate assumes a reactant diffusion-limited reaction in agreement with much of the literature (see for example (Laidler, 1987)). On the other hand, the dissolution rate follows a first-order kinetic expression with rate constant according to the Arrhenius law.

The formation and dissolution of mixed boron-silicon clusters is described by:



$$B_{p-1}Si_q + B_i \underset{\hat{d}}{\overset{\hat{g}}{\rightleftharpoons}} B_pSi_q$$
$$B_pSi_{q-1} + Si_i \underset{\hat{d}}{\overset{\hat{g}}{\rightleftharpoons}} B_pSi_q \quad (9)$$

where p, q are integers larger than or equal to 1. The formation and dissolution rates of mixed clusters again follow diffusion-limited and first-order kinetics, respectively, as in the pure cluster dynamics.

The TED model requires a set of activation energies associated with the diffusivities and kinetic rate constants for the boron activation reaction and cluster dissociation dynamics. These activation energies are difficult to directly measure experimentally and determine computationally. Experimental and *ab initio* density functional theorem (DFT) estimates of the activation energies are scattered throughout the literature. For many of the activation energies, the published values show significant variation. To resolve problems in regard to conflicting estimates in the literature, maximum likelihood (ML) estimation was applied to determine the most likely values and the standard deviations from the published parameter estimates (Gunawan*, et al.*, 2003).

Maximum *a posteriori* estimation takes a Bayesian approach which combines experimental data with the *a priori* information, in this case, from maximum likelihood estimation of published experimental and/or DFT values (Gunawan*, et al.*, 2003). Figure 2 presents the after-anneal experimental data used in the MAP estimation along with simulation profiles using the MAP parameters employing various RTA programs. Figure 3 shows the agreement between the TED model using the MAP estimates and the experimental observations compiled from the literature (Agarwal*, et al.*, 1999).

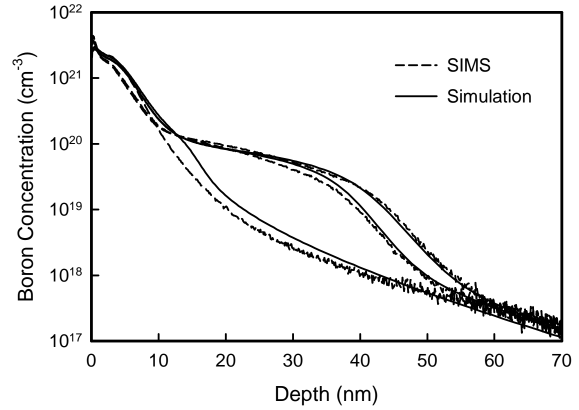


Fig. 2. Comparison of experimental and simulation boron profiles using the TED model with the MAP parameters. The junction depth is defined as the spatial penetration of boron at a total concentration of 10^{18} atoms/cm^3

4. OPTIMAL CONTROL FORMULATION

In the literature, control of transient enhanced diffusion through manipulation of RTA programs adopted an *ad hoc* approach through trial and error (Jain*, et al.*, 2002), due to incomplete understanding of TED mechanisms and correspondingly undependable models for describing dopant diffusion and activation. In contrast, this work employs a model-based control approach for designing an optimal temperature program that minimizes the junction depth while maintaining a suitable sheet resistance. The optimization variable is the RTA temperature profile, in particular, the heating and cooling profiles and the annealing temperature. A conventional RTA only employs a radiative cooling step, but there exists evidence (Agarwal, 2000;

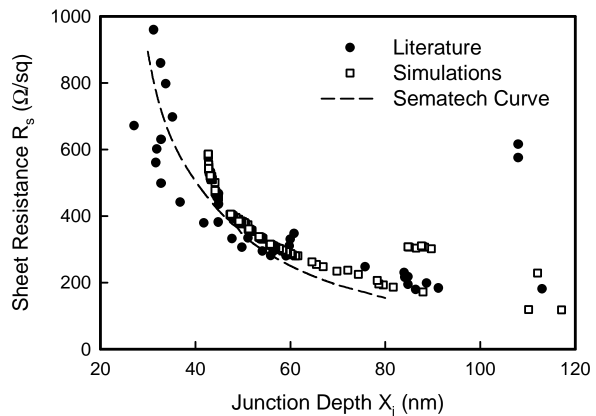


Fig. 3. Comparison of junction depth and sheet resistance data from experimental works employing various annealing programs as summarized in the Sematech curve, and from the TED simulations.

Agarwal, *et al.*, 1999) supporting the importance of the ramp down trajectory, especially in high heating rate applications (>150 ºC/s).

The optimal control objective chosen here is to minimize the junction depth while maintaining a satisfactory sheet resistance, which is equivalent to the minimization problem:

$$\min_{\substack{T(t)\\ R_s \le R_{s,\max}\\ \beta_{\min} \le \frac{dT}{dt} \le \beta_{\max}\\ T \le T_{\max}}} X_j \tag{10}$$

where X_j denotes the junction depth, R_s denotes the sheet resistance, and $T(t)$ is the RTA temperature trajectory. The sheet resistance is given by:

$$R = \frac{1}{q\int \mu(C)C(x)\,dx}\left[\frac{\Omega}{\text{sq}}\right] \tag{11}$$

where q denotes the carrier charge, $\mu(C)$ denotes the mobility (concentration dependent), and $C(x)$ is the spatial concentration of charge carrier (*i.e.,* activated dopant). The following empirical formula gives the mobility $\mu(C)$ for boron (Zeghbroeck, 2002):

$$\mu(C) = 44.9 + \frac{425.6}{1+\left(\frac{C}{2.23\times10^{17}}\right)^{0.719}}\left[\frac{\text{cm}^2}{\text{Vs}}\right] \tag{12}$$

In this work, it is desired to produce junctions with the sheet resistance below $R_{s,\max}$ of 350 Ω/sq. The constraints on the temperature gradient, *i.e.,* $\beta_{\min}$ and $\beta_{\max}$, describe the limits for the cooling and heating rates, respectively. The state-of-the-art lamp-based RTA can produce heating rates up to 400 ºC/s (Shishiguchi, *et al.*, 1997), while recent advances in RTA technology can achieve cooling rates up to 200 ºC/s (Vortek Industries Ltd., 2002). The maximum temperature of thermal anneal $T_{\max}$ is set to the melting point temperature of silicon at 1410 °C.

5. WORST CASE ANALYSIS

Worst case analysis (Ma and Braatz, 2001) provides tools for quantifying the robustness of the optimal control performance to uncertainties in model parameters and control implementation. The information can be used to assess whether a more accurate model and thus more experiments are needed, or to give the desired performance and accuracy of the lower level control loops and control equipment, respectively. The parametric and control uncertainties are described as norm bounded perturbations $\boldsymbol{\delta u}$ and $\boldsymbol{\delta\theta}$, that is,

$$\mathrm{E}_\theta = \left\{\boldsymbol{\theta} : \boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \boldsymbol{\delta\theta},\ \left\|\boldsymbol{W}_\theta \boldsymbol{\delta\theta}\right\| \le 1\right\} \tag{13}$$

$$\mathrm{E}_u = \left\{\boldsymbol{u} : \boldsymbol{u} = \hat{\boldsymbol{u}} + \boldsymbol{\delta u},\ \left\|\boldsymbol{W}_u \boldsymbol{\delta u}\right\| \le 1\right\} \tag{14}$$

where $\boldsymbol{W}_\theta$ and $\boldsymbol{W}_u$ are positive-definite weighting matrices. This formulation includes uncertain parameters lying within a hyperellipsoid as well as independent bounds on each element.

For brevity, only the simplest techniques for the worst case analysis of batch processes is summarized here. A first-order expansion of the performance objective with respect to the model parameters gives

$$\delta\Phi = \boldsymbol{L}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) = \boldsymbol{L}\boldsymbol{\delta\theta} \tag{15}$$

where $\boldsymbol{L}$ denotes the sensitivity coefficients given by

$$L_i = \left.\frac{\partial\Phi}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{16}$$

Based on this expansion, the worst-case deviation of the performance is defined by (Ma and Braatz, 2001)

$$\delta\Phi_{wc} = \max_{\|\boldsymbol{W}_\theta \boldsymbol{\delta\theta}\| \le 1} \left|\boldsymbol{L}\boldsymbol{\delta\theta}\right| \tag{17}$$

Similar worst case analysis with respect to the control implementation inaccuracies requires a second-order series expansion:

$$\delta\Phi = \boldsymbol{M}\boldsymbol{\delta u} + \boldsymbol{\delta u}^T \boldsymbol{H}\boldsymbol{\delta u} \tag{18}$$

where

$$M_j = \left.\frac{\partial\Phi}{\partial u_j}\right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} \tag{19}$$

$$H_{ij} = \left.\frac{\partial^2\Phi}{\partial u_i \partial u_j}\right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} \tag{20}$$

Then the worst-case performance deviation due to control errors is:

$$\delta\Phi_{wc} = \max_{\boldsymbol{\delta u}_{\min} \le \boldsymbol{\delta u} \le \boldsymbol{\delta u}_{\max}} \left|\boldsymbol{M}\boldsymbol{\delta u} + \boldsymbol{\delta u}^T \boldsymbol{H}\boldsymbol{\delta u}\right| \tag{21}$$

This optimization problem is equivalent to

$$\max_{\mu_\Delta(N) \ge k} k \tag{22}$$

where k is any real number, the perturbation $\Delta = \operatorname{diag}\{\Delta_r, \Delta_r, \delta_c\}$ consists of independent real scalar blocks Δ_r and a complex scalar δ_c, and

$$\boldsymbol{N} = \begin{bmatrix} 0 & 0 & k\boldsymbol{w} \\ k\boldsymbol{H} & 0 & k\boldsymbol{H}\boldsymbol{z} \\ \boldsymbol{z}^T\boldsymbol{H} + \boldsymbol{M} & \boldsymbol{w}^T & \boldsymbol{z}^T\boldsymbol{H}\boldsymbol{z} + \boldsymbol{M}\boldsymbol{z} \end{bmatrix} \tag{23}$$

where

$$w = \tfrac{1}{2}\left(\boldsymbol{\delta u}_{\max} - \boldsymbol{\delta u}_{\min}\right) \tag{24}$$

$$z = \tfrac{1}{2}\left(\boldsymbol{\delta u}_{\max} + \boldsymbol{\delta u}_{\min}\right) \tag{25}$$

and $\boldsymbol{\delta u}_{\max}$ and $\boldsymbol{\delta u}_{\min}$ are the upper and lower bounds for the control implementation inaccuracies. Tight upper and lower bounds for k can be computed in polynomial-time using iterative μ-calculations or skewed-μ analysis.

6. RESULTS AND DISCUSSION

The wafers were implanted with 2×10^{15} ions/cm^2 of boron at 0.60 keV with 0° tilt, which gave a junction depth of 40 nm. The total boron was assumed to be initially 20% substitutional boron and 80% interstitial boron (Kobayashi*, et al.*, 2001). The initial conditions for Si interstitials agreed with the "+1" model, where Si interstitial concentration tracked the total boron concentration. The clusters and the B_s-Si_i complex were assumed not present initially. Boundary conditions at the surface for all species assumed no flux (*i.e.*, $J_i|_{\text{surface}} = 0$) with no surface Fermi level pinning (Jung*, et al.*, 2001). The optimization is solved by extending the golden search method (Press, *et al.*, 1992) to multidimensional problems.

Figure 4 presents the optimal RTA programs using linear and quadratic parameterizations of the temperature trajectory, which give junction depths of 51.3 and 48.2 nm, respectively (see Fig. 5), and the same sheet resistance of 350 Ω/sq. The optimal linear heating and cooling rates were 400 ºC/s and 200 ºC/s, respectively, indicating that the optimal RTA program was to effectively heat and cool as quickly as possible to the annealing temperature of 1111 ºC, in agreement with experimental studies (Agarwal, 2000; Mannino*, et al.*, 2001). The use of a high annealing temperature with fast heating and cooling can be explained by the lower effective activation energy of TED compared to boron activation.

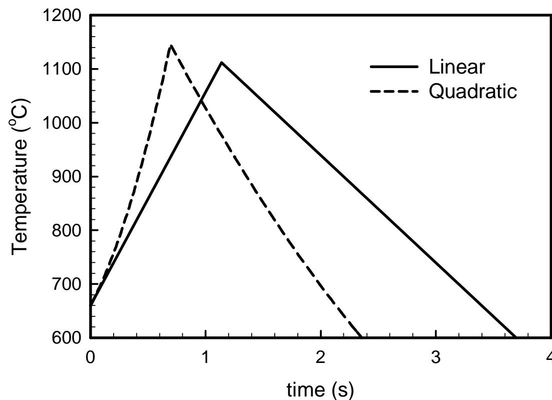


Fig. 4. Optimal RTA programs employing linear and quadratic parameterizations.

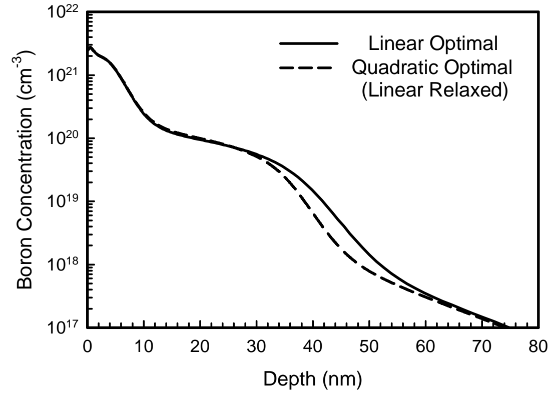


Fig. 5. Simulations of after-anneal Boron profiles employing the optimal RTA programs.

