

Model-Based Uniformity Control for Epitaxial Silicon Deposition

Aaron E. Gower-Hall, Duane S. Boning, *Member, IEEE*, Peter Rosenthal, and Ann Waldhauer

Abstract—Semiconductor fabrication requires tight specification limits on thickness and resistivity for epitaxially deposited silicon films. Our testbed system for integrated, model-based, run-to-run control of epi films incorporates a Centura tool with an epi deposition chamber, an in-line epi film thickness measurement tool, and off-line thickness and resistivity measurement systems. Automated single-input-single-output, run-to-run control of epi thickness has been successfully demonstrated. An advanced multi-objective controller is described, which seeks to provide simultaneous epi thickness control on a run-to-run basis using the in-line sensor, as well as combined thickness and resistivity uniformity control on a lot-to-lot basis using off-line thickness and resistivity sensors.

Control strategies are introduced for performing combined run-to-run and lot-to-lot control, based on the availability of measurements. Techniques are proposed for dealing with multiple site measurements of multiple film characteristics taken at different sampling rates, as well as the use of time-based inputs and rate-based models. These concepts are widely applicable for semiconductor fabrication processes.

Index Terms—Epitaxial growth, process control, run-to-run control, semiconductor process modeling, uniformity control.

I. INTRODUCTION

SEMICONDUCTOR fabrication requires an increasingly expensive and integrated set of processes. The bounds on efficiency and repeatability for each process step continue to tighten under the pressure of economic forces and product performance requirements. Integrated run-to-run process control is one approach to meeting these challenges, and has been successfully demonstrated with processes such as CMP and plasma etch, among others [1], [2]. An in-line sensor for measuring epitaxial silicon (epi) film properties provides crucial information to enable integrated run-to-run control of the epi deposition process [3]. Epi deposition is typically the first step in a process flow, where the crystal lattice of the bulk silicon is extended through growth of a new silicon layer with a different doping level than the bulk. Matching target epi film thickness and resistivity parameters is crucial for the subsequent fabrication of properly functioning devices.

Manuscript received March 12, 2001; revised December 31, 2001. This work was supported by the National Science Foundation under the SBIR program and by a NIST ATP grant.

A. E. Gower-Hall and D. S. Boning are with the Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge MA 02139 USA.

P. Rosenthal is with MKS Instruments, On-Line Products Group formerly On-Line Technologies), East Hartford, CT 06108 USA.

A. Waldhauer is with Applied Materials, Santa Clara, CA 95054 USA.

Publisher Item Identifier 10.1109/TSM.2002.801368.

A testbed system is under joint development with the cooperation of MKS Instruments (On-Line Products Group), Applied Materials, and MIT, using an epi deposition chamber on a Centura tool, off-line thickness and resistivity measurement systems, and an in-line film thickness measurement tool. Integrated single-input-single-output (SISO), model-based run-to-run control of epi film thickness has been successfully demonstrated using the in-line metrology tool and automatic updates of the deposition time [4], [5].

There has been considerable research at MIT involving run-to-run control of semiconductor processes [1], [2], [4]–[9]. Implementations of controllers, dynamic models, and model update algorithms are explored by various researchers, including an “Internal Model Approach” [10], [11], “Robust Run by Run Control” [12], [13], “Linear EWMA” [1], “Neural Network Based EWMA” [14], [15], “Probabilistic EWMA” [16], “Adaptive EWMA” [17], [18], and “PCC” [1]. The system used in this work is a multi-input, multi-output, model-based controller with an Exponentially Weighted Moving Average (EWMA) model update algorithm. The underlying model is of the following matrix form:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \quad (1)$$

where $\mathbf{x}$ is the input vector and $\mathbf{y}$ is the output vector. Thus all that is needed to completely specify the model are the slope terms ($\mathbf{A}$) and the offset, terms ($\mathbf{b}$). Between each run the EWMA controller updates the offset terms based on the previous inputs, the measured outputs, and some controller parameters. Based on the new model and a set of target outputs, the controller uses “inverse model” techniques to suggest a new set of inputs for the next run. Investigations concerning the theoretical and practical issues of EWMA run-to-run control algorithms are found in a number of sources, including detailed analyses in [1], [8], [18].

Currently, a multi-objective model-based controller is under development, which seeks to provide simultaneous run-to-run control of not only epi thickness, but resistivity as well. Furthermore, thickness and resistivity averages and *uniformities* are to be driven to target, as opposed to simply averaged results or single measurement sites. This work considers sensor readings of film thickness and resistivity at multiple sites across the wafer, which makes uniformity control possible.

There are three major stages to consider for developing and implementing model-based run-to-run control: 1) selection of process parameters and a Design of Experiments (DOE); 2) process optimization; and 3) run-to-run control setup and testing. Parameter selection involves an analysis of the process to determine the outputs one wishes to monitor and control,

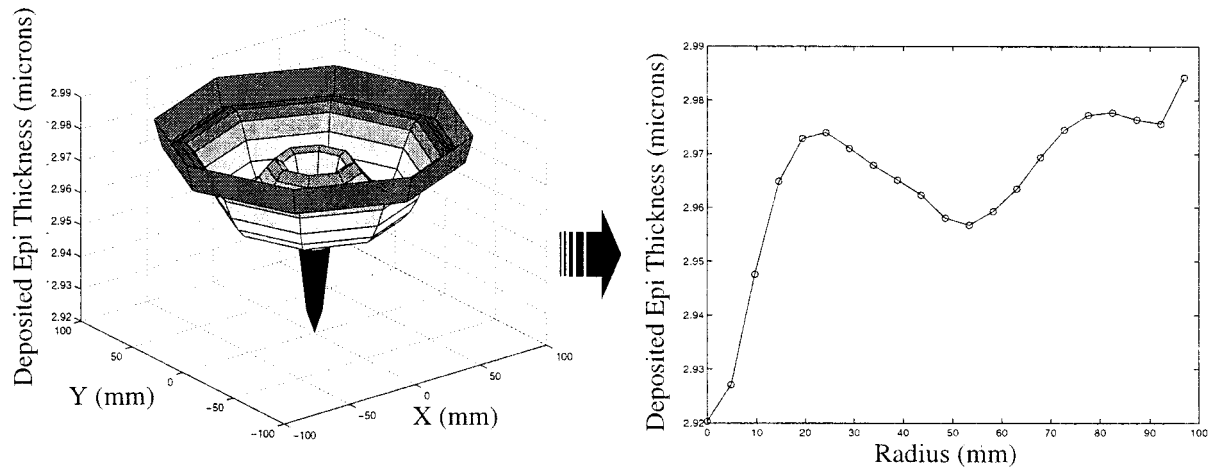


Fig. 1. Compression of 2-D radial wafer map sensor data into a single radial scan.

as well as the inputs one will use to drive those outputs. A set of designed experiments is then selected and performed to provide input-output data. Next, these data are analyzed to build models of the relationships between the inputs and outputs. These models are then used in conjunction with a cost function to select inputs that minimize the cost. Finally, run-to-run control models and a model update strategy are created for the optimized process. An integrated controller is implemented, and a series of runs are performed to test its response. Sections II–IV explore the three stages required for demonstrating epi deposition uniformity control. Sections V and VI present results and conclusions.

II. SELECTION OF PROCESS PARAMETERS AND DESIGN OF EXPERIMENTS

The success of this work depends on gathering information about the epi deposition process and sensor capabilities. The following sections outline an ordered methodology for selecting process parameters of interest and gathering experimental data. Section II-A describes guidelines for determining which sensor outputs will be monitored. Section II-B covers the subsequent selection of equipment settings (inputs). Section II-C discusses the development of a DOE to capture relations between the inputs and outputs.

A. Selecting System Outputs

The primary goal of this work is to demonstrate run-to-run uniformity control of epi film thickness and resistivity via in-line and off-line sensor measurements. The in-line thickness measurements, which are taken while the wafer is in the cool-down chamber, represent the only outputs that are available for true run-to-run control (where control decisions are made after each and every wafer run). Off-line measurement stations can be used for additional lot-to-lot control.