Although experimental studies (Agarwal, 2000; Mannino*, et al.*, 2001) had alluded to using fast heating and cooling rates, the determination of the maximum annealing temperature was made through extensive trial and error. In contrast, the optimal control formulation using the TED model can directly determine the annealing temperature, and therefore reduce the number of costly experiments.

The quadratic parameterization was applied to the heating ramp, while the cooling rate was kept at the optimal linear case. The slope of the quadratic profile was limited to 1000 ºC/s, which gave the optimal trajectory with an annealing temperature of 1144 ºC. The quadratic heating profile only gave minimal improvement of the junction depth over a linear heating profile. If the optimization problem for linear heating profile was solved using a relaxed constraint on the heating rate $\beta_{\max}$ of 1000 ºC/s, the optimal annealing temperature increased to 1146 ºC giving a junction depth of 48.2 nm (see Fig. 5). In other words, if the same maximum heating rate is used, the quadratic and linear parameterizations give the same minimum junction depth. Since heating rate is the true constraint in practice, this indicates that there is no benefit to using quadratic over linear heating and cooling profiles.

Worst case analysis was applied to the linear optimal trajectory. The model parameter uncertainties were quantified by the MAP covariance estimate (Gunawan*, et al.*, 2003). The analysis on control implementation inaccuracies used control trajectory perturbations of 5 ºC, 10 ºC, and 15 ºC at five temperatures along the heating and cooling ramps (660 ºC, 800 ºC, 950 ºC, 1050 ºC, 1100 ºC, 1112 ºC). Table 1 presents the worst-case junction depth increases due to uncertainties in the model parameters and control implementation.

Table 1. Worst-case junction depth increases (in nm) from parameter uncertainty and control errors.

$\delta\theta$	$\lvert\delta u\rvert \le 5$ ºC	$\lvert\delta u\rvert \le 10$ ºC	$\lvert\delta u\rvert \le 15$ ºC
0.13	1.89	5.15	9.78

The analysis results indicate that the deviations from the optimal junction depth were minimal for model parameter uncertainties and moderate to significant for control inaccuries. These results indicate that the MAP estimation gave parameter estimates with sufficient accuracy for use in optimal control studies. The typical RTA controllers make ~20 control moves every second (Bratschun, 1999), which translates to every 20 ºC in the heating step. Further accounting of nonuniformity of temperature across the wafer, the control inaccuracies could exceed 15 ºC at any given time. The analysis indicates that existing feedback controllers for implementing RTA programs need improvement as future junction depth requirements necessitate further reduction of the junction depth.

7. CONCLUSIONS

This paper has shown that the optimal RTA program for minimizing TED while achieving the desired sheet resistance consisted of fast linear heating and cooling profiles, as suggested in many experimental studies. Worst case analysis on the optimal junction depth deviations suggested the need of improvements in existing RTA controllers and advances in RTA technology to ensure temperature uniformity across the wafer.

REFERENCES

Agarwal, A. (2000). *Ultra-shallow junction formation using conventional ion implantation and rapid thermal annealing*. Paper presented at the *Int. Conf. on Ion Impl. Tech*., Austria.

Agarwal, A., Gossmann, H.-J., & Fiory, A. T. (1999). Effect of ramp rates during rapid thermal annealing of ion implanted boron for formation of ultra-shallow junctions. *J. Elec. Mat.,* **28**(12), 1333-1339.

Bratschun, A. (1999). The application of rapid thermal processing tehcnology to the manufacture of integrated circuits - An overview. *J. Elec. Mat.,* **28**, 1328-1332.

Collart, E. J. H., Murrell, A. J., Foad, M. A., van den Berg, J. A., Zhang, S., Armour, D., Goldberg, R. D., Wang, T. S., Cullis, A. G., Clarysse, T., & Vandervorst, W. (2000). Cluster formation during annealing of ultra-low-energy boron-implanted silicon. *J. Vac. Sci. & Tech. B,* **18**(1), 435-439.

Downey, D. F., Falk, S. W., Bertuch, A., F., & Marcus, S. D. (1999). Effects of "fast" rapid thermal anneals on sub-keV boron and BF_2 ion implants. *J. Elec. Mat.,* **28**(12), 1340-1344.

Gelpey, G., Elliot, K., Camm, D., McCoy, S., Ross, J., Downey, D. F., & Arevalo, E. A. (2002). *Advanced annealing for sub-130 nm junction formation.* Paper presented at the *201st Meeting of the ECS*, Philadelphia, PA.

Gunawan, R., Jung, M. Y. L., Seebauer, E. G., & Braatz, R. D. (2003). Maximum *a posteriori* estimation of transient enhanced diffusion energetics. *AIChE J.*, in press.

Haynes, T. E., Eaglesham, D. J., Stolk, P. A., Gossmann, H. J., Jacobson, D. C., & Poate, J. M. (1996). Interactions of ion-implantation-induced interstitials with boron at high concentrations in silicon. *Appl. Phys. Lett.,* **69**(10), 1376-1378.

Jain, S. C., Schoenmaker, W., Lindsay, R., Stolk, P. A., Decoutere, S., Willander, M., & Maes, H. E. (2002). Transient enhanced diffusion of boron in Si. *J. Appl. Phys.,* **91**(11), 8919-8941.

Jung, M. Y. L., Gunawan, R., Braatz, R. D., & Seebauer, E. G. (1999). Detailed TED modeling of transient enhanced diffusion in implanted Si. *AIChE Annual Meeting*, Dallas, TX. Paper 189d.

Jung, M. Y. L., Gunawan, R., Braatz, R. D., & Seebauer, E. G. (2001). Surface Fermi level pinning: An electrical "valve" in transient enhanced diffusion. *Materials Research Society Spring Meeting*, San Francisco, CA. Paper J4.21.

Kobayashi, H., Nomachi, I., Kusanagi, S., & Nishiyama, F. (2001). Lattice site location of ultra-shallow implanted B in Si using ion beam analysis. In: *Si Front-end Processing - Physics & Technology of Dopant-Defect Interactions III,* pp. J5.3. MRS, Inc., Warrendale, PA.

Laidler, K. J. (1987). *Chemical Kinetics*. Harper & Row, New York, NY.

Law, M. E., & Tasch, A. (2000). Florida object oriented process simulator (FLOOPS) 2000.

Ma, D. L., & Braatz, R. D. (2001). Worst-case analysis of finite-time control policies. *IEEE Trans. on Control Sys. Tech.,* **9**(5), 766-774.

Mannino, G., Stolk, P. A., Cowern, N. E. B., Boer, W. B. d., Dirks, A. G., Roozeboom, F., Berkum, J. G. M. v., Woerlee, P. H., & Toan, N. N. (2001). Effect of heating ramp rates on transient enhanced diffusion of ion-implanted silicon. *Appl. Phys. Lett.,* **78**(7), 889-891.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY.

Shishiguchi, S., Mineji, A., Hayashi, T., & Saito, S. (1997). Boron implanted shallow junction formation by high-temperature/short-time/high-ramping-rate (400 °C/sec) RTA. *Symposium on VLSI Technology*, Japan.

Stolk, P. A., Gossmann, H. J., Eaglesham, D. J., Jacobson, D. C., Rafferty, C. S., Gilmer, G. H., Jaraiz, M., Poate, J. M., Luftman, H. S., & Haynes, T. E. (1997). Physical mechanisms of transient-enhanced dopant diffusion in ion-implanted silicon. *J. Appl. Phys.,* **81**(9), 6031-6050.

Vortek Industries Ltd. (2002). Private communication.

Zeghbroeck, B. V. (2002). http://ece-www.colorado.edu/~bart/book/.

APPLICATION OF REDUCED-RANK MULTIVARIATE METHODS TO THE ANALYSIS OF SPATIAL UNIFORMITY OF SILICON WAFER ETCHING

Pratik Misra[1], Michael Nikolaou[1], Andrew D. Bailey III[2]

1. Dept. of Chemical Engineering, University of Houston
Houston, TX 77204-4004
2. Lam Research Corporation
4650 Cushing Parkway Fremont CA 94538

Abstract: We provide a smooth introduction to reduced-rank multivariate analysis, and show how it can be used to monitor images of etched silicon wafers. Results from two industrial case studies are presented and discussed.

Copyright © 2002 IFAC

Keywords: Plasma etching, principal component analysis, singular value decomposition, image processing.

1 INTRODUCTION

Spatially uniformity is necessary for high yields in a number of crucial processes of the semiconductor manufacturing industry, such as etching or deposition of thin films and chemical-mechanical planarization (CMP). In plasma etching, good spatial uniformity is the result of both appropriate design of etching tools as well as development of successful recipes. For either of these tasks, the designer or operator must be able to assess spatial uniformity characteristics, understand similarities and differences between tools or recipes, and apply criteria for the monitoring of spatial uniformity from tool to tool or run to run. Because uniformity is usually expressed in terms of a single number (e.g., 3σ/[average etch depth]) very different spatial uniformity profiles may result in the same numerical value of uniformity (Figure 1), thus masking important information that could be useful in a number of ways related to tool or recipe performance.

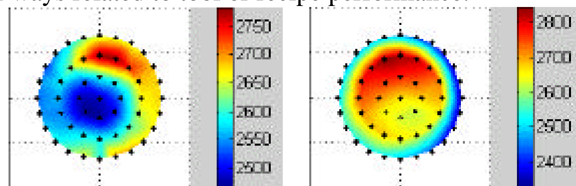


Figure 1 – Etch rate profiles on 300-mm wafer surface, interpolated over 49 measurement points (black dots). Both wafers correspond to virtually the same numerical uniformity value, but exhibit very different etch patterns.

In this presentation we provide a brief tutorial overview of the fundamentals of reduced-rank analysis, a topic that has found widespread use in chemical engineering. We show how it can be applied to the analysis, comparison, monitoring, and control of images corresponding to etch patterns of silicon wafers. Similar rank reduction techniques, especially Karhunen-Loeve (KL) transform, have been used to study spatiotemporal patterns on catalyst surface by Krischer et al.(1993) and in analyis and control of paper machines by Rigopoulos and Arkun (1996).

2 COMPRESSION OF COLLINEAR DATA VIA SVD

2.1 Basic case: Deterministic signals, no noise

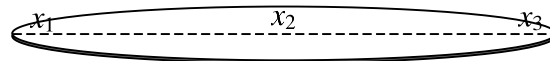


Figure 2 – Etch rate measurement points

An unrealistic but instructional example setting

Suppose that etch rates, x_1, x_2, x_3 are exactly measured at three points (edge/center/edge) along the diameter of a wafer, as shown in Figure 2. We want to know if the etch profiles are similar and etching process consistent.

Noiseless data are collected

Note that, for now, the data are assumed to be exact, i.e. *there is no measurement noise*. A set of data collected is shown in the matrix **X** below, and Figure 3.

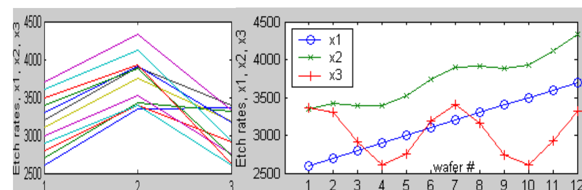


Figure 3 – Hypothetical etch rate profiles for 12 wafers (left) and Hypothetical local etch rates vs. wafer # (right).

$$\mathbf{X} = \overbrace{\begin{bmatrix} 2600 & 3348 & 3361 \\ 2700 & 3423 & 3311 \\ 2800 & 3392 & 2907 \\ 2900 & 3393 & 2609 \\ 3000 & 3527 & 2757 \\ 3100 & 3745 & 3182 \\ 3200 & 3900 & 3400 \\ 3300 & 3919 & 3163 \\ 3400 & 3882 & 2740 \\ 3500 & 3934 & 2614 \\ 3600 & 4118 & 2927 \\ 3700 & 4327 & 3324 \end{bmatrix}}^{x_1 \quad x_2 \quad x_3} \triangleq [\mathbf{x}_1\, \mathbf{x}_2\, \mathbf{x}_3] \tag{1}$$

Data collinearity and computation of matrix rank

Are the variables x_1, x_2, x_3 linearly dependent? i.e. is there a nonzero vector $\mathbf{a} \triangleq [a_1\ a_2\ a_3]^T$ such that

$$a_1 x_1 + a_2 x_2 + a_3 x_3 = 0 \Leftrightarrow \mathbf{x}^T \mathbf{a} = 0 \tag{2}$$

If so, the data satisfy the relationship (**model equation**)

$$a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + a_3 \mathbf{x}_3 = 0 \Leftrightarrow \mathbf{X}\mathbf{a} = \mathbf{0} \text{ for } \mathbf{a} \neq \mathbf{0} \tag{3}$$

A numerically robust method to check whether eqn. (3) is valid is the singular value decomposition (SVD). Detailed treatment of SVD can be found in a number of standard texts such as Horn and Johnson (1985). In SVD a matrix of rank r is decomposed as

$$\begin{aligned} \mathbf{X} &= \underbrace{s_1 \mathbf{u}_1}_{\text{"score"}1} \underbrace{\mathbf{v}_1^T}_{\text{"loading"}1} + \cdots + \underbrace{s_r \mathbf{u}_r}_{\text{"score"}r} \underbrace{\mathbf{v}_r^T}_{\text{"loading"}r} \\ &\equiv \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \equiv \sum_{i=1}^{r} \mathbf{y}_i \mathbf{v}_i^T \end{aligned} \tag{4}$$

Application of SVD (e.g. in Matlab ®) to the data matrix $\mathbf{X}$, eqn. (4) yields that the rank of $\mathbf{X}$ is 2, and the matrix $\mathbf{X}$ can be decomposed as

$$\begin{aligned} \mathbf{X} = & \underbrace{19973 \begin{bmatrix} -0.26882 \\ -0.2727 \\ -0.26381 \\ -0.25876 \\ -0.26976 \\ -0.29076 \\ -0.30431 \\ -0.30144 \\ -0.2919 \\ -0.29302 \\ -0.30999 \\ -0.32996 \end{bmatrix}_{12}}_{\text{"score"}1,\, \mathbf{y}_1} \underbrace{\begin{bmatrix} -0.54865 & -0.65112 & -0.52443 \end{bmatrix}_1}_{\text{"loading"}1,\, \mathbf{v}_1^T} \\ & + \underbrace{1233.7 \begin{bmatrix} 0.53687 \\ 0.44752 \\ 0.14112 \\ -0.10007 \\ -0.068095 \\ 0.1339 \\ 0.20911 \\ 0.005124 \\ -0.31245 \\ -0.44865 \\ -0.31508 \\ -0.13062 \end{bmatrix}_{12}}_{\text{"score"}2,\, \mathbf{y}_2} \underbrace{\begin{bmatrix} -0.52217 & -0.22301 & 0.82317 \end{bmatrix}_1}_{\text{"loading"}2,\, \mathbf{v}_2^T} \\ & + \underbrace{\mathbf{0}^{12\times 1}}_{\text{"score"}3,\, \mathbf{y}_3} \underbrace{\begin{bmatrix} -0.65293 & 0.72548 & -0.21764 \end{bmatrix}_1}_{\text{"loading"}3,\, \mathbf{v}_3^T} \end{aligned} \tag{5}$$

The above eqn. (5) implies that each row of the matrix $\mathbf{X}$ can be written as a linear combination of the row vectors *loading*1 and *loading*2, i.e.

$$\underbrace{[x_1\ x_2\ x_3]}_{\mathbf{x}^T} = \underbrace{y_1}_{\text{"score"}1} \underbrace{\boxed{\mathbf{v}_1^T}}_{\text{"loading"}1} + \underbrace{y_2}_{\text{"score"}2} \underbrace{\boxed{\mathbf{v}_2^T}}_{\text{"loading"}2} \tag{6}$$

Because $\mathbf{V}$ is orthonormal, eqn. (6) yields the sought eqn. (2), i.e.

$$\mathbf{x}^T \mathbf{v}_3 = 0 . \tag{7}$$

Loadings can be interpreted as basic shapes that can be used to represent the raw data

Note that the row vectors *loading*1 and *loading*2 in eqn. (6) are **the same** for all rows of data triplets x_1, x_2, x_3; they appear to be related to the system and not to any individual wafer. Therefore, *loading*1 and *loading*2 can be *interpreted* as two basic shapes (Figure 4), whose linear combination (sum weighted by score entries) can produce any of the 12 measured shapes.

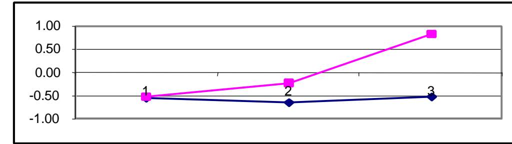


Figure 4 – Loadings, eqn. (5). The two shapes attempt to capture the curvature in the etch rate profile.

Monitoring scores gives a complete picture of the data

It follows from the preceding discussion that one can simply observe the scores (compressed data, values of *principal components* – hence PCA), to capture all information about the original data. In other words, instead of looking at

Figure 3, one can look at Figure 5.

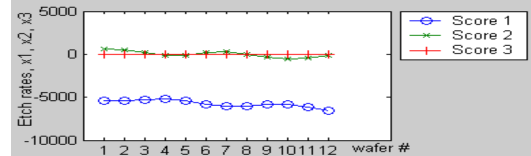


Figure 5 – Scores for the data in Figure 2, according to eqn. (5). Note that Score 3 is identically 0, which is precisely the equation sought in eqn. (2).

$rank(\mathbf{X}) = 2$ *implies data points fall on a plane*

Figure 6 shows 3-D plots of the data from two different viewpoints. The second viewpoint clearly shows that data fall on a plane.

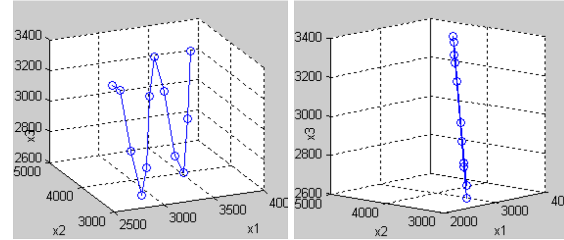


Figure 6 – 2-D world in 3-D data ("collinearity").

Loadings can also be thought of as weights used to relate original data to scores (compressed data)

If the score vectors $\mathbf{y}_1$, $\mathbf{y}_2$ are thought of as corresponding to two new variables, y_1, y_2, then y_1, y_2 are related to x_1, x_2, x_3 as follows: Because the loadings

are orthonormal, we can post-multiply eqn (4). by $\mathbf{v}_j$ to get

$$\mathbf{X}^{m\times n}\mathbf{v}_j^{n\times 1} = \underbrace{s_j \mathbf{u}_j^{m\times 1}}_{\text{"score" } j} \triangleq \mathbf{y}_j \quad (8)$$

or, row by row,

$$y_j = [x_1 \cdots x_n]\mathbf{v}_j \equiv \mathbf{x}^T\mathbf{v}_j = \mathbf{v}_j^T\mathbf{x} \quad (9)$$

or, in vector/matrix form,

$$\mathbf{y} = \mathbf{V}^T\mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{V}\mathbf{y} \quad (10)$$

(The new variables $\mathbf{y}$ are also called ***principal components***, see section 2.3.)

Thus, for this particular example we get, using eqn.(9), that the two nonzero score variables are

$$y_1 = [x_1\ x_2\ x_3]\begin{bmatrix}-0.54865\\-0.65112\\-0.52443\end{bmatrix},\quad y_2 = [x_1\ x_2\ x_3]\begin{bmatrix}-0.52217\\-0.22301\\0.82317\end{bmatrix} \quad (11)$$

and that the last score variable should be trivially equal to zero, i.e.

$$y_3 = [x_1\ x_2\ x_3]\begin{bmatrix}-0.65293\\0.72548\\-0.21764\end{bmatrix} = 0 \quad (12)$$

which is the same as eqn. (7).

This gives us the *second interpretation of loadings*: They are the vectors of coefficients by which we weight the original variables in linear combinations that produce a new set of variables (the "scores").

The preceding findings about **X** *can be used to monitor the system*

If the system etches subsequent wafers in the same way, it is reasonable to expect that data points (x_1, x_2, x_3) will be produced that are related as before, i.e. by eqn. (2). That means, equivalently, that if one first constructs 2 new variables y_1, y_2 in terms of eqn. (9) then the value of the *residual error* (cf. eqn. (6))

$$\mathbf{e}^T \triangleq \underbrace{[x_1\ x_2\ x_3]}_{\mathbf{x}^T} - \left(\underbrace{y_1}_{\text{"score"}1} \overbrace{\boxed{\mathbf{v}_1^T}}^{3}_{\text{"loading"}1} + \underbrace{y_2}_{\text{"score"}2} \overbrace{\boxed{\mathbf{v}_2^T}}^{3}_{\text{"loading"}2} \right) \quad (13)$$

$$= (\mathbf{x} - \mathbf{P}\mathbf{P}^T\mathbf{x})^T$$

for each new data triplet should be equal to zero, or, equivalently,

$$\|\mathbf{e}\|^2 \triangleq \mathbf{e}^T\mathbf{e} = 0 \Leftrightarrow \mathbf{x}^T(\mathbf{I} - \mathbf{P}\mathbf{P}^T)\mathbf{x} = 0 \quad (14)$$

where the matrix **P** consists of the first r columns of **V**. (The reason for using eqn. (14), instead of simply $\mathbf{e} = \mathbf{0}$, is that it can easily be extended to handle noisy data, as will be shown below).

Consider now the new data shown in Figure 7 .

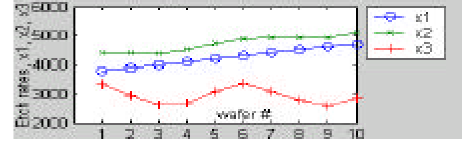


Figure 7 – Data set from 10 new wafers.

Applying the test of eqn. (14) to the data shown above yields the results of Figure 8. It is clear that two data points (#7 and #8) do not fall on the zero line as they should. These points indicate that the behavior of the system that etched these wafers is different from before.

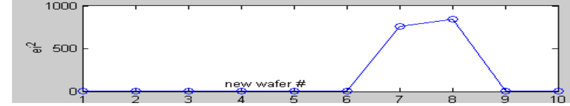


Figure 8 – (Errors)^2 for 10 new data sets, Figure 7.

2.2 Noisy signals

SVD on the noisy counterpart of **X** *reveals similar relationship among* x_1, x_2, x_3.

Table 1 – Noisy data

#	x_1	x_2	x_3
1	2585	3373	3353
2	2874	3586	3374
3	2809	3311	2861
4	2759	3355	2562
5	3175	3602	2763
6	3071	3753	3258
7	3424	3933	3486
8	3368	3974	3263
9	3526	3887	2709
10	3523	4034	2735
11	3546	4209	2910
12	3666	4381	3417

If measurements of x_1, x_2, x_3 are obtained with measurement noise as shown in the data of Table 1, SVD on the data of Table 1 yields singular values of 20219, 1206.5, 226.15 (cf. eqn.(5)). The eigenvalues (singular values squared) are shown in Figure 9. The smallest singular value is two orders of magnitude smaller than the largest one, indicating that it is probably equal to zero. But the second singular value is also an order of magnitude smaller than the largest singular value. Is it really non-zero or zero? How many singular values should be retained? What is the underlying rank of the data? How many singular values of **X** are really nonzero?

Let us call the noiseless data matrix Ξ and

$$\mathbf{X} = \Xi + \mathbf{E} \quad (15)$$

where **E** is a matrix that contains measurements errors. Note that for the data in Table 1

$$rank(\mathbf{X}) = 3 > rank(\Xi) = 2 \quad (16)$$

The singular values of **X**, $\sigma_{\mathbf{X}}$, can be bounded by bounds such as (Horn and Johnson, 1985):

$$|s_i(\mathbf{X}) - s_i(\Xi)| \le \|\mathbf{E}\|_{i2} = s_{\max}(\mathbf{E}) \quad (17)$$

Two simple criteria for detecting the number of essentially nonzero singular values of **X** are

(a) visual inspection of the singular value plot such as in Figure 9, and

(b) fidelity of reconstruction of the original data in **X**

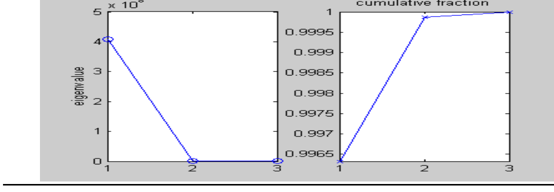


Figure 9 – Squared singular values (eigenvalues) for data in Table 1. (a) individual, (b) cumulative.

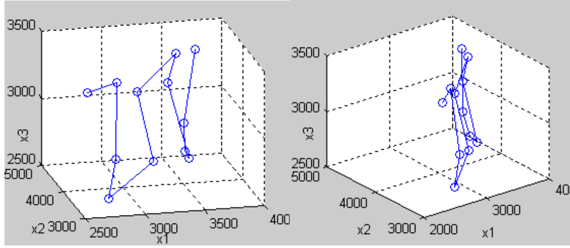


Figure 10 – 2-D world in noisy 3-D data.(cf. Figure 6).

Singular values quantify the goodness of data fit by a matrix of reduced rank

If only a "small" number of principal components is important, what is the best estimate of Ξ (with rank $r < n$) given the data in $\mathbf{X}$? Answering this question will allow us to construct scores and loadings, and to monitor the system, in the same way as we did in the noiseless case. The difference is that what should have been ideally zero errors, eqn. (13) should now be "small" (more in the sequel).

To find the best estimate $\hat{\Xi}$ of Ξ given $\mathbf{X}$ we can minimize the distance between Ξ and $\mathbf{X}$, i.e. find

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\| \tag{18}$$

When the norm in (18) is *induced 2 norm* or *Frobenius* norm, the solution is given by SVD as

$$\hat{\Xi} = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T \tag{19}$$

Moreover, the optimal difference can be shown to be

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\|_{i2} = \|\mathbf{X}-\hat{\Xi}\|_{i2} = s_{r+1} \tag{20}$$

and

$$\min_{rank(\Xi)=r<n} \|\mathbf{X}-\Xi\|_{F} = \|\mathbf{X}-\hat{\Xi}\|_{F} = \sqrt{\sum_{i=r+1}^{n} s_{r+i}^2} \tag{21}$$

Note that the singular vectors (loadings) of $\mathbf{X}$ could be very different from the singular vectors (loadings) of Ξ (Stewart, 1991). Figure 11, shows loadings for $\mathbf{X}$. Comparison with Figure 4 shows little difference.