The system enables multiple (N -site) in-line thickness measurements to be taken along the wafer's diameter, instead of just a single center point. We also use an off-line thickness measurement station that provides a densely sampled wafer map. While the full wafer map measurements are not available for run-to-run control, they can be used in a lot-to-lot feedback loop

to make run-to-run control more effective. They are also important for evaluating the system's ability to perform run-to-run control based on only the in-line measurements.

Resistivity readings are available from an off-line measurement station; there is currently no *in situ* or in-line sensor. Thus resistivity must be controlled in a lot-to-lot fashion for this work.

Wafer rotation during epi deposition yields highly radially symmetric films, both in thickness and resistivity. Thus wafer map measurements may be compressed (averaged) into radial scans, as shown in Fig. 1. A model-based controller can update process settings between runs to control the radial uniformity of thickness and resistivity, rather than simply controlling their averages.

B. Selecting System Inputs

After determining the sensor readings (outputs) that will be monitored, the machine settings (inputs) that are most likely to affect those outputs must be selected. This does not imply that the exact relation between the inputs and outputs is known. There should be either empirical or theoretical evidence that changes in the inputs correlate with changes in the outputs. Preliminary work was done by Applied Materials to determine which process settings are likely to affect the epi film thickness and resistivity. These factors include those listed in Table I.

C. Developing a DOE

Once the system outputs and inputs are chosen, a set of designed experiments is selected and performed to model the relations between the inputs and outputs. First, "reasonable" operating ranges for the inputs are selected. Specifically, upper and lower bounds are determined, which define a "hyper-box" within the input space. These bounds are based on a number of criteria, including the following:

- Machine limitations;
- Safety limitations;
- Undesirable effects on uncontrolled outputs;
- Settings beyond which the resulting cost function is known to increase.

This last criterion implies some knowledge from previous experimentation and the process optimization methodology. Ide-

TABLE I
INPUT FACTORS AND DOE LEVELS

Factor	Lower axial ($-\alpha$)	Lower (-1)	Center (0)	Upper (+1)	Upper axial ($+\alpha$)
Deposit time (sec)	N/A	N/A	50	N/A	N/A
Deposit temperature ($^{\circ}\text{C}$)	1090 (-1.500)	1100	1120	1140	1150 (1.500)
Dopant mixing ratio (%)	20 (-1.500)	30	50	70	80 (1.500)
Dopant main flow (sccm)	40 (-1.571)	80	150	220	260 (1.571)
% Lower power	45 (-1.667)	47	50	53	55 (1.667)
% Inner power	40 (-1.429)	43	50	57	60 (1.429)
Dilutant (H_2) flow (slm)	30 (-1.500)	35	45	55	60 (1.500)
Trichlorosilane (TCS) flow (slm)	10 (-1.667)	10.8	12	13.2	14 (1.667)
Center gas flow valve (Accusett inner setpoint)	115 (-1.667)	123	135	147	155 (1.667)
Outer gas flow valve (Accusett outer setpoint)	60 (-1.667)	68	80	92	100 (1.667)

ally a set of optimal outputs can be achieved with machine settings that are near the center of the bounded input space. Table I also lists the operating ranges for the DOE.

A Central Composite design is selected for this work, as it is both commonly used and effective for response surface modeling. Noting that deposit time is used as a multiplicative factor on estimates of deposition rate (a process modeling strategy that is described in more detail later), there are nine factors under consideration for modeling the deposition rate and resistivity. The Central Composite DOE calls for a two level full-factorial design for all of the input factors, plus axial and center points. This means that a nine factor DOE requires over 512 experiments. Ideally all of these experiments could be performed, but processing that many wafers is probably unnecessary. A one-quarter fractional factorial design is used, which can be augmented later with more experiments, if needed.

III. PROCESS OPTIMIZATION

The process optimization methodology is based on the input and output parameters that have been selected. In the most generic case, a cost function is created in terms of the outputs (and possibly the inputs), which defines the relative quality of a set of outputs (and inputs). Specifically, a lower cost implies a better set of process parameters, meaning that a globally minimum cost is the desired operating point.

The following sections describe the strategies and issues involved in process optimization for the MIMO epi control problem. There are three main steps involved: selection of an appropriate cost function (Section III-A), empirically modeling the cost function in terms of the process inputs (Section III-B), and performing a constrained optimization over the cost model to find the inputs that yield the lowest cost (Section III-C).

A. Cost Functions

Often the cost is based primarily on the outputs and only slightly, if at all, on the inputs. In the epi control problem, a minimum cost solution should occur when all measured sites have

achieved the target thickness and resistivity, and all of the inputs are within the valid operating ranges. A constrained optimization capability eliminates the need to worry about input ranges, so the cost function need only be concerned with the optimality criteria.

Cost functions take many forms, but a weighted sum of squared errors from the output targets is extremely common and often analytically meaningful. A general form for this is

$$\text{Cost} = \sum_i [W_i (T_i - o_i)]^2. \quad (2)$$

This function sums over all outputs (i), the product of a weighting term (W_i^2) and the squared difference between the output target (T_i) and the actual output (o_i). The weighting terms scale the importance of meeting various output targets. The weights could reflect trade-offs between comparing outputs of different types, or trade-offs between the relative importance of outputs that are the same type.

There are many possible cost functions for the epi control problem, but

$$\text{Cost} = \sum_i [W_{ti} (T_{ti} - t_i)]^2 + \sum_j [W_{\rho j} (T_{\rho j} - \rho_j)]^2 \quad (3)$$

is a convenient and meaningful choice. T_{ti} and $T_{\rho j}$ are the target thickness and target resistivity, respectively for each of the thickness (i) and resistivity (j) measurement sites, while t_i and ρ_j are the measured thickness and resistivity values, respectively. W_{ti} and $W_{\rho j}$ are relative weighting constants between the individual thickness and resistivity site measurements. The weights encapsulate the trade-offs between achieving target thickness uniformity versus achieving target resistivity uniformity. Any differences between the number of thickness and resistivity sites will affect the weights, as well as the relative magnitudes of target thicknesses and target resistivities.

The weights are also likely to be affected by the spatial organization of the outputs. There is a weighting structure based on the relative area that each output site represents, as shown in Fig. 2. For the epi control problem we make use of data from points along a radius. Clearly, sites near the outer edge represent a larger area than those near the center. Therefore it makes sense to associate a higher cost for these sites when they deviate from the target. Using these weights is analogous to finding a cylindrical shell approximation for error volume. Better approximations may be obtained using thin plate spline error volume estimation techniques, as presented in [19]. However, the use of a dense sampling pattern can eliminate the need for a more complicated weighting scheme.

B. Cost Function Modeling

The optimization process makes use of a function (model) of the cost in terms of the inputs. This function is back-solved to find a minimum cost solution. DOE data are used to fit a model of the cost as a function of the inputs. There are many ways to do this, especially when the system outputs include multiple site measurements. One approach involves empirically modeling the cost directly as a function of the DOE input settings. This is