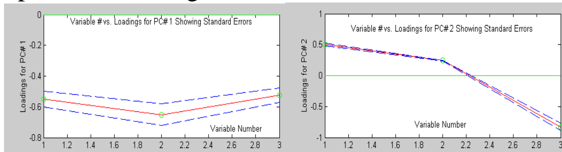


Figure 11 – Loadings(with error bounds) for noisy data of Table 1 (by PLS-toolbox®) (cf. Figure 4).

Process monitoring by looking at residual errors

Once the relationship among x_1, x_2, x_3 has been identified by the counterpart of eqn. (7) with noisy loading $\mathbf{v}_3$, the value of the *residual error* (i.e. counterpart of eqn. (13) for noisy loadings) for each new data point (x_1, x_2, x_3) arriving in the future can be checked. If the relationship among x_1, x_2, x_3 remains the same, then the residual error should be "small". This leads to the counterpart of eqn. (14) for noisy data. Specifically, if the residual error is normally distributed (very often a reasonable assumption) then $\|\mathbf{e}\|^2 = \mathbf{e}^T\mathbf{e}$ follows a chi-square distribution, from which one can construct Q-confidence as (cf. eqn. (14))

$$\mathbf{e}^T\mathbf{e} = \mathbf{x}^T(\mathbf{I}-\mathbf{P}\mathbf{P}^T)\mathbf{x} < d^2 \tag{22}$$

2.3 Stochastic signals

For multiple random variables principal components *are uncorrelated new variables, a few of which capture most variance*

SVD can provide additional insight if the vector variable $\mathbf{x}$ is stochastic. The analysis is known as *principal component analysis* (PCA) (Jolliffe, 1986).

Consider the random variable vector $\mathbf{x} \triangleq [x_1 \cdots x_n]^T$, and assume that $E[\mathbf{x}] = \mathbf{0}$ [1] where E denotes expected value. Denote the covariance matrix of $\mathbf{x}$ by

$$\mathbf{C} = E[\mathbf{x}\mathbf{x}^T] \in \Re^{n\times n} \tag{23}$$

It can be shown that we can use the modal matrix $\mathbf{A} \triangleq [\mathbf{a}_1 \ \cdots \ \mathbf{a}_n]$ of $\mathbf{C}$ (i.e. the matrix whose columns are the orthonormal eigenvectors of $\mathbf{C}$) to construct a new, zero-mean, vector random variable $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{A}^T\mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{A}\mathbf{y} \tag{24}$$

(**principal components**) that has the following important property

$$\mathrm{var}(y_i) = \max_{\substack{\|\mathbf{a}_i\|_2=1 \\ E[y_i y_{j<i}]=0}} \mathrm{var}\left(\mathbf{a}_i^T\mathbf{x}\right) = \lambda_i, \tag{25}$$

That is, each principal component, y_i is a weighted sum of the original variables $x_1, \ldots, x_n$, (eqn. (24)) such that

(a) its variance is maximal and equal to the i-th eigenvalue of the original covariance matrix $\mathbf{C}$ (eqn. (25)), and

(b) y_i is orthogonal to all previous principal components y_{i-j}, $i \geq 2$, $j = 1,\ldots,i-1$ (eqn. (25)).

---

[1] If the average of $\mathbf{x}$ is not zero, a new deviation variable can trivially be defined as $\mathbf{x}-E[\mathbf{x}]$. There is much higher chance that deviation variables (as opposed to original variables) are linearly dependent. Indeed, if the variables $\mathbf{x}$ satisfy the relationship $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, Taylor series expansion around $E[\mathbf{x}]$ yields

$$\mathbf{0} = \mathbf{f}(\mathbf{x}) \approx \underbrace{\mathbf{f}(E[\mathbf{x}])}_{=\mathbf{0}} + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right|_{\mathbf{x}=E[\mathbf{x}]} (\mathbf{x}-E[\mathbf{x}]) \triangleq \mathbf{B}\cdot\Delta\mathbf{x}$$

which implies linearly dependent $\Delta\mathbf{x}$.

SVD on covariance estimate produces values of principal components

Because the matrix **C** is unknown, it has to be estimated from data. The best estimate of **C** is

$$\mathbf{C} \approx \frac{1}{m-1}\mathbf{X}^T\mathbf{X} \qquad (26)$$

where **X** is a matrix that contains the data for each random variable in a column. Then, the eigenvalue/eigenvector pairs $(\kappa, \mathbf{w})$ of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ are estimates of the eigenvalue/eigenvector pairs $(\lambda, \mathbf{a})$ of **C**, which implies that

(a) the eigenvectors **w** of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ (hence the estimates of eigenvectors of **C**) are equal to the singular vectors **v** of **X** (eqn.(4)), and

(b) the eigenvalues of $\frac{1}{m-1}\mathbf{X}^T\mathbf{X}$ (hence the estimates of eigenvalues of C) are equal to $(m-1)$ times the squares of the singular values of **X**

Consequently, one can look at the values of

$$\frac{\sigma_i^2}{\sigma_1^2+\cdots+\sigma_r^2} = \frac{\sigma_i^2}{E[\mathbf{x}^T\mathbf{x}]} = \frac{\lambda_i}{\lambda_1+\cdots+\lambda_r} \quad i=1,\cdots,r \qquad (27)$$

to assess what percentage of the total variance of **x**, is captured by each of the principal components. By looking at the first few principal components, one can monitor the system that produces the data

(a) visually, e.g., by plotting PC1 vs. wafer #, PC2 vs. wafer #, etc. or PC1 vs. PC2 vs. PC3.

(b) numerically, by monitoring statistics such as the Hotelling statistic [5].

Principal components are directly related to multivariate SPC

If the zero-mean vector random variable **x** has (non-degenerate) covariance **C**, then one can construct the Hotelling (scalar) random variable

$$\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x} = \underbrace{\mathbf{x}^T\mathbf{A}}_{\mathbf{y}^T}\Lambda^{-1}\underbrace{\mathbf{A}^T\mathbf{x}}_{\mathbf{y}} \triangleq \mathbf{y}^T\Lambda^{-1}\mathbf{y} = \sum_{i=1}^{n}\frac{y_i^2}{\lambda_i} \qquad (28)$$

i.e. the Hotelling random variable is the sum of n independent random variables, y_i^2/λ_i. If some eigenvalues are zero, then we stop the summation in eqn. (28) at r, the rank of **C**, to ensure $\lambda_i \neq 0$.

3 CASE STUDY 1

Etch profiles (49 measurement points $x_1,\ldots,x_{49}$) from 9 different etching tools were collected, thus creating a 9×49 matrix **X**. Figure 12 indicates that 2 or 3 principal components result in less that 10% or 5% error, respectively. Corresponding scores are shown in Figure 13. Loadings are shown as weights in Figure 14 and as basis surfaces in Figure 15. The quality of reconstruction of the original data by 3 principal components is excellent, in that it captures curvature characteristics, as indicated by the samples shown in Figure 16.

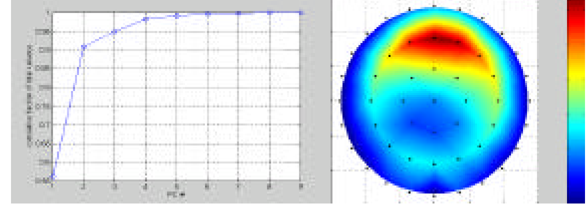


Figure 12 – Cumulative fraction of total variance captured by principal components (left) for variables $x_1,\ldots,x_{49}$ scaled by subtraction of sample averages $\bar{x}_1,\ldots,\bar{x}_{49}$ (right).

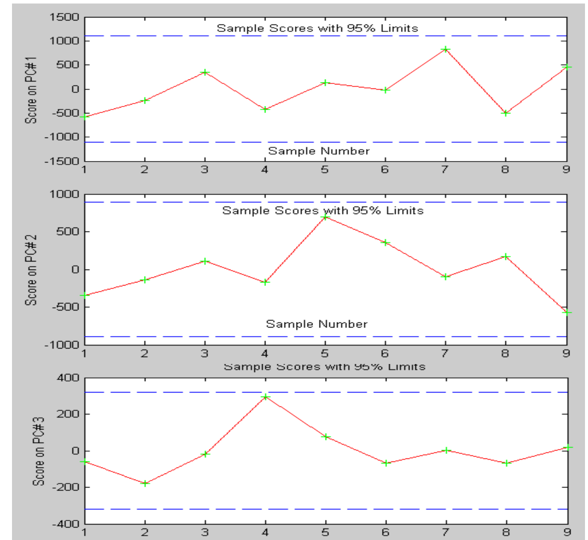


Figure 13 – Scores for the first 3 principal components (cf. Figure 5). (Confidence bounds by PLS-toolbox®)

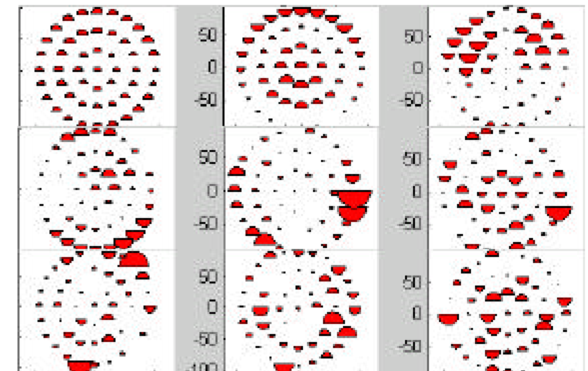


Figure 14 – Loadings as weighting coefficients for all 9 principal components. Semi-disk size and orientation denote magnitude and sign, respectively.

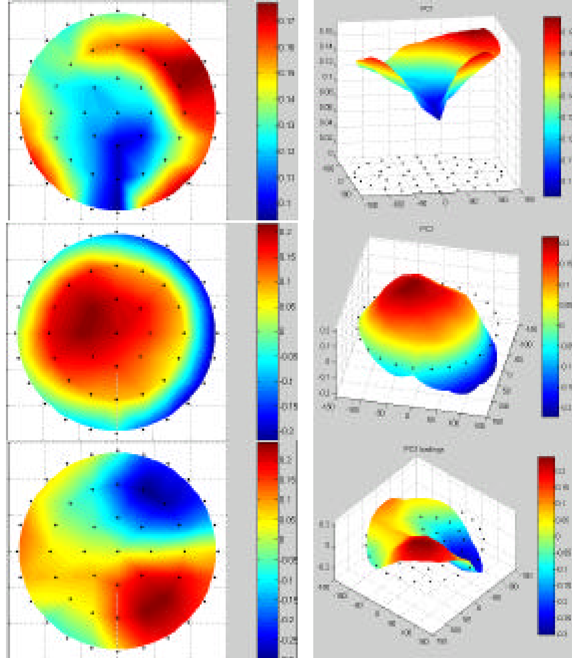


Figure 15 – Top and angle views of loadings as contour surfaces for the first 3 principal components.

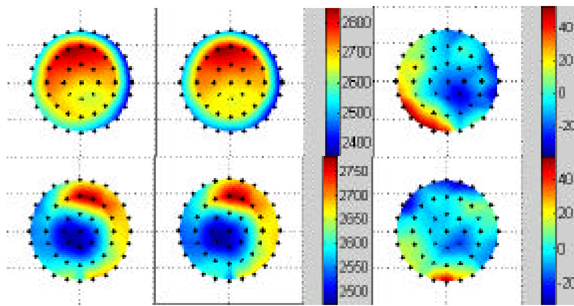


Figure 16 – Original etch profile (column 1), etch profile reconstructed from 3 principal components (column 2) and approximation error (column 3) for two sample wafers (cf. Figure 1)

4 CASE STUDY 2

18 200-mm silicon wafers were etched in an inductively coupled plasma reactor at Lam Research Corporation's facilities in Fremont, CA. Etch rates were measured at 49 points on the wafer, and a 18×49 data matrix **X** was constructed. Three principal components account for 99.94% of variation in data and are considered significant. The three loadings are shown in Figure 17. The scores are shown in Figure 18.

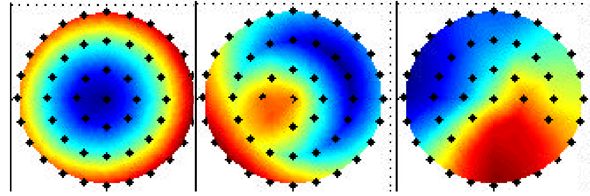


Figure 17 – Loadings of 3 principal components.

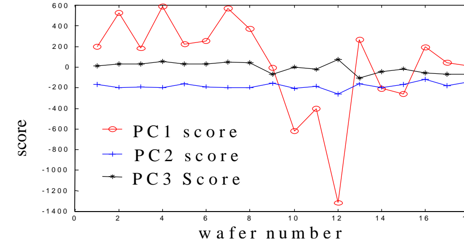


Figure 18 – Scores for PCs for experimental data

It can be observed that PC1 score varies far more than PC2 or PC3 score. There is a strong linear correlation between PC1 score and u_1, u_2 with $R^2 = 0.9686$ and $F = 200.24$. This implies that the first shape can be easily removed from the etch patterns and indeed we can see for wafer 9, PC1 score is almost zero. This information can be used to design better recipes.

5 CONCLUSIONS

Silicon wafer images depicting etch depth uniformity can be analyzed efficiently and effectively using reduced-rank multivariate methods. Two industrial case studies exemplify the basics summarized in this work.