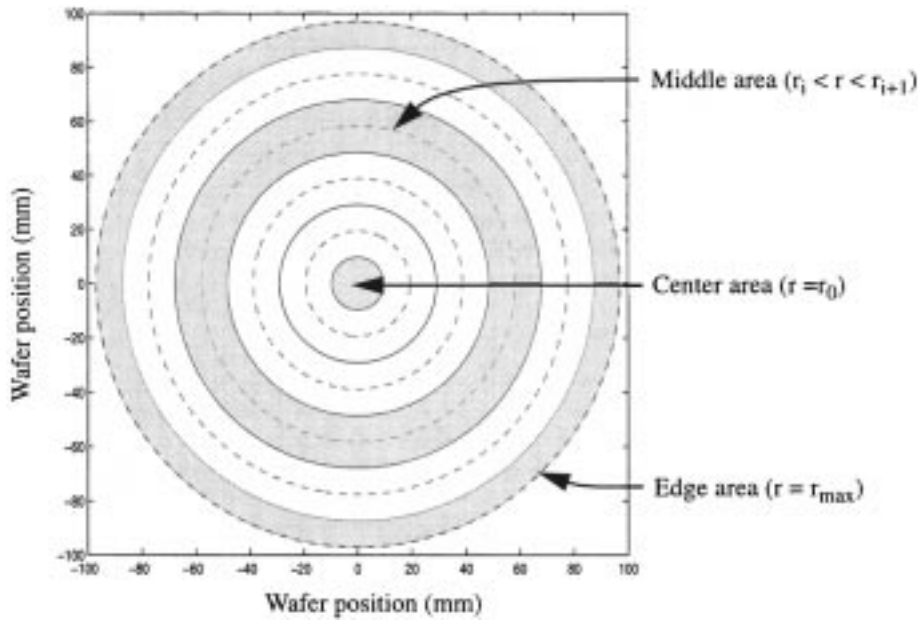


Fig. 2. Sample output weighting based on representative area.

known as a single response surface (SRS) methodology, which would simply generate the cost function

$$\text{Cost} = c(\mathbf{x}) \quad (4)$$

where $\mathbf{x}$ is the vector of process settings. That is, the model does not receive any information from individual sites; the cost is modeled directly in terms of the inputs.

A multiple response surface (MRS) methodology builds individual models of each site, then uses the cost function's form or expression to combine the outputs of the models. The weighted MRS cost function for the epi process becomes

$$\text{Cost} = \sum_i [W_{ti} (T_{ti} - t_i(\mathbf{x}))]^2 + \sum_j [W_{\rho j} (T_{\rho j} - \rho_j(\mathbf{x}))]^2. \quad (5)$$

There are a number of issues involved in choosing between an SRS and an MRS strategy, some of which are described in [4], [5], [20]. While one must be careful in using MRS strategies, this work makes use of MRS models primarily because they are required for implementing uniformity control.

Time-Based Control: An important modeling issue is the effect of time-based settings. For thickness measurements it is often more effective to model the etch or deposition rate, then multiply by time to arrive at final thicknesses. This is better than trying to include time as a "regular" input and create a (linear) model that includes time as an additive term rather than the multiplicative factor it actually is. This added twist results in a cost function of the form

$$\text{Cost} = \sum_i [W_{ti} (T_{ti} - (r_i(\mathbf{x}) \cdot d))]^2 + \sum_j [W_{\rho j} (T_{\rho j} - \rho_j(\mathbf{x}))]^2 \quad (6)$$

where $r_i(\mathbf{x})$ is the model of the i th site's deposition rate as a function of the inputs, and d is the deposition time. It is assumed that deposition time does not affect resistivity in the same

manner. Within reasonable ranges, the final epi resistivity will be essentially independent of the film thickness, and thus independent of the deposition time. This cost function model should achieve a fairly accurate representation of the process and will reasonably map into a linear model-based run-to-run control strategy. The complications involved with using time-based inputs for run-to-run control will be introduced in Section IV.

C. Cost Function Optimization

Using a MRS model of the cost function in terms of the process settings and target outputs, a set of inputs that yield a minimum cost are found through a constrained optimization. While the run-to-run controller will use linear models of deposition rate and resistivity, the DOE-based process optimization could use higher order models. A general optimization process such as this is provided by a number of software packages. In particular, we use MATLAB for modeling and optimization. The details of constrained optimization are beyond the scope of this paper and are expected to be automated through the use of a software package. This applies to the initial process optimization step, but the integrated run-to-run controller contains its own constrained optimization routines for the linearized, time-based models.

IV. RUN-TO-RUN CONTROL

Having selected process parameters, performed a DOE, and created and optimized a set of site models, the run-to-run controller is initialized to begin processing. As with previous run-to-run control work at MIT, linear process models whose offsets are updated in an EWMA manner are used in the controller. However, the deposited thickness site models have been modified to take advantage of the time-based structure described earlier.

Because this controller is used for our run-to-run control experiments, it is worthwhile to note that the run-to-run process

models might require a DOE from a smaller, more linear region of the input space. This will be the case if the original DOE space requires highly nonlinear models.

Even before the DOE is specified and analyzed, the run-to-run controller's cost function and model update procedure are considered. It is important to match the cost function and optimization strategies between initial DOE-based optimizations and the subsequent run-to-run control optimizations. Section IV-A looks at cost function optimization for the run-to-run controller, Section IV-B discusses the controller's model update strategy, and Section IV-C describes the issues involved in combining lot-to-lot and run-to-run feedback data.

A. Cost Function Optimization

The run-to-run controller's cost function modeling structure should be identical to that shown in Section III (6):

$$\text{Cost} = \sum_i [W_{ti} (T_{ti} - (r_i(\mathbf{x}) \cdot d))]^2 + \sum_j [W_{\rho j} (T_{\rho j} - \rho_j(\mathbf{x}))]^2. \quad (7)$$

It is important to carry this cost function into the run-to-run controller to insure a proper "hand-off" from the DOE-based process optimization. However, the site models for the run-to-run controller, $r_i(\mathbf{x})$ and $\rho_j(\mathbf{x})$, are restricted to be linear, as in

$$r_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x} + b_i \text{ and } \rho_j(\mathbf{x}) = \mathbf{A}_j \mathbf{x} + b_j. \quad (8)$$

This restriction implies that the run-to-run site models are linearized versions of the DOE optimization models. Ideally the linear models capture (nearly) all of the important behavior from the DOE models. If they fit well, linear models may be generated directly from the DOE data. If this is not the case, the nonlinear models can be linearized at an optimal operating point, and the allowable input space is truncated such that the linear models are a "good enough" fit for that region, as defined by the desired stability criteria. (This is the case where a follow-up DOE is recommended.)

For each run, the controller must be able to select inputs $\mathbf{x}$ and d that optimize the cost function, given the current site models. Software routines have existed in the MIT controller for quite some time that can efficiently minimize a weighted sum-of-squared-error cost function based on linear site models. However, the inclusion of time-based models implies that these routines cannot be directly applied to the optimization problem.

The optimization is currently solved by holding one set of inputs (either $\mathbf{x}$ or d) constant, then solving for the other. When the deposition time, d , is held constant, the "normal" run-to-run control solver may be used to find the process settings, $\mathbf{x}$. The weighting and target terms are first adjusted to account for the time factor (d). Specifically,

$$\hat{W}_{ti} = d \cdot W_{ti}, \text{ and } \hat{T}_{ti} = \frac{T_{ti}}{d}. \quad (9)$$

Once the vector $\mathbf{x}$ is chosen, the optimal deposition time may be found by minimizing the cost function with respect to d . In this second stage of the optimization, the inputs ($\mathbf{x}$), and thus

the rates ($r_i(\mathbf{x})$) are known and can be treated as constants. This leads to the following straightforward solution for d :

$$d = \frac{\sum_{i=1}^N \hat{W}_{ti}^2 \cdot \hat{T}_{ti} \cdot r_i}{\sum_{i=1}^N \hat{W}_{ti}^2 \cdot r_i^2}. \quad (10)$$

Note that the optimal deposition time is only a function of terms corresponding to deposition rate outputs. Intuitively, resistivity models should not affect the selection of a deposition time when all other inputs are held constant.

In fact, a complete nonlinear optimization scheme is achieved by iterating between the solution of optimal nontime- inputs $\mathbf{x}$ (while holding constant the deposition time), and the solution of an optimal deposition time d (while holding constant the nontime inputs). Essentially this strategy guarantees that each step decreases the cost function toward a minimum, which could be local or global. A minimum will be found unless process setting bounds are hit first. A full development of solvers for these types of time-based models is found in [5].

B. Model Updates

As described earlier, the run-to-run controller's dynamic site models consist of affine functions of deposition rates and resistivities in terms of the equipment settings. The models contain a slope matrix, $\mathbf{A}$, and a vector of offset terms, b , as in (1). The slope matrix is determined from the DOE and is not modified by the EWMA controller.