REFERENCES

Horn, RA, and CR Johnson, *Matrix Analysis*, Cambridge University Press (1985).

Dewilde, P., and Ed. F. Deprettere, "Singular Value Decomposition: An Introduction," in *SVD and signal processing: algorithms, applications, and architectures*, edited by Ed. F. Deprettere, North-Holland, 3-41. (1988).

Rigopoulos A. and Y. Arkun, "Principal Component Analysis in Estimation and Control of Paper Machines", *Computers Chem. Engg.*, **20**, S1059-S1064, 1996.

Krischer. K., R. Rico-Martinez, I.G. Kevrikidis, H.H. Rotermund, G Ertl, and J.L. Hudson, "Model Identification Of A Spatiotemporally Varying Catalytic Reaction", *AICHE J.*, 39, 89-98, 1993.

Stewart, G., "Perturbation Theory for the Singular Value Decomposition", in *SVD and Signal Processing, II*, R. J. Vacarro ed., Elsevier, Amsterdam (1991).

Jolliffe, IT, *Principal Component Analysis*, Springer-Verlag (1986).

Acknowledgement – The first two authors gratefully acknowledge partial support for this work from Lam Research Corporation and the National Science Foundation through a GOALI grant.

REAL-TIME FEEDBACK CONTROL OF CARBON CONTENT OF ZIRCONIUM DIOXIDE THIN FILMS USING OPTICAL EMISSION SPECTROSCOPY

Dong Ni [1], **Yiming Lou** [1], **Panagiotis D. Christofides** [1],
Sandy Lao [2] and **Jane P. Chang** [2]

Department of Chemical Engineering
University of California, Los Angeles, CA 90095-1592

Abstract: In this work, we present a methodology for real-time carbon content feedback control of a plasma-enhanced metal organic chemical vapor deposition process using optical emission spectroscopy. Initially, an estimation model of carbon content of ZrO_2 thin films based on real-time optical emission spectroscopy data is presented. Then, a feedback control scheme, which employs the proposed estimation model and a proportional-integral controller, is developed to achieve carbon content control. Using this approach, a real-time control system is developed and implemented on an experimental electron cyclotron resonance high density plasma-enhanced chemical vapor deposition system at UCLA to demonstrate the effectiveness of real-time feedback control of carbon content. Experimental results of the deposition process under both open-loop and closed-loop operations are shown and compared. The advantages of operating the process under real-time feedback control in terms of higher productivity, reduced process variation and lower carbon content are demonstrated.

1. INTRODUCTION

The decrease of microelectronic device dimensions has motivated the replacement of silicon dioxide with oxides of higher dielectric constant (κ) as a dielectric layer in metal oxide semiconductor (MOS) devices. This is because for silicon dioxide layers thinner than about 1.6 nm, direct tunnelling currents through the oxide result in an exponential increase of leakage current. Significant leakage current increases the power dissipation and deteriorates the device performance and circuit stability for very large scale integrated (VLSI) circuits (Iwai and Momose, 1998; Lo *et al.*, 1997). In addition, since the minimum dimension of capacitors for 1-4 Gb dynamic random access memory (DRAM) generations falls into the deep submicron range, it is questionable whether acceptable charge storage can be achieved with SiO_2 within such small size regime.

The alternative is to use layers of a "new" high-κ dielectric, with the same equivalent oxide thickness or capacitance. A large number of high-κ candidate materials have been extensively studied. Among these candidate materials, ZrO_2 (as well as HfO_2) has several important properties which make it a leading candidate for an alternative dielectric. The dielectric constant of ZrO_2 is relatively high among the binary-metal oxides ($\kappa \sim 25$), and its thermal stability on Si is very good. Moreover, studies have indicated that pure ZrO_2 next to Si (with an ultra thin intervening SiO_x layer) remains stable up to 900°C (Copel *et al.*, 2000). In addition, ZrO_2 films have superior chemical resistance, good mechanical strength and a low leakage current level.

A variety of techniques can be used to prepare metal oxide thin films. Plasma-enhanced chemical vapor deposition (PECVD) is one of the most prominent means of preparing dielectric thin films, especially for memory devices applications, because of such advantages as low process temperature, high film growth rate and wide flexibility of deposition conditions. The use of metal-organic (MO) chemicals as precursors in PECVD of metal oxide thin films enables uniform film growth over large areas and complex surface ge-

[1] Process Control Group.
[2] Electronic Materials Synthesis and Plasma Processing Lab.

ometries. However, a potential problem of using MO precursors is the possibility of incorporation of impurities in the deposited thin film. One of the most important impurity species is carbon, which is abundant in the precursors. The incorporation of high concentration of carbon in the deposited film can negatively affect device performance by changing the dielectric constant and the leakage current density (Chaneliere *et al.*, 1998).

In general, carbon can be incorporated in the films either by forming carbides or oxides with the deposited metal or oxygen or by occupying intergranular positions among the grains of the main deposited compound in the form of cyclic or aliphatic species. Carbon incorporation can even occur simultaneously in multiple states depending on precursor, material to-be-deposited and operating conditions (Vahlas *et al.*, 1998; Maury *et al.*, 1996). Therefore, the development and implementation of real-time feedback control systems for carbon content control could improve the operation and use of MO precursors in the deposition of high-κ materials. Previous work on control of PECVD processes has mainly focused on control of deposition spatial uniformity (Armaou and Christofides, 1999) (see also (Armaou *et al.*, 2001) for results on control of plasma etching).

In this work, we present a methodology for real-time carbon content feedback control of a plasma-enhanced MOCVD process using optical emission spectroscopy (OES). Initially, an estimation model of carbon content of ZrO_2 thin films based on real-time OES data is presented. Then, a feedback control scheme, which employs the proposed estimation model and a proportional integral controller, is developed to achieve carbon content control. Using this approach, a real-time control system is developed and implemented on an experimental electron cyclotron resonance (ECR) high density PECVD system at UCLA to demonstrate the effectiveness of real-time feedback control of carbon content. Experimental results of the deposition process under both open-loop and closed-loop operations are shown and compared. The advantages of operating the process under real-time feedback control in terms of higher productivity, reduced process variation and lower carbon content are demonstrated.

2. ECR HIGH-DENSITY PECVD REACTOR

The schematic of the experimental ECR PECVD reactor system is shown in Figure 1. It consists of an ECR type microwave source, a reactor chamber, a pumping system, a pressure control system, a gas delivery system, an OES system and a computer-based real-time process control system.

Figure 2 shows the internal configuration of the reactor chamber. A 6-inch-diameter cylindrical stainless-steel chamber is surrounded by two circular coaxial electromagnets, which are 7 inches apart. An ASTeX ECR source is on top of the chamber. Microwave at 2.45 GHz is generated from the source and transmitted

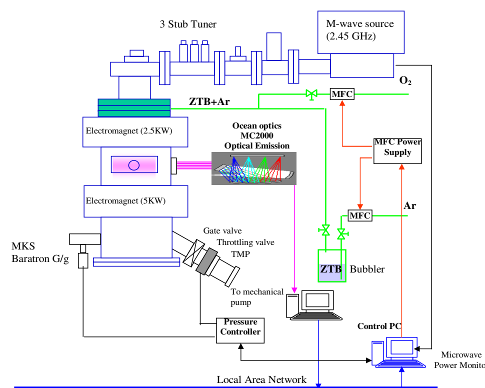


Fig. 1. Schematic diagram of the ECR plasma-enhanced CVD system used in this study.

into the chamber through a 3/8-inch thick vacuum-sealed quartz window and a high-density plasma is generated. A gas diffusion ring is located just below the top quartz window to conduct uniform distribution of the gases. A 4 inch-diameter anodized aluminum substrate holder is centered inside the chamber. The distance between the substrate holder and the top quartz window is adjustable in the range of 6.5 inches to 12 inches. The substrate holder is also connected with a 13.56 MHz radio frequency (RF) power supply tuned by a matching network; this allows controlling the ion impinging energy by applying bias voltage to the substrate.

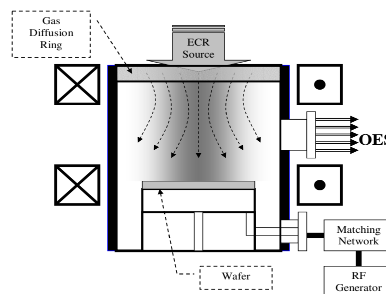


Fig. 2. Internal configuration of ECR PECVD chamber.

The chamber is pumped by a 140 l/s Alcatel 5150CP turbo-molecular pump (TMP) backed by a mechanical pump. The base pressure is measured with an HPS I-Mag cold cathode ion gauge. The chamber pressure can be controlled and varied between the base pressure and atmospheric pressure. The MKS 651C pressure controller takes the measurement of chamber pressure by an MKS 626A Baratron gauge as input and manipulates an MKS 253B throttle valve, thereby allowing to control the pressure independently from the gas flow rates.

We chose zirconium tetra-tert-butoxide $[Zr(OC_4H_9)_4]$ (ZTB) as our MO precursor because it has a sufficiently high vapor pressure (0.26 mbar at 60 ^oC) (Frenck *et al.*, 1991). A bubbler, which is kept at constant temperature (65 ^oC), is used for precursor delivery because ZTB is a liquid at room temperature (boiling point=90 ^oC). Ar is used as a carrier gas of the precursor vapor and the gas line is heated to 80 ^oC

to prevent the condensation of precursors. O_2 is used as an oxidant and mixed with Ar and ZTB at a point 8 inches away from the entrance to the reactor.

Throughout this study, the electric currents are fixed at 120 A for the 5 kW top magnet and 150 A for the 2.5 kW bottom magnet. The distance between the top quartz window and the substrate holder is kept constant at 6.5 inches and no bias is applied to the substrate.

3. OPTICAL EMISSION SPECTROSCOPY SYSTEM

Optical emission spectroscopy (OES) is the central real-time measurement tool used in this study. We use an Ocean Optics MC2000 OES system with five channels covering the wavelength range from 200 nm to 1000 nm to analyze the plasma. Each channel consists of independent optic setups including slits, gratings, a 2048-element linear silicon charged coupled diode (CCD) array and an optic fiber cable. The configurations of individual channels are shown in Table 1. The best optical resolution [full width at half maximum] for this system is 1.4 Å with a 10 μm slit width in the ultraviolet (UV) range. The integration time can be set within the range of 3 ms to 60 s. A sapphire window with minimal UV absorption is used as the OES port. The emission spectra are taken 1 in. above the substrate surface in this study so that gas phase information near the wafer surface can be collected.

Table 1. OES channel configurations of wavelength range, start pixel (SP), end pixel (EP) and resolution (in full width at half maximum [FWHM]).

CH	Range (nm)	SP	EP	Res.[FWHM]
0	196.14 ~ 354.44	4	2044	1.5
1	327.23 ~ 464.27	0	2047	1.4
2	437.93 ~ 617.89	0	2047	1.8
3	585.70 ~ 868.81	0	2046	2.8
4	786.50 ~ 1039.51	1	2047	2.5

Table 2. Transitions and wavelengths of atomic emissions observed.

(Striganov and Sventitskii, 1968)

Species	Wavelength (nm)	Transition
Ar	750.39	$4s'^{\,o}_{(1/2)} - 4p'_{(1/2)}$
C	247.85	$2p^{2\,1}S - 3s^1 4P^o$
H_β	486.13	$2p^2P^o - 4d^2D$
O	777.42	$3s^5S^o - 3p^5P$
Zr	350.93	
	351.96	
Zr^+	339.20	N/A
	343.82	
	349.62	

The major atomic emission peaks and molecular band heads observed in this study are summarized in Table 2 and Table 3, respectively. The analog signals produced by optical channels are captured by an Ocean Optics ADC1000 high-speed ISA-bus A/D converter installed in a Pentium PC. The OES data are then transmitted through fast ethernet to the computer used for real-time process control.

Table 3. Transitions and wavelengths of molecular emissions observed.

(Pearse and Gaydon, 1976)

Species	Wavelength (nm)	Transition
C_2	516.52	$A^3\Pi_g - X^3\Pi_u$
CH	431.42	$A^2\Delta - X^2\Pi$

4. FEEDBACK CONTROL SYSTEM: DESIGN AND IMPLEMENTATION

The carbon content of the thin film can not be measured directly in real-time, and thus, estimates of the carbon content, which are obtained based on plasma composition in the reactor chamber by OES, are used in the feedback control system. Previous spectroscopic study of the reaction plasma (Cho *et al.*, 2001) in this ECR PECVD system has shown that the carbon content in the film has a quasi-linear relationship with respect to the optical emission intensity ratio of C_2 molecules and O atoms in the reacting gas. This can be explained by the fact that carbon molecules are mostly responsible in forming the precursors for carbon incorporation into the film. This result suggests that the information of optical emission intensity ratio of C_2/O can be utilized to estimate the carbon content in the zirconium dioxide film in real-time.

In this work, a mathematical model is constructed to estimate the carbon content of the film based on the optical emission intensity ratio which is obtained through OES in real-time. Following the previous experimental results (Cho *et al.*, 2001), the relationship between the carbon content in the surface layer and the optical emission intensity ratio can be written as follows:

$$X_C^s(t) = A\gamma(t) \tag{1}$$

where X_C^s is the atomic concentration (%) of carbon in the surface of the film, A is a constant which is related to the operating condition of the specific experimental system (experimentally determined for our current chamber condition to be 11.92) and γ is the optical emission intensity ratio of C_2/O.