The controller tracks output disturbances and drifts by updating the offset terms. After a set of measurements is presented to the controller, it corrects for model errors by changing the offset for each output model. The EWMA weights (λ) determine the amount of model update after each run (n), based on the equations

$$b_i[n] = \lambda_i \cdot \hat{b}_i[n] + (1 - \lambda_i) \cdot b_i[n-1], \hat{b}_i[n] = y_i[n] - \mathbf{A}_i \mathbf{x}[n] \quad (11)$$

where $y_i[n]$ is the measured output after processing a wafer with equipment settings $\mathbf{x}[n]$. These update formulas are used for both the deposition rates (y_i) and the resistivities (y_j).

Optimal EWMA weights for a given process are selected as a function of the statistical noise in the output measurements and the amount of actual deterministic drift or disturbances the system exhibits; a thorough treatment is found in [1]. Estimates of output noise and drift should be determined by the DOE runs and historical data from the equipment. However, for demonstration of the controller's reaction to a staged disturbance event, the controller typically uses a "large" EWMA constant for each output (e.g., as in the SISO run-to-run control demonstration, where λ was set to 0.7).

C. Combining Lot-to-Lot and Run-to-Run Feedback Data

An important issue for run-to-run control is the combined use of in-line (run-to-run) and off-line (lot-to-lot) sensors. Measurements from the DOE give a full set of data for each run, but many of these will not be available on a run-to-run basis in an actual production setting. Model update strategies and feedback loops

must work together to ensure that sensor data are used effectively.

Consider the epi thickness measurements, which are available from two different sensors. An in-line thickness sensor provides a few measurements along a single diameter, yielding an even smaller number of radially averaged site measurements. An off-line thickness sensor provides a densely sampled wafer map, yielding many radial site thickness averages. In other words, one sensor samples more densely in time, while the other sensor samples more densely in space. The question then arises, how should the modeling and control strategies make use of combined feedback from these two sensors?

More generally, this problem presents itself when multiple sensors provide correlated measurements at different sampling rates. If the sensors return correlated data at the same time intervals, then appropriate strategies can be applied to extract important features before the modeling steps are performed. If the sensors return uncorrelated data, then the modeling steps are performed independently and do not rely on the availability of both measurements at the same time.

Note that the resistivity measurements fall into the latter category; resistivity is not correlated with film thickness. An epi film of a given resistivity can exist for essentially any thickness, and vice versa. Therefore, given information about one characteristic, we can make no (or very few) assumptions about the other. Thus, the resistivity site models are updated only when resistivity data come back, which is on a lot-to-lot basis.

The two different thickness measurement systems, however, provide the more difficult case. Measurements of the same sites on the same wafers are almost perfectly correlated. However, the in-line sensor provides thickness readings for wafers that the off-line sensor does not see, and the off-line sensor provides thickness readings for sites that the in-line sensor does not see. How does one effectively use the combined information provided by these two systems? There are many possible ways to deal with this problem, but one of the most promising techniques involves creating a single set of models that can be updated by either sensor's output.

Joint Sensor Monitoring: The concept of "joint sensor monitoring" is introduced, where a single set of output models is maintained for all sensors with correlated measurements. Broadly speaking, the goal is to create models and model update strategies that do not require all of the measurements to be available at once, yet capture the important information from each sensor.

Implementing such a strategy is aided by first considering an ideal scenario, where all sensor data are available at the same time, and at the highest measurement frequency provided by any of the sensors. As mentioned earlier, this type of system presents a case where there is no longer an issue of combining lot-to-lot and run-to-run data. The simplified problem requires the engineer to decide what output models to build, and how to generate those outputs from the combined set of measurements. This is the type of analysis performed in Section II-A, when it was decided to build models of radial sites. These virtual outputs are not directly provided by the sensor, but are formed by compressing the full set of measurements into a smaller set of outputs that are less noisy and more controllable. Generally, the

same types of analyses can be used to merge correlated measurements into a set of virtual output signals. Ideally the virtual outputs exhibit better noise characteristics than the pure measurements and capture only the features that can actually be manipulated or controlled.

For the epi thickness modeling and control problem there are two different sets of outputs, one is a densely sampled radial thickness profile, while the other is a lightly sampled version of the same profile. Essentially the in-line sensor provides a subset of the spatial information from the off-line sensor. If one is to consider a situation where both of these sensors provide data after every run, then we would clearly keep the densely sampled profile and simply average in or ignore the lightly sampled profile. Thus, the selection of virtual outputs is quite simple; we will keep the densely sampled radial profile.

Given a mapping from sensor measurements to virtual outputs, the next step is to define appropriate strategies for mapping individual sensor measurements to the full set of virtual outputs. Presumably this would be difficult to do for some or all of the sensors. Remember that the measurements must be correlated in some manner for them to be grouped together, so knowing the outputs from one sensor does provide information about what the other sensor(s) would measure. Based on the methods for generating virtual outputs, one should be able to create a highly probable set of combined outputs, given the measurements from a single sensor. However, the correlation could be weak, so generating a set of virtual outputs based on only one sensor will likely introduce noise and errors. This cannot be avoided.

We will see that there is a way to improve upon this basic strategy, but for now consider formulating the epi thickness control problem using these concepts. First, it is easy to see what should happen when a full radial scan is produced by the off-line sensor. The virtual outputs are simply equal to the sampled radial profile, since that's what they were defined to be. The interesting problem is how to generate a full radial scan from the small number of sites sampled by the in-line sensor. A logical and relatively simple suggestion is to simply "connect the dots" and generate a piecewise linear output extrapolation of the available sites, as demonstrated in Fig. 3. More advanced extrapolations, such as those using splines, might also be considered [19].

After generating a new set of virtual outputs from the available sensor data, the corresponding output models are updated. The question is, does the methodology described above really utilize all of the available information for generating virtual outputs? There is in fact another source of information about the relative thicknesses of the measured and unmeasured radial sites. Weighted histories for all of the thickness outputs are contained in the current thickness models. While the above derivation for extrapolating thicknesses is reasonable, given that nothing else is known, it does not take into account information contained in the current models.

This insight leads to a more effective update strategy: given a set of input parameters, a set of output models, and a mapping from sensor measurements to virtual outputs, we must define appropriate methods for mapping individual sensor measurements to a full set of virtual outputs. Built into the current models are past measurements from all of the sensors, plus the expected relations between the various inputs and outputs of the system.

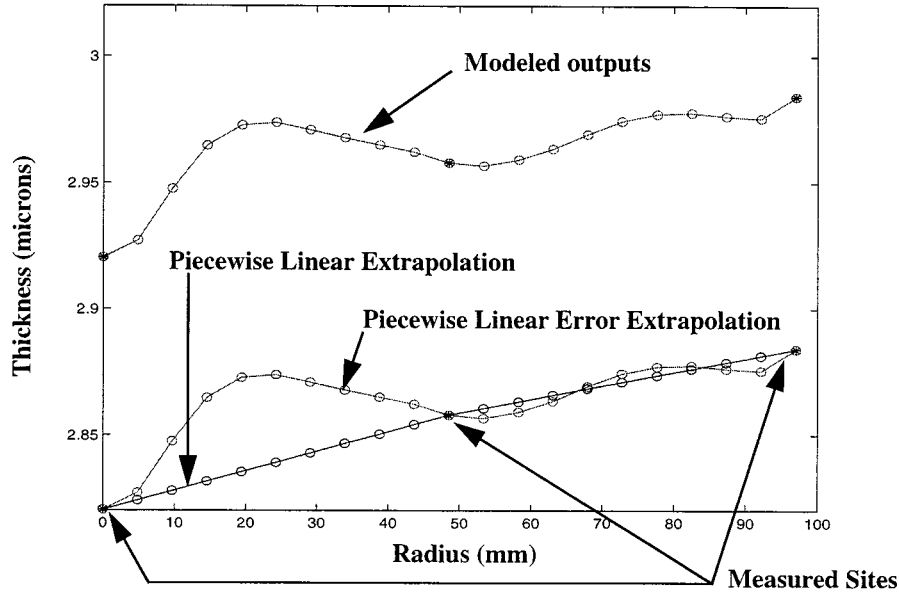


Fig. 3. Example of virtual radial profiles generated from in-line measurements (and process models).