Under the assumption that the film growth rate remains constant, the carbon content of the whole film is obtained using the following formula:

$$X_C(t) = \frac{\int_{t_0}^{t} X_C^s(s)ds}{t - t_0} \tag{2}$$

where X_C is the atomic concentration (%) of carbon in the bulk of the deposited film at time t and t_0 is the time in which the deposition starts. In this case, we treat X_C as the time average of X_C^s. Combining Eqs. 1 and 2, the following estimation model is obtained:

$$X_C(t) = A\frac{\int_{t_0}^{t} \gamma(s)ds}{t - t_0} \tag{3}$$

We note that although the deposition process is a batch process in nature, an optimal operating recipe can not be obtained since no accurate mathematical model describing the relationship between the optical emission intensity ratio and the inlet mass flow rate is currently available. Thus, the control problem for the process is formulated as a set-point regulation problem; this approach is further justified by our experimental results which clearly show that the response time of the closed-loop system is significantly smaller than the total deposition time.

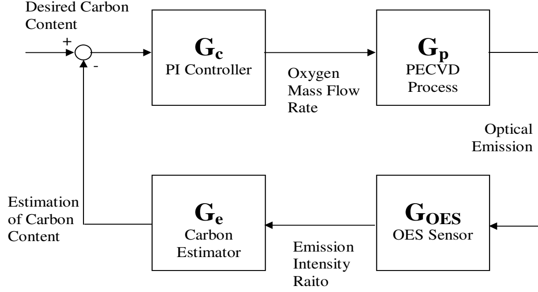


Fig. 3. Block diagram of the closed-loop system under the proposed carbon content controller.

Figure 3 shows the structure of the closed-loop system under the proposed carbon content controller. The input to the controller is the difference between the desired carbon content and the estimated carbon content and the controller manipulates the inlet oxygen mass flow rate. The sensor block G_{OES} can be treated as a pure time delay since it takes a fixed amount of integration time for the OES system to obtain good signal-to-noise ratios and transfer the OES data through the network. The G_e block is the carbon estimator described above. The G_c block is the controller based on the proportional-integral (PI) control algorithm (described below in detail). The G_p block is the process block describing the relationship between the change of oxygen mass flow rate and the optical emission intensity ratio γ of the plasma. G_p is identified experimentally and the identification procedure will be discussed in detail in subsection 5.1 below.

To eliminate unnecessary control actions, which may interfere with the plasma and lead to poor closed-loop performance, the control objective is to stabilize the carbon content value close to the desired set-point (i.e., within a certain tolerance ε). A PI control algorithm is used to achieve this objective of the following form:

$$\frac{f_{Ar}(t)}{f_{O_2}(t)} = U(t) = K_c\hat{e}(t) + K_i\int_{t_0}^{t}\hat{e}(\mu)d\mu + \bar{R}_f \quad (4)$$

$$\hat{e}(t) = \begin{cases} e(t) & |e(t)| > \varepsilon \\ 0 & |e(t)| \leq \varepsilon \end{cases} \quad (5)$$

where U is the output of the controller (i.e., the mass flow ratio of Ar/O_2), $\bar{R}_f$ is a steady state bias expressed in terms of the mass flow ratio of Ar/O_2 at steady state, f_{O_2} is the oxygen mass flow rate, f_{Ar} is the Argon mass flow rate which scales with the precursor vapor flow rate, e is the difference between the estimated carbon content and the set-point value, K_c is the proportional gain and K_i is the integral gain. The input of the controller $\hat{e}(t)$ is defined as in Eq.5 where ε is the tolerance within which we want to approach the desired set-point.

MATLAB simulations of the entire process model were performed to obtain reference values of the controller parameters to be used in the real-time computer control system. The reference values were initially computed by using the Ziegler Nichols (ZN) tuning method (e.g. (Coughanowr, 1991)) and then adjusted based on simulation results to achieve a desired closed-loop response.

The computer process control system was implemented on an Intel Pentium III 700 MHz PC with 512 MByte RAM. All the programs used in this study were written in LabVIEW language and National Instruments LabVIEW for Windows Version 6.1 was used as runtime platform.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 *Open-loop system*

The objective of the open-loop experiments is to study the dynamic behavior of the deposition process based on real-time OES measurements.

The first set of experiments (3 independent runs) were performed to study the relationship between the steady-state value of γ and the mass flow ratio of Ar/O_2, R_f. The experimental results are shown in Figure 4; each data point is obtained by setting R_f at a fixed value and measuring γ after 200 s to guarantee that the process has reached steady-state. The experimental results in Figure 4 suggest that the optical emission intensity ratio varies proportionally with respect to the cubic of the mass flow rate ratio; this relationship is shown by the dotted line and can be mathematically expressed as follows:

$$\gamma_{ss} = K_p R_f^3 \quad (6)$$

where γ_{ss} is the steady-state value of the optical emission intensity ratio and K_p is a constant which depends on the processing chamber conditions and the carrier gas flow rate.

In the second experiment, the process dynamics are identified by varying the mass flow ratio R_f in a way shown in the top curve in Figure 5 and measuring γ in real time using OES; the experimental results are presented in Figure 5. It can been seen that the process can be approximated by a first-order system which has a small time constant.

Using the experimental results shown in Figures 4 and 5, we constructed a *Simulink* model shown in

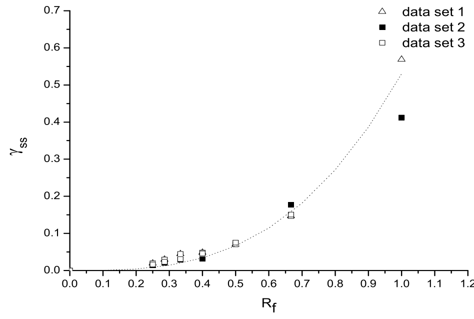


Fig. 4. Experimental data of R_f vs. γ_{ss} from different depositions for fixed argon flow rate 8 sccm, chamber pressure at 40 mTorr and microwave power 300 W.

Figure 6 within a MATLAB environment to simulate the process; $R_f(t)$ is the input and $\gamma(t)$ is the output. The model parameters were identified from the experiments to be K_p=0.53 and τ_p=10 s.

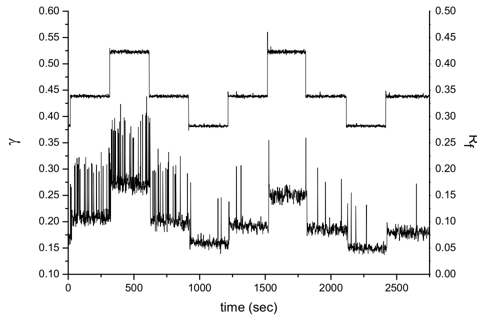


Fig. 5. Response curve of γ for step changes in R_f for argon flow rate 8 sccm, chamber pressure 40 mTorr and microwave power 300 W.

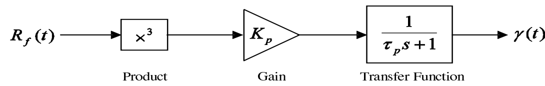


Fig. 6. Simulink representation for the process dynamics.

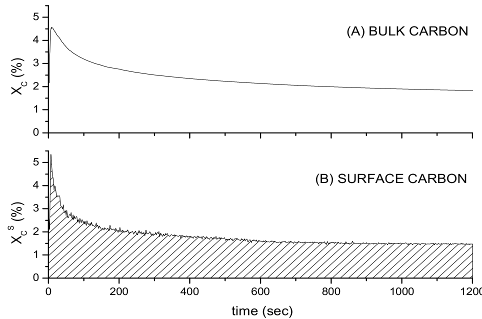


Fig. 7. Profiles of bulk (A) and surface (B) carbon concentration of a ZrO_2 film computed based on real-time OES measurements during an open-loop deposition with microwave power 300 W, chamber pressure 40 mTorr, *Ar* flow rate 8.4 sccm and O_2 flow rate 8 sccm.

Figure 7 shows the evolutions of the carbon concentration of the surface (A) and of the bulk (B) of a ZrO_2 film during a typical open-loop deposition. The carbon concentrations are computed based on real-time OES measurements using the proposed estimation model. It can be observed that the starting stage of the deposition has relatively higher carbon incorporation. This corresponds to the OES measured high C_2 emission intensity and low O emission intensity during the initial stage of the deposition, as shown in Figure 8. Low O emission intensity indicates a low O concentration in the plasma; this may cause incomplete oxidation of the precursor, which leads to a high concentration of C_2 in the plasma during the initiation of the deposition process.

It can also be noticed in Figure 7 that the carbon concentration of the bulk of the film changes throughout the deposition process. This is not only because the bulk carbon concentration is an average value, but also because the carbon incorporation rate varies with time. This time variation may be explained by the continuous increase of O concentration in the plasma due to the complex and competing serial oxidation and dissociation processes (Cho *et al.*, 2002). As a result, reaction products with different compositions are generated and different amount of carbon is incorporated into the film at different times during the deposition process. Due to the existence of these uncertainties in the deposition process, the profile of bulk concentration of carbon shown in Figure 7 is not reproducible in our experiments; this suggests that it is very difficult to obtain a desired carbon concentration with open-loop operation.

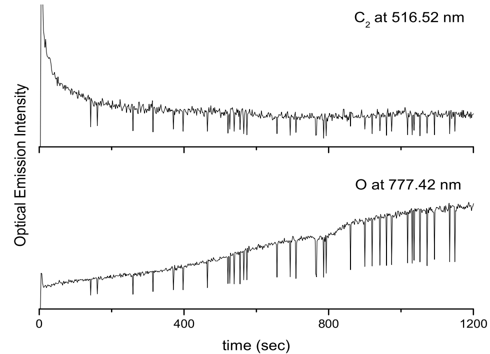


Fig. 8. Profiles of C_2 and O optical emission intensity during an open-loop deposition with microwave power 300 W, chamber pressure 40 mTorr, *Ar* flow rate 8.4 sccm and O_2 flow rate 8 sccm.

5.2 *Evolution of the closed-loop system*

Using the developed real-time feedback control system, a carbon content-controlled deposition experiment was performed (see (Ni *et al.*, 2003) for more results of closed-loop deposition experiments). Figure 9 shows a 20-minute long controlled-deposition which was carried out with microwave power fixed at 300 W, chamber pressure controlled at 40 mTorr and Ar flow rate set at 8 sccm. The carbon content controller was implemented with a set-point value for the atomic car-

bon concentration of 1.4%, proportional gain K_c=1.0, integral gain K_i=0.05 and error tolerance ε=0.03%.

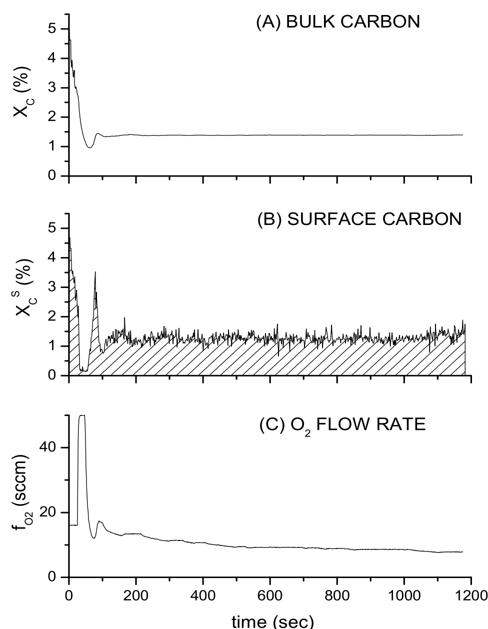


Fig. 9. Profiles of bulk (A) and surface (B) carbon concentration of a ZrO_2 film computed based on real-time OES measurements and profile of manipulated oxygen flow rate (C) during a controlled deposition experiment with microwave power 300 W, chamber pressure 40 mTorr and Ar flow rate 8.4 sccm.

From the bulk carbon concentration curve in Figure 9, we can see that the carbon content of the film was controlled very closely to the desired value of 1.4% in spite of the initial plasma disturbance mentioned above (this result was also verified through off-line XPS analysis of the deposited film; see (Ni *et al.*, 2003) for details). The response time is relatively small compared to the deposition duration which supports our set-point regulation formulation of this control problem.

Comparing the bulk carbon concentration profile of the thin films under closed-loop (Figure 9, top plot) and open-loop (Figure 7, top plot) conditions with the same initial deposition conditions, it can be clearly seen that the carbon content of the film was reduced by more than a factor of 5 under closed-loop control. Moreover, compared to open-loop operation, the controlled process is more robust with respect to disturbances caused by system start-up, and mass flow rate and plasma variations.

6. REFERENCES

Armaou, A. and P. D. Christofides (1999). Plasma-enhanced chemical vapor deposition: Modeling and control. *Chem. Eng. Sci.* **54**, 3305–3314.

Armaou, A., J. Baker and P. D. Christofides (2001). Feedback control of plasma etching reactors for improved etching uniformity. *Chem. Eng. Sci.* **56**, 1467–1475.

Chaneliere, C., J. L. Autran, R. A. B. Devine and B. Balland (1998). Tantalum pentoxide (Ta_2O_5) thin films for advanced dielectric applications. *Mater. Sci. Eng., R.* **22**, 269–322.

Cho, B. O., J. Wang and J. P. Chang (2002). Metalorganic precursor decomposition and oxidation mechanisms in plasma-enhanced ZrO_2 deposition. *J. Appl. Phys.* **92**(8), 4238.

Cho, B. O., S. Lao, L. Sha and J. P. Chang (2001). Spectroscopic study of plasma using zirconium tetra-tert-butoxide for the plasma enhanced chemical vapor deposition of zirconium oxide. *J. Vac. Sci. Tech. A* **19**(6), 2751.

Copel, M., M. Gribelyuk and E.P. Gusev (2000). Structure and stability of ultrathin zirconium oxide layers on Si(001). *Appl. Phys. Lett.* **76**, 436–438.

Coughanowr, D. R. (1991). *Process Systems Analysis and Control*. McGraw-Hill. New York.

Frenck, H. J., E. Oesterschulze, R. Beckmann, W. Kulisch and R. Kassing (1991). Low temperature remote plasma-enhanced deposition of thin metal oxide films by decomposition of metal alkoxides. *Mater. Sci. Eng.A* **139**, 394–400.

Iwai, H. and H. S. Momose (1998). Ultra-thin gate oxides performance and reliability. *IEDM Tech. Dig.* pp. 163–166.

Lo, S. H., D. A. Buchanan, Y. Taur and W. Wang (1997). Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's. *IEEE Electron. Device Lett.* **18**, 209–211.

Maury, F., L. Gueroudji and C. Vahlas (1996). Selection of metalorganic precursors for MOCVD of metallurgical coatings: Application to Cr-based coatings. *Surf. Coat. Technol.* **87**, 316–324.

Ni, D., Y. Lou, P. D. Christofides, L. Sha, S. Lao and J. P. Chang (2003). A method for real-time control of thin film composition using OES and XPS. In: *Proceedings of the American Control Conference*. Denver, Colorado.

Pearse, R. W. B. and A. G. Gaydon (1976). *The Identification of Molecular Spectra*. Wiley. New York.

Striganov, A. R. and N. S. Sventitskii (1968). *Tables of Spectral Lines of Neutral and Ionized Atoms*. IFI/Plenum. New York.

Vahlas, C., F. Maury and L. Gueroudji (1998). A thermodynamic approach to the CVD of chromium and of chromium carbides starting from Cr(C_6H_6)(2). *Chem. Vap. Deposition* **4**, 69–76.

DESIGN, SIMULATION, AND EXPERIMENTAL TESTING OF A SPATIALLY CONTROLLABLE CVD REACTOR

Jae-Ouk Choo[#], Raymond A. Adomaitis[#], Gary W. Rubloff*, Laurent Henn-Lecordier*, and Yijun Liu*

[#]*Department of Chemical Engineering and Institute for Systems Research*
**Department of Materials and Nuclear Engineering and Institute for Systems Research*
University of Maryland
College Park, MD 20742

Abstract: Most conventional chemical vapor deposition systems do not have the spatial actuation and sensing capabilities necessary to control deposition uniformity, or to intentionally induce nonuniform deposition patterns for single-wafer combinatorial CVD experiments. In an effort to address this limitation, a research program is underway focusing on developing a novel CVD reactor system that can explicitly control the spatial profile of gas-phase chemical composition across the wafer surface. This paper discusses the development of a simulator for the three-segment prototype that has recently been constructed and the results of preliminary experiments performed to evaluate the performance of the prototype in depositing tungsten films. *Copyright © 2003 IFAC*

Keywords: Semiconductor processing; Chemical vapor deposition; Distributed parameter systems; Simulation.

1. INTRODUCTION

Chemical Vapor Deposition (CVD) is one of the essential processes in semiconductor manufacturing because of its ability to deposit thin smooth films conformally onto submicron-scale features. CVD processes have evolved together with the semiconductor industry, from early bell-jar CVD reactors to the current cold-wall single-wafer reactor (Xia *et al*., 2000). Although current conventional CVD reactors produce thin smooth films successfully, their configurations lack the 2-dimensional spatial controllability necessary to counteract non-uniformity generators such as reactant depletion.

Significant research effort has been put into improving uniformity of deposited thin films (Wang *et al*., 1986; Moffat and Jensen 1988; Kleijn *et al*., 1989; Moslehi *et al*., 1995; van der Stricht *et al.*, 1997; Yang *et al*., 1999; Theodoropoulos *et al*., 2000). Most of these studies have focused on optimizing the process operating parameters (gas input rate, wafer temperature, and reactor pressure) and apply to reactor configurations that do not allow any adjustments to their overall geometry (Wang *et al*., 1986; Moffat and Jensen 1988; Kleijn *et al*., 1989; Moslehi *et al*., 1995; van der Stricht *et al.*, 1997). Because current configurations of CVD reactors mostly lack precise control actuators for gas delivery to the wafer surface, the spatial control of film characteristics becomes limited by inflexible CVD reactor configurations. While some research has focused on distributing precursor gases across the wafer surface with pre-specified spatial variation (Yang *et al.*, 1999; Theodoropoulos *et al*., 2000), the reactor designs were mainly motivated by the goal of decreasing gas phase reactions in MOCVD processes using designs that separate precursors to improve film uniformity.

As a response to these perceived CVD reactor design shortcomings, we have developed a novel CVD reactor intended to improve across-wafer 2-dimensional controllability. This new CVD reactor introduces a segmented showerhead design featuring

individually controllable gas distribution actuators, a design that reverses the residual gas flow by directing it up through the showerhead (henceforth referred to as the reverse-flow design), and sampling ports for in-situ gas sampling.

In this paper, we describe the proof of this novel design concept by simulation and a sequence of experiments performed using a prototype reactor. We refer to this design as the Programmable CVD Reactor Concept because of the potential of real-time control of gas phase composition across the wafer surface. An illustration of the three-zone prototype Prgrammable CVD reactor is shown in Fig. 1.

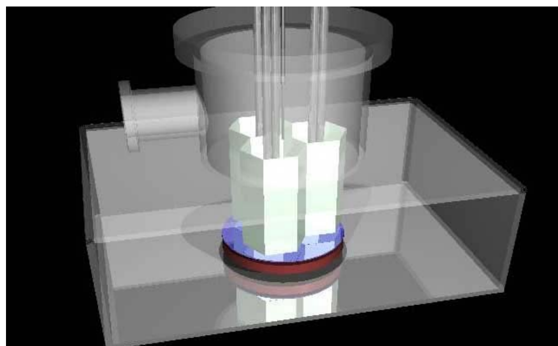


Fig. 1. The Programmable CVD reactor illustrating the segmented showerhead structure.

2. CVD REACTOR PROTOTYPE DESIGN

The major design feature of the Programmable CVD reactor is its segmented showerhead. A schematic diagram of the reactor is shown in Fig. 2.

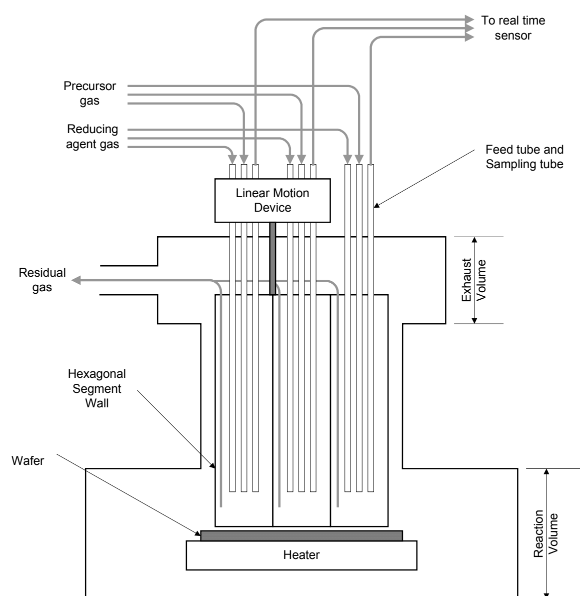


Fig. 2. Schematic diagram showing a vertical cross-section of the Programmable CVD reactor and its feed gas delivery system.

The effect of the segmented showerhead design is to discretize the region above the wafer surface into individually controllable regions. Because each segment is fitted with separate feed gas lines, the precursor gas composition in the area of wafer surface corresponding to each segment can be individually adjusted.

To enhance film uniformity in the wafer area corresponding to each segment and to reduce interaction between segments, residual gas is recirculated up through each segment of the showerhead and mixed in a common exhaust volume above the showerhead honeycomb structure (Fig. 3). The reverse-flow of exhaust gas means diffusional transport dominants in the region above the wafer and below the bottom of the showerhead segments. This design feature increases the controllability of across wafer gas composition relative to conventional CVD reactors, which normally draw residual gas across the wafer surface (e.g., Moslehi *et al.,* 1995).

Showerhead/wafer spacing is controlled with the linear motion device shown in Figure 2. The sampling tube of each segment can be used to transport a small amount of gas to a real time in-situ sensor, such as a mass spectrometer. From the residual gas analysis of each segment, approximate film thickness and the composition of film deposited on each area corresponding to each segment can be determined. Also, this sampling tube and sensor can be used to diagnose process operation during a deposition run.

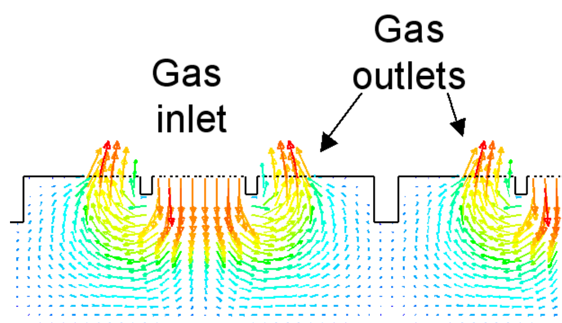


Fig. 3. Periodic gas flowfields generated by the recirculating showerhead design.

To test the feasibility of the Programmable CVD concept, we have constructed a prototype reactor by modifying one reaction chamber of an Ulvac-ERA1000 CVD cluster tool. The Ulvac-ERA1000 CVD cluster tool located on the University of Maryland's campus is a commercial CVD tool used for selective tungsten deposition. In its original configuration, the hydrogen reducing gas entered through a quartz showerhead above the wafer; wafer heating was provided by a ring of heating lamps above the showerhead. As part of the programmable reactor modification, substrate heating was used in place of lamp heating, and the quartz showerhead was replaced by a new assembly consisting of a three-segment honeycomb structure equipped with two feed tubes and one sampling tube per each segment. (Fig. 4.)

Fig. 4. A photograph of the three-zone prototype showerhead mounted on the Ulvac reactor chamber.

The prototype system is designed to deposit tungsten films using the hydrogen reduction process; this deposition process was chosen to test the prototype reactor because the Ulvac reactor originally was designed for tungsten deposition, and the reactions for tungsten deposition by hydrogen reduction have well known mechanisms and rate expressions (Arora and Pollard, 1991; Kleijn 2000; Kleijn *et al.*, 1991; Kleijn and Werner, 1993). Additionally tungsten deposition remains a commercially important manufacturing process (Ireland, 1997).

2. PROTOTYPE EXPERIMENTAL TESTS

A number of initial experiments were performed using the three zone prototype to validate the design assumptions and collect data for developing a detailed process simulator. Typical operating conditions for the first experiments consisted of a 0.5 torr chamber pressure, a wafer temperature of 350°C, and 20 minute deposition times. In all cases where the showerhead/wafer spacing was small (e.g., 1mm), distinct hexagonal film patterns were produced (Fig. 5), and as anticipated, the pattern because more diffuse as the showerhead/wafer increased.

In one particular set of experiments, pure Ar was fed to Segment 1 at a flowrate of 50sccm; 50sccm of WF_6 was fed to segment 2, and 50sccm of H_2 was fed to segment 3. The film thickness in the region below each segment was determined by sheet resistance measurements using a four-point probe; the interpolated results are shown in Fig. 6. While some W deposition should take place directly under Segment 2 (where pure WF_6 is fed) due to the Si reduction mechanism, it is interesting to note that some W deposition takes place under the remaining two segments. An explanation for this phenomenon is presented in the following sections on segment simulator development.

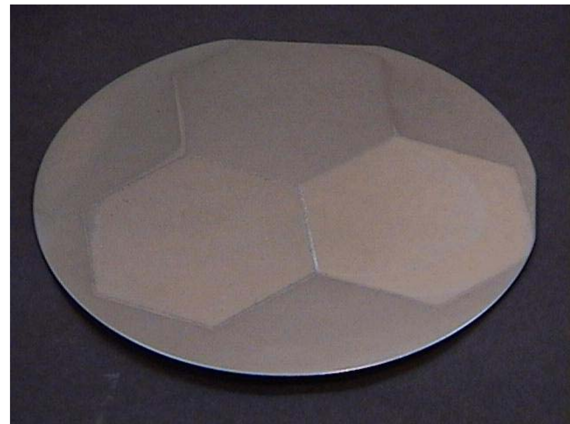


Fig. 5. Typical tungsten deposition pattern produced by the prototype Programmable CVD reactor.

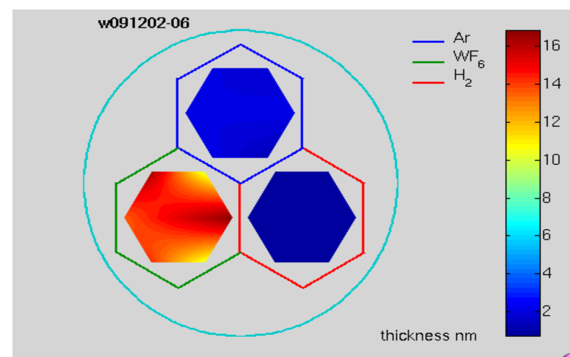


Fig. 6. Deposition thickness profiles corresponding to an experiment in which pure Ar, WF_6, and H_2 were fed to the individual segments.