Clearly, an output estimator should take these models into account when extrapolating virtual outputs from a subset of measurements.

Algorithms for generating virtual outputs are highly dependent on the structure of the problem. Additionally, there are likely to be many reasonable and mathematically sound methods available for any given case. Consider the example in Fig. 3, which shows the measured outputs, modeled outputs, and two types of virtual outputs. Notice that the original piecewise linear extrapolation assumes straight line profiles between the measured sites, ignoring information provided by the model. Over time, this type of strategy gradually eliminates any information that the models provide about unmeasured sites. Between updates from off-line measurements, this system becomes one that simply attempts to maintain the subset of outputs that are measured by the in-line sensor. This will be true regardless of the output extrapolation technique.

A better strategy for extrapolating the full set of outputs is to assume that the modeled outputs provide deterministic components, while the measured sites provide information about the deviation, or error. Instead of directly using the piecewise linear extrapolation of the measured sites, a piecewise linear extrapolation of the deviation (error) can be added to the modeled outputs. Fig. 3 also shows what the new set of virtual outputs looks like for the given example. Instead of virtual outputs with the form

$$\mathbf{x}_v = f(\mathbf{x}_m) \quad (12)$$

we have

$$\mathbf{x}_v = \hat{\mathbf{x}}_v + f(\hat{\mathbf{x}}_{v=m} - \mathbf{x}_m) \quad (13)$$

where $\mathbf{x}_v$ are the virtual outputs, $\mathbf{x}_m$ are the measured outputs, and $\hat{\mathbf{x}}_v$ are the modeled (expected) outputs. This error extrapolation strategy has a number of useful features:

- The error surfaces (curves) often exhibit smaller (in magnitude) and/or lower frequency variations than the

actual output surfaces. Thus fewer samples can be used to achieve a given level of confidence, and simpler extrapolation techniques (like piecewise linear methods) can be used.

- The exact in-line measurements are still part of the extrapolated virtual outputs. This is important because the thickness sensors (both in-line and off-line) are very accurate. The predicted (modeled) outputs should not influence these sites.
- This strategy uses (presumably) more accurate assumptions about the relative spatial behavior between the measured sites. It relies primarily on the model, and thus the information garnered from the DOE and previous off-line sensor measurements, to help fill in the gaps where there are no measurements.
- Finally, it is easy to see from the last figure that the error extrapolation method responds appropriately when there is a simple mean shift across the entire profile. If all of the in-line measurements differ from the model predictions by a fixed amount, then this strategy simply shifts the profile so that the measured sites line up properly, but the shape is left unchanged. (As we will see later from the results of the DOE, this type of shift or drift is the largest type of disturbance found in the system.)

V. EXPERIMENTAL RESULTS

As described in Section II, a large set of designed experiments is performed to gather information about the system and build process models. This section provides a description and analysis of the designed experiments. Section V-A details the experimental design and execution, while Sections V-B–V-D analyze the resulting data in terms of process stability and noise, process model construction, and process optimization. Section V-E presents simulated run-to-run control scenarios using the resulting noise characteristics, time-based models for thickness, and linear models for resistivity.

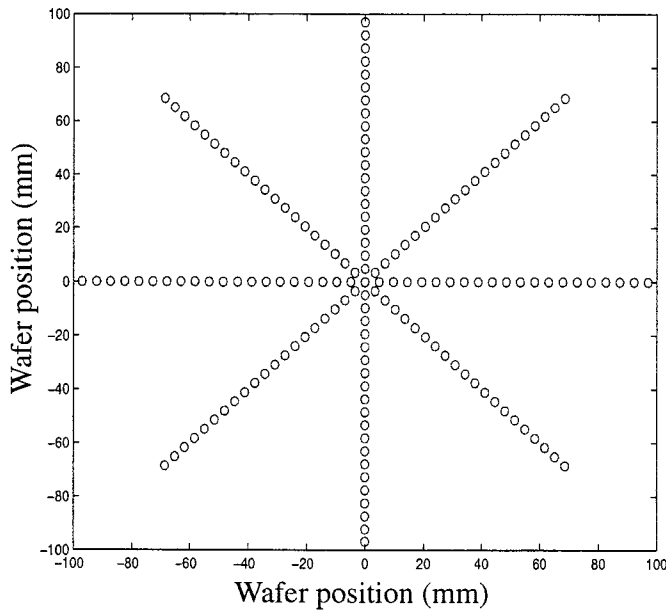


Fig. 4. DOE thickness measurement map.

A. Experimental Design and Execution

The DOE is composed of 128 two-level fractional factorial design points, 18 axial points, eight center points, and four additional replicates of a randomly chosen fractional factorial design point, for a total of 158 runs. There are nine tunable parameters, excluding deposition time, which is held constant across all of the experiments. Table I specifies the actual input levels that are used in the design. Input levels for the axial runs are set at the $\pm\alpha$ levels.

After depositing epi on the wafers, each are measured by off-line thickness and resistivity sensors. 164 thickness measurements are taken using four diameter scans with 41 sites each, as shown in Fig. 4. All measurements at each radius (four measurements for the center point and eight measurements for all other radii) are averaged to provide 21 radial site thicknesses. These Measurements are highly repeatable and accurate.

Due to repeatability and throughput issues for the available mercury probe, four repetitions of a single diameter scan (with 21 sites) are taken for resistivity measurements, as shown in Fig. 5. All four scans are reviewed and the "best" one is selected for each wafer. (Process engineers at Applied Materials are the primary source for determining which results are "best.") This scan is similarly averaged, where there is a single center point and two measurements for all other radii, resulting in 11 radial site resistivities. Only 10 sites are retained for analysis since the outermost resistivity measurements demonstrate large variations that are not well modeled. Due to sensor availability limitations, data for only 113 of the 158 experiments are used for model construction.

B. Process Stability and Noise

Process stability and noise are addressed by analyzing experimental replicates and monitor wafers. Pairs of center points are run at the beginning of each of the four DOE lots, yielding a total of eight center point replicates. There are also five replicates of a randomly selected fractional factorial design point, which are

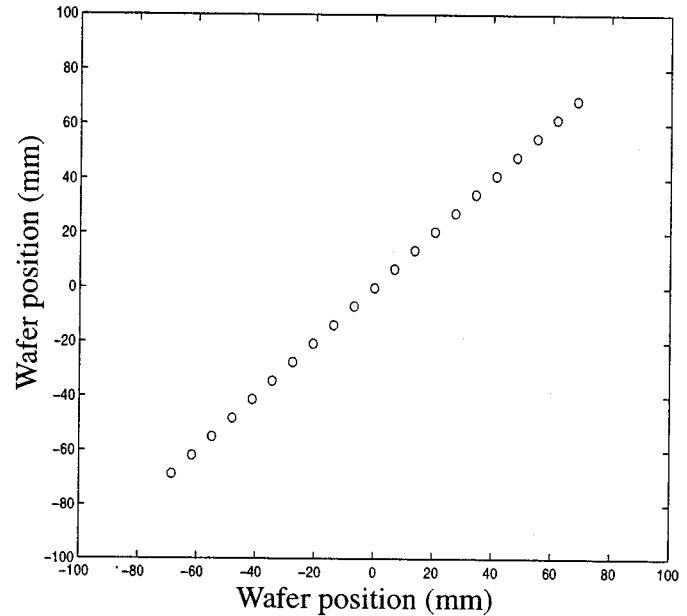


Fig. 5. DOE resistivity measurement map.

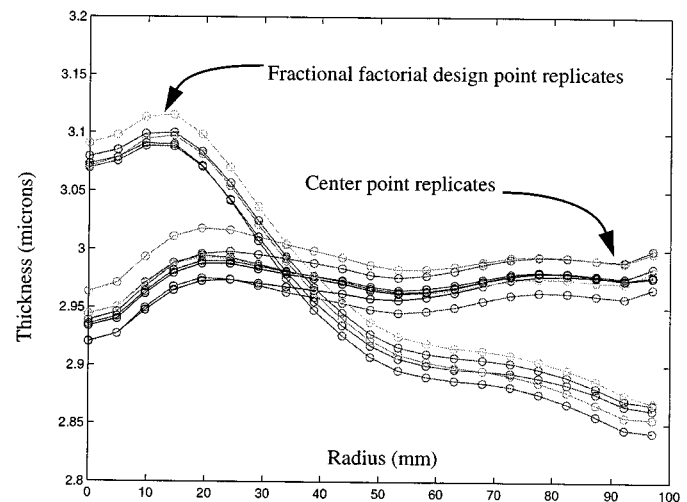


Fig. 6. DOE thickness measurements: replicates.

also run at the start of each lot. Monitor wafers are run before the start of each lot to provide information about process stability and drift during the course of the lots. Each monitor wafer is processed with the same settings that are used for the last wafer in its lot.

Examination of the measurements reveals that deposited thickness appears to be quite stable over time and after processing at a variety of equipment settings. Also, the data demonstrate considerable variation in thickness profiles across the different DOE process settings, indicating that effective process optimization and control are possible. The resistivity data show much noisier results, and the lack of a full data set limits the ability to compensate for any process shifts or drifts. The next two subsections consider process noise and stability for thickness and resistivity.

Thickness: Process noise is analyzed using the experimental replicates. Fig. 6 shows the radial thickness profiles for these replicates. The noise can be broken down into a mean offset

TABLE II
THICKNESS: NOISE ESTIMATES FROM THE DOE AND PROCESS MODELS

Estimate based on:	Type of Noise	Std Dev of Noise (microns)
Center Replicates	Mean shift noise ($\hat{\sigma}_\mu$)	0.0136 (0.46%)
Center Replicates	Site noise ($\hat{\sigma}_{site}$)	0.0041 (0.14%)
Fractional Factorial Repl.	Mean shift noise ($\hat{\sigma}_\mu$)	0.0104 (0.35%)
Fractional Factorial Repl.	Site noise ($\hat{\sigma}_{site}$)	0.0035 (0.12%)
1st Order process models	Mean shift noise ($\hat{\sigma}_\mu$)	0.0267 (0.90%)
1st Order process models	Site noise ($\hat{\sigma}_{site}$)	0.0096 (0.32%)
2nd Order process models	Mean shift noise ($\hat{\sigma}_\mu$)	0.0159 (0.53%)
2nd Order process models	Site noise ($\hat{\sigma}_{site}$)	0.0049 (0.16%)

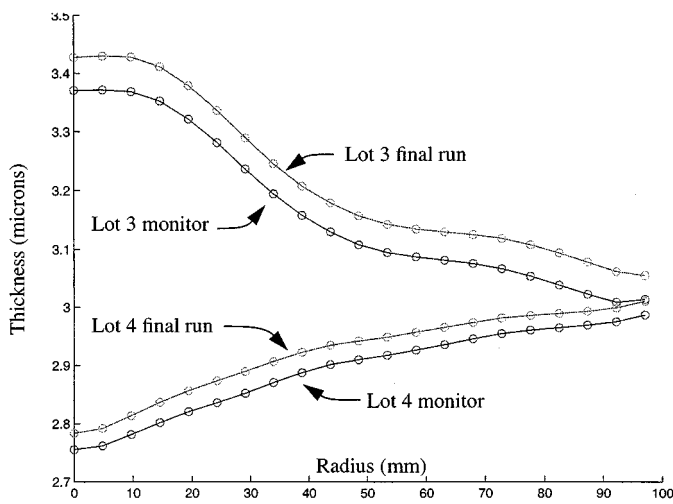


Fig. 7. DOE thickness measurements. Lots 3 and 4: monitor and final runs.

across all radial sites, combined with individual site noise. Table II shows the estimates for these values; clearly the mean offset noise is the most significant factor.

Monitor wafer data tend to demonstrate a small, but repeatable trend toward increasing deposition rate over the course of processing a lot. Fig. 7 shows the monitor and final run radial profiles for the third and fourth lots of the DOE. Profiles vary from center thick to center thin, indicating that there is great opportunity for process modeling, optimization, and control.

Resistivity: The lack of a complete data set for resistivity limits the ability to estimate noise characteristics for the system. Only six of the eight center point replicates and four of the five factorial replicates are available. Also, there is only a single lot whose monitor and final wafer are both valid.

Fig. 8 shows the data for the center point and fractional factorial design point replicates. Considering the scale for these plots, it is clear that there is much more variation in these outputs, both in terms of mean shift noise and individual site noise. Estimates of output noise are given in Table III.

The sample standard deviation of the noise appears to be considerably higher for resistivity than that found in the thickness measurements. It is unclear how much of the noise is due to the system (actual process noise) and how much is due to the sensor

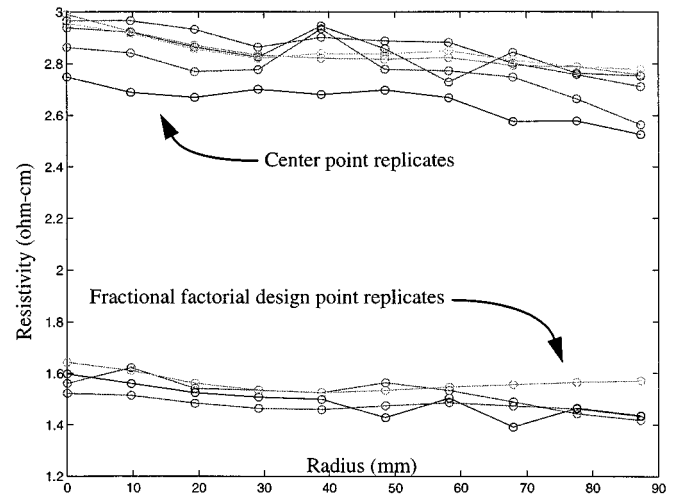


Fig. 8. Resistivity: Center point and factorial design point replicates.

TABLE III
RESISTIVITY: NOISE ESTIMATES FROM THE DOE AND PROCESS MODELS

Estimate based on:	Type of Noise	Std Dev of Noise (ohm-cm)
Center Replicates	Mean shift noise ($\hat{\sigma}_\mu$)	0.0805 (2.87%)
Center Replicates	Site noise ($\hat{\sigma}_{site}$)	0.0376 (1.34%)
Fractional Factorial Repl.	Mean shift noise ($\hat{\sigma}_\mu$)	0.0389 (2.57%)
Fractional Factorial Repl.	Site noise ($\hat{\sigma}_{site}$)	0.0322 (2.13%)
1st Order process models	Mean shift noise ($\hat{\sigma}_\mu$)	0.6117 (18.0%)
1st Order process models	Site noise ($\hat{\sigma}_{site}$)	0.1635 (4.8%)
2nd Order process models	Mean shift noise ($\hat{\sigma}_\mu$)	0.3117 (9.16%)
2nd Order process models	Site noise ($\hat{\sigma}_{site}$)	0.1481 (4.35%)

(measurement noise). The difficulty in getting the sensor to provide repeatable measurements indicates that a substantial part of the noise is probably introduced by the mercury probe.

C. Process Model Construction

Models for each of the 21 radial deposition rate sites and the 10 radial resistivity sites are generated via a statistically driven polynomial response surface modeling methodology. (Remember that the 21 radial thickness sites use the time-based model, where deposition rate multiplied by deposition time yields final thickness.) First- and second-order models are considered for the DOE's deposition rate and resistivity data.