3. MODELING AND SIMULATION

Significant effort is being put into developing process simulation tools to assess the effectiveness of the segmented design and to determine the operating conditions for future experiments. Because of the Programmable CVD reactor's reverse-flow design, reactants in the gas mixture in the common exhaust volume can diffuse back into the segments. Therefore, to sustain the pre-specified gas compositions at the bottom of each segment, the back diffusion through the segment should be suppressed below an acceptable level by the convective upward flux contribution.

A steady-state 1-dimensional segment model (for each segment) combined with a well-mixed common exhaust volume model was used to assess the ability of the segmented structure to maintain significant segment-to-segment gas composition differences near the wafer surface. The geometry of a single segment, together with the notation used in the model development, is shown in Fig. 7. A schematic diagram of common exhaust volume is shown in Fig. 8 where the each shaded area represents the top of the each segment of the showerhead.

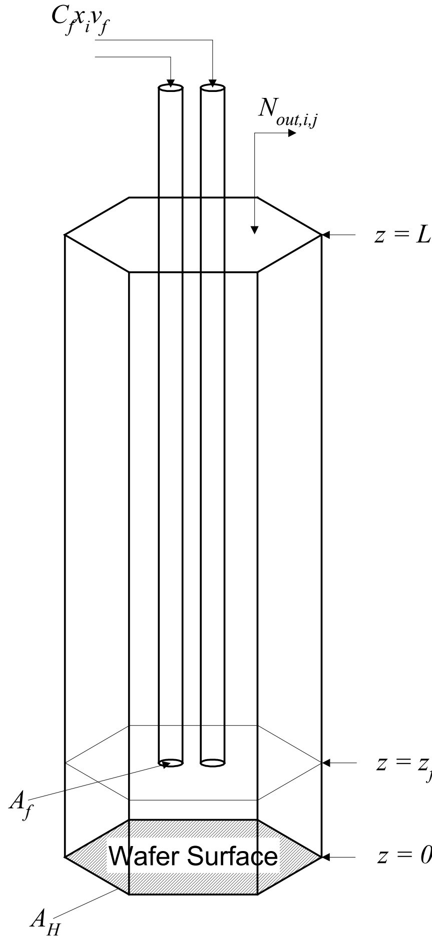


Fig. 7. Schematic diagram of a single showerhead segment.

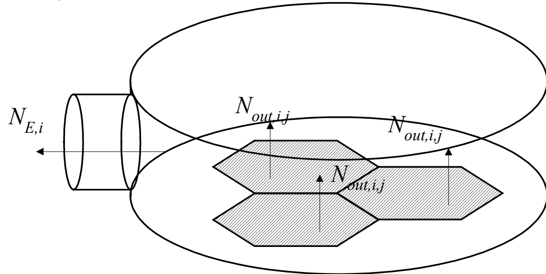
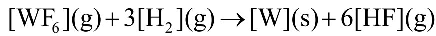


Fig. 8. Schematic diagram of the common exhaust volume.

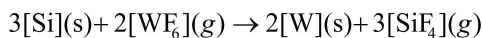
3.1 Film deposition reactions

The overall reaction of tungsten deposition by hydrogen reduction is



$$[WF_6](g) + 3[H_2](g) \rightarrow [W](s) + 6[HF](g)$$

The gas phase reactions associated with this deposition process are negligible due to low reactor pressure during the process operation (Arora and Pollard, 1991; Kleijn *et al.*, 1991). Surface reactions by Si reduction will occur during the film nucleation step



$$3[Si](s) + 2[WF_6](g) \rightarrow 2[W](s) + 3[SiF_4](g)$$

This is a self-limiting reaction and typically accounts for a 10-200 nm of W film thickness (Groenen *et al.*, 1994).

3.2 Multi-component reactant transport.

In the limit of zero distance between the segment wall bottom and wafer surface, the mass balance for each species in each segment can be written as

$$\frac{\partial Cx_i}{\partial t} = -\nabla N_i + F_i \text{, where } i = 1,2,\ldots,n \qquad (2)$$

where C is the total gas concentration in mol/m^3, N_i is the total flux of species i in $mol/(m^2\ s)$, and F_i is a forcing function accounting for the change in flux due to fresh feed of species i from the segment feed tube. At steady-state, (2) can be rewritten as

$$N_i = \begin{cases} \alpha_i \mathrm{R}_{\mathrm{kin}} & (0 \le z \le z_f) \\ \alpha_i \mathrm{R}_{\mathrm{kin}} + C_f x_i v_f \left(\frac{A_f}{A_H} \right) & (z_f \le z \le L) \end{cases} \qquad (3)$$

where $\mathrm{R}_{\mathrm{kin}}$ represents the rate of either of the deposition reactions and α_i is the corresponding stoichiometric coefficient of species i in that reaction.

The multicomponent gas species transport can be expressed by the Stefan-Maxwell equation. The binary diffusion coefficient is estimated by the Chapman-Enskog kinetic theory and Neufield method (Kleijn and Werner, 1993; Reid *et al.*, 1987). Neglecting any effect of pressure and forced diffusion, the Stefan-Maxwell equation is written as

$$\nabla x_i = \sum_{j=1, j \ne i} \frac{1}{CD_{ij}} \left(x_i \overline{N}_j - x_j \overline{N}_i \right) \qquad (4)$$

where

$$\overline{N}_i = N_i + \frac{\mathbf{D}_i^T}{M_i} \nabla \ln T$$

and $\mathbf{D}_i^T$ is the multicomponent thermal diffusion coefficient defined in Kleijn and Werner (1993)

To examine whether back-diffusion of WF_6 from the common exhaust volume through the segments where Ar and H_2 are the only feed gas species can account for the W deposition in these segments, we compute the maximum WF_6 concentration possible for these segments by setting N_i, the total species flux by combined thermal and normal diffusion, to zero at $z=0$ in (3). The boundary condition at the segment top is based on the assumption that the gases leaving each showerhead segment are mixed perfectly in the common exhaust volume (Fig. 8) giving

$$x_{i,j} = \frac{N_{out,i,j}}{\sum_{i=1}^{n} N_{E,i}} \quad \text{at } z=L \quad \text{where } N_{E,i} = \sum_{j=1}^{ns} N_{out,i,j} \qquad (5)$$

and where subscript i denotes species number and j refers to segment number. A linear temperature profile was assumed for the gas located in the region

between the wafer and the bottom of the feed gas tube, where it was assumed the gas entered at room temperature.

Given boundary conditions (5), the species moles fractions $x_{i,j}$ can be represented as a function of spatial position by a truncated global trial function expansion. A Galerkin projection method is used to spectrally discretize the system; a Newton-Raphson procedure then is used to solve the resulting set of algebraic equations. For these simulations, it was found that a truncation number of 20 was sufficient to obtain converged solutions to this boundary-value problem.

4. SIMULATION RESULTS

Results of this solution procedure, with simulation conditions set to match the experimental conditions that produced the wafer shown in Fig. 5 are shown in Fig. 9. In this Figure, the wafer surface is located at *z=0* (the left axis limit) and the segment top is to the right; the vertical line represents the location of the bottom of the feed tube bundle inside each segment.

The top plot shows that for the Segment 1, in which pure Ar is fed, the major gas component is Ar; however, there is significant back-diffusion of H_2 and it appears that sufficient WF_6 diffuses back into the segment to account for the W film found in the experiments. The effect of thermal diffusion is clearly evident in these plots: note how the H_2 profile increases near the heated wafer surface. Similarly, we observe the back-diffusion of WF_6 into the H_2-fed segment, and the dominance of WF_6 in the segment where only WF_6 is fed. We conclude from this simulation that the potential for significant back diffusion of WF_6 can account for the thin W film deposited in Segments 2 and 3, where no W-containing species were fed.

5. CONCLUSIONS

A novel, spatially controllable CVD reactor design has been developed and a prototype reactor was constructed by modifying a commercial CVD cluster tool. Preliminary simulations and experiments demonstrate the feasibility of producing spatially-patterned film characteristics by controlling gas phase reactant composition directly above the wafer surface

This approach to thin-film manufacturing control opens the door to a new generation of CVD reactor design, allowing single-wafer combinatorial studies and precise across-wafer uniformity control in a single reactor design. Current research focuses on developing a detailed process simulator that will be used to fully exploit the capabilities of this new reactor system.

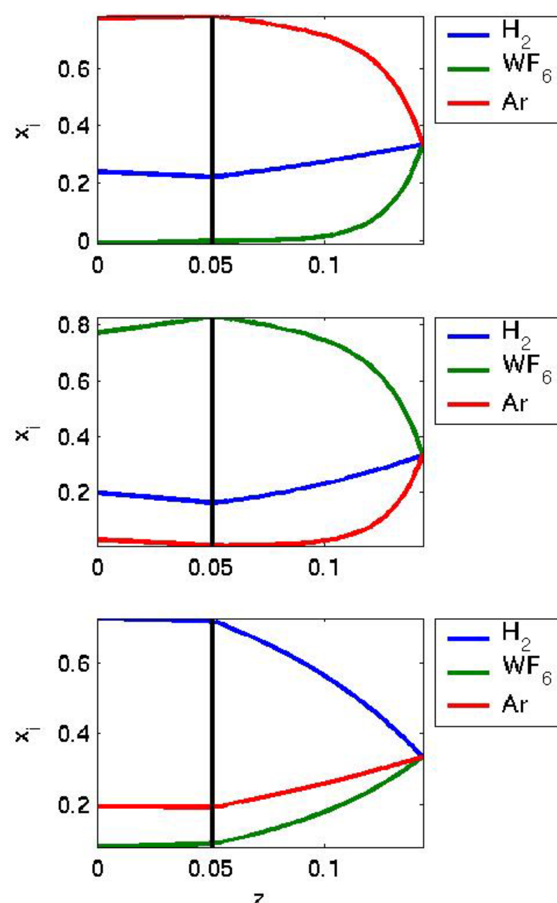


Fig. 9. Reactant gas composition profiles for the 3 segments corresponding pure Ar feed to Seg. 1, WF_6 to Seg. 2, and H_2 to Seg. 3; all at 50sccm.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the National Science Foundation through grant CTS-0085632 for construction of the prototype and simulation work, the continued support of NSF through CTS-0219200, and National Institute of Standards and Technology for fabricating several showerhead components.

REFERENCES

Arora, R. and R. Pollard, (1991) A Mathematical model for chemical vapor deposition process influenced by surface reaction kinetics: application to low-pressure deposition of tungsten, *J. Electrochem. Soc.* **138**(5) 1523.

Groenen, P.A.C., J.G.A. Holscher, and H.H. Brongersma, (1994) Mechanism of the reaction of WF_6 and Si, *Applied Surface Sci* **78**, 123.

Ireland, P.J., (1997) High aspect ratio contacts: A review of the current tungsten plug process, *Thin Solid Films* **304**, 1.

Kleijn, C.R., (2000) Computational modeling of transport phenomena and detailed chemistry in chemical vapor deposition – a benchmark solution, *Thin Solid Films* **365**, 294.

Kleijn, C.R., C.J. Hoogendoorn, A. Hasper, J. Holleman and J. Middelhoek, (1991) Transport phenomena in tungsten LPCVD in a single-wafer reactor, *J. Electrochem. Soc.* **138**, 509.

Kleijn, C.R., Th.H. van der Meer, and C.J. Hoogendoorn, (1989) A mathematical model for LPCVD in a single wafer reactor, *J. Electrochem.l Soc.* **136**, 3423.

Kleijn C.R. and C. Werner, (1993) *Modeling of chemical vapor deposition of tungsten films*, Basel; Boston: Birkhäuser Verlag.

Moffat, H.K. and K.F. Jensen, (1988) Three-dimensional flow effects in silicon CVD in horizontal reactor, *J Electrochem Soc.* **135**, 459.

Moslehi, M.M, C.J. Davis and R.T. Matthews, (1995) Programmable multizone gas injector for single-wafer semiconductor processing equipment, United State Patent #5,453,124.

Reid, R.C., J.M. Praunitz and B.E. Poling, (1987) The properties of gases and liquids (4th edition), New York, McGraw-Hill.

Theodoropoulos, C., T.J. Mountziaris, H.K. Moffat and J. Han, (2000) Design of gas inlets for the growth of gallium nitride by metalorganic vapor phase epitaxy, *J. Crystal Growth* **217**, 65.

van der Stricht, W., I. Moerman, P. Demeester, J.A. Crawley and E.J. Thrush, (1997) Study of GaN and InGaN films grown by metaloganic chemical vapor deposition, *J. Crystal Growth* **170**, 344

Wang, C.A., S.H. Groves and S.C. Palmateer, (1986) Flow visualization studies for optimization of OMVPE reactor design, *J. Crystal Growth* **77**, 136.

Xia, L., P.W. Lee, M. Chang, I. Latchford, P.K. Narwankar, R. Urdahl, 'Chapter 11. Chemical Vapor Deposition', (2000) *Handbook of Semiconductor Manufacturing Technology, Yoshi Nishi, Robert Doering, ed.*, New York: Marcel Dekker. .

Yang, C., C. Huang, G. Chi and M. Wu, (1991) Growth and characterization of GaN by atmosphere pressure metalorganic chemical-vapor deposition with a novel separate-flow reactor, *J. Crystal Growth* **200**, 39.