Deposition Rate: Each of the 21 deposition rate (site) models is a direct function of all nine (nontime) inputs. Based on the ANOVA information from the resulting linear deposition rate models, the inputs that most strongly affect the outputs are Center gas flow valve, Outer gas flow valve, deposition temperature, H₂ Flow, and TCS Flow. Figs. 9 and 10 show example plots of the modeled deposition rates as functions of radius and the gas flow valves. For each simulation, the selected input is varied throughout its DOE range, while the remaining eight process settings are held at their DOE center point values.

Notice that the rate as a function of radius is highly nonlinear for the plots. However, we must remember that the site models

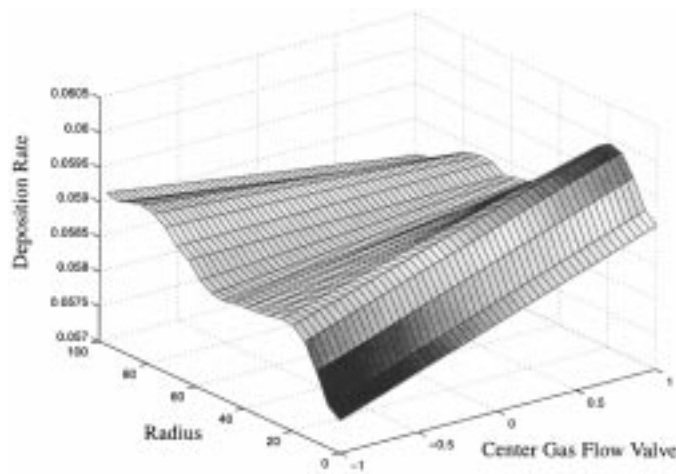


Fig. 9. Dep Rate: Uniformity as a function of center gas flow valve.

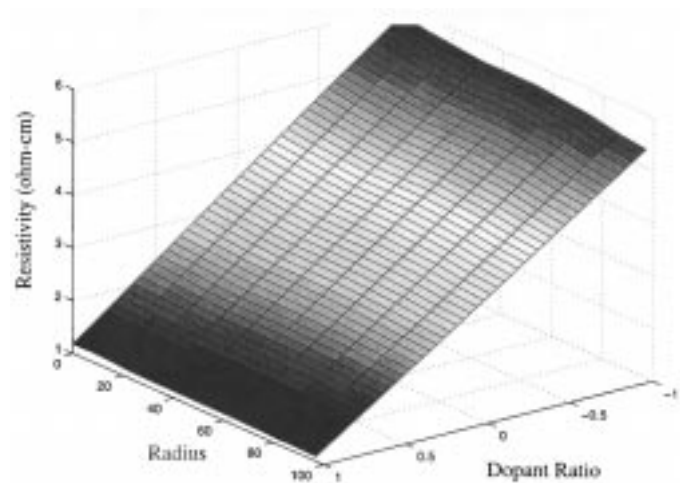


Fig. 11. Resistivity: Uniformity as a function of dopant ratio.

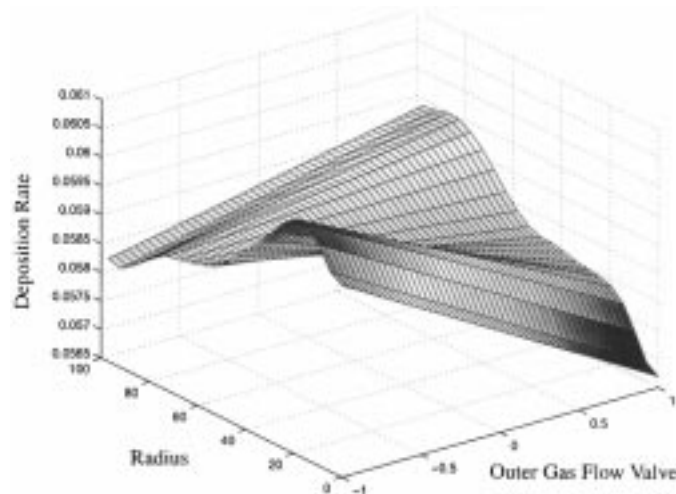


Fig. 10. Dep Rate: Uniformity as a function of outer gas flow valve.

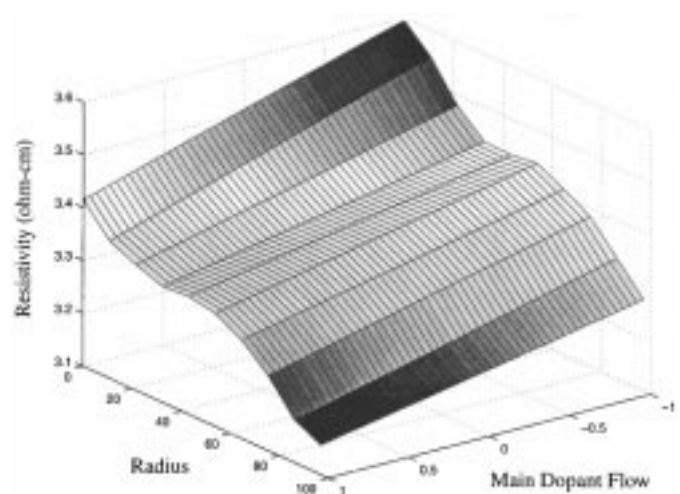


Fig. 12. Resistivity: Uniformity as a function of main dopant Flow.

absorb this nonlinearity. The outputs do not directly model the rate as a function of radius; each radial site is a separate model. Thus, in these two figures, the models only contain terms from the process settings and not the radius.

The quality of these deposition rate models can be considered using a number of different criteria. First, an (ANOVA-based) R-squared analysis provides estimates for the percentage of output variation that is accounted for by the models. The affine models capture about 97.2% of the variation, and the second-order models capture about 99.1%. These results also show that the second-order models represent the DOE data extremely accurately. However, it is more important to note that the affine models also provide excellent modeling capability, suggesting that the proposed time-based controller with affine rate models will work well.

These models also provide a means for estimating the amount of noise in the system. Model error for each DOE run can be broken down into mean shift noise and site noise (see Table II), as was done with the noise estimates from the DOE replicates. The second-order models demonstrate noise characteristics that are very close to those found in the replicate data. For this reason, the second-order models provide an ideal

level of model complexity. That is, higher order models could better fit the data, but they would probably be fitting to the noise in the system, and therefore providing worse process models. The first-order models show somewhat less than twice as much noise as the replicate-based estimates. This indicates that some model-based noise will be generated while using them for control, but we expect the system to perform adequately.

Resistivity: Each of the 10 resistivity (site) models is a direct function of all nine (nontime) inputs. Based on the ANOVA information from the resulting linear resistivity models, the inputs that most strongly affect the outputs are Dopant Ratio, H_2 flow, TCS flow, and deposit temperature. Dopant main flow, while not providing large coefficients for the resistivity models, is forced to be included because it is a well known control knob for fine-tuning resistivity. It should also be noted that Dopant Ratio has by far the strongest effect on the outputs, and that all others have substantially weaker effects than those seen in the deposition rate models. Figs. 11 and 12 show the modeled resistivity as a function of radius for two of the aforementioned inputs. (For viewing purposes, the directions for increasing radius and input factor settings are reversed from those used in the thickness uniformity plots.) For each simulation, the selected

input is varied throughout its DOE range, while the remaining eight process settings are held at their DOE center point values.

As with the thickness models, the resistivity modeling quality can be considered using the (ANOVA-based) R-squared analysis. With the available DOE data, the first-order (affine) models capture an average of 91.9% of the variation, while the second-order polynomials capture approximately 97.6%. These results show that neither first-order, nor second-order models represent the resistivity data as well as the thickness data. There are any number of possible reasons, such as high measurement or process noise, or the response surface might simply require a high-order model. However, with only 113 (or even 158) data points it is impossible to build a generic third-order model for this system. While the second-order models do provide better data fitting capability, it is hoped that the first-order models will be good enough for closed loop control.

Once again, model error from the DOE runs is broken down into mean shift noise and site noise for comparison against the replicate-based noise estimates. These results are shown in Table III. As one might suspect from the R-squared analysis, both the first- and second-order models provide relatively noisy estimates of the DOE data. The noise is larger than that predicted by the replicates. This information gives us the opportunity to run simulations with models of different fidelity and systems with different noise characteristics.

D. Process Optimization

The resulting site models for deposition rates and resistivities are combined with a cost function to model the cost in terms of the ten process settings. As presented in Section III in (6), the cost function is

$$\text{Cost} = \sum_i [W_{ti} (T_{ti} - (r_i(\mathbf{x}) \cdot d))]^2 + \sum_j [W_{\rho j} (T_{\rho j} - \rho_j(\mathbf{x}))]^2. \quad (14)$$

This function uses time-based models for thickness and may be combined with any type of underlying deposition rate and resistivity models, such as the first- and second-order polynomials described above. Constrained optimizations of the cost function provide optimal operating points for initiating run-to-run control experiments. The bounds for each input are set to the factorial design limits (± 1) shown in Table I, which are presumed to contain the input space where the models are most accurate.

We find that the first-order models lead to a unique solution within the bounded region, while the second-order models result in three different local minima. Second-order models are likely to provide better simulated results because they are able to find lower cost solutions. However, second-order models make the optimization problem much more difficult due to increased complexity and computational demands, plus the increased likelihood of obtaining multiple locally optimal solutions.

E. Run-to-Run Control Simulations

To demonstrate control, thickness and resistivity models are initialized to uniform radial profiles at target values. Process

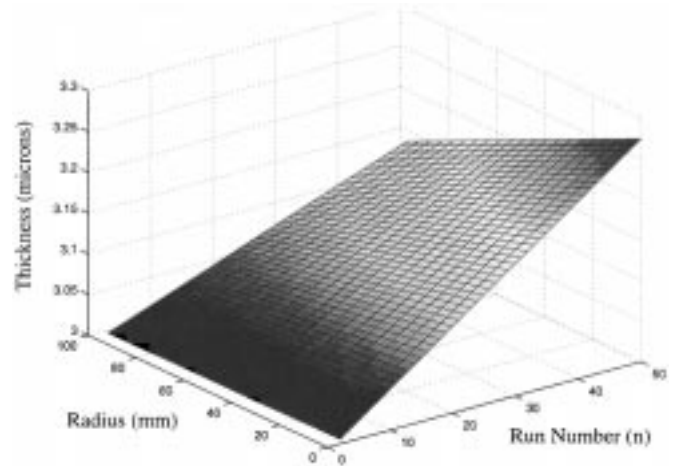


Fig. 13. Noiseless system: Uncontrolled thickness.

drift is added to the simulated system, while the controller attempts to maintain target output values. This scenario is considered first with a system that presents no noise to the controller and allows the controller to apply input changes with very fine (nearly continuous) precision. Next, the controller acts on the same noiseless system, but is only allowed to use discretized inputs, where the discretization levels are selected to match those available on the actual processing equipment. The controller is then tested with a noisy system, where the process noise is similar to that found through the DOE replicate data. This system is again controlled using both continuous and discretized inputs.

The Noiseless Continuous System: The time-based controller is first tested with a drifting system that has no process or sensor noise. In addition, the controller is allowed to select inputs from a (nearly) continuous set of input levels. The noiseless, continuous system provides a best case scenario, or benchmark, for how well the system could perform if the process models and sensors provide perfect information.

For convenience, the target thicknesses and resistivities are uniformly set to 3.00 microns and 3.36 $\Omega\cdot\text{cm}$, respectively. These values are selected to be near the average outputs predicted by the process models when the center point inputs are applied to the system.

Fig. 13 shows a drifting process where the uncontrolled thickness becomes gradually thicker with time (run number). Note also that sites near the center drift more quickly than those near the edge. Thus the films are gradually becoming thicker and less uniform. The full multi-objective controller is demonstrated by simultaneously engaging a drifting resistivity model in the system. Fig. 14 shows a drifting process where the mean resistivity across the whole wafer increases linearly with time, while the uniformity remains constant. It is important to note that the drifting effects shown in Figs. 13 and 14 are taking place simultaneously, and thus the controller is compensating for both drifts simultaneously as well. Figs. 15 and 16 show the measured thickness and resistivity uniformities achieved by using the run-to-run controller, when applied to this drifting process. The controller uses EWMA weights of 0.7 for all outputs.

As shown by Fig. 15, the controller is able to correct for the increasingly nonuniform thickness and maintain outputs very

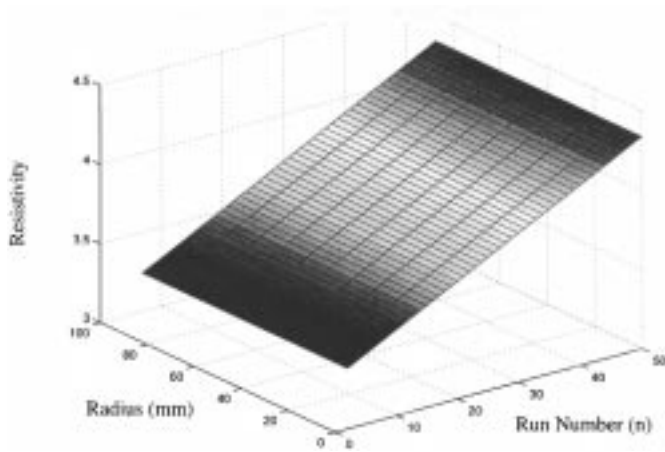


Fig. 14. Noiseless system: Uncontrolled resistivity.

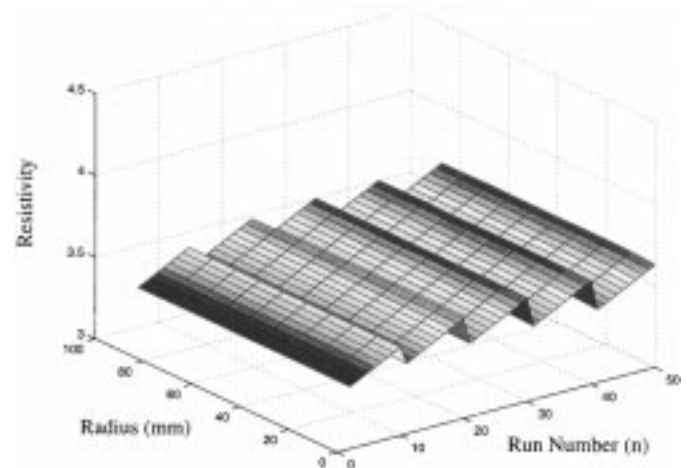


Fig. 16. Noiseless continuous system: Controlled resistivity.

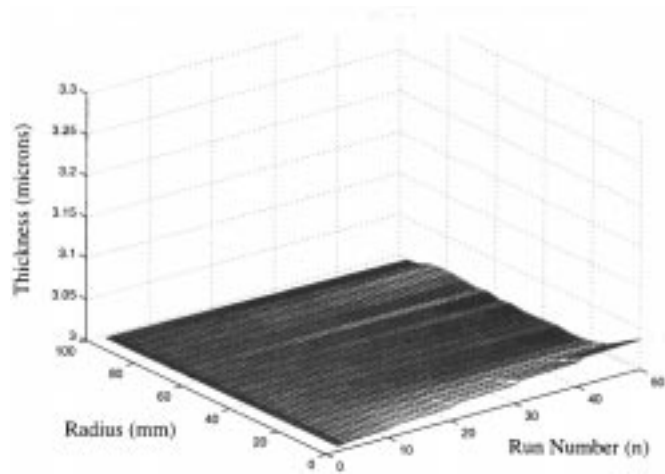


Fig. 15. Noiseless continuous system: Controlled thickness.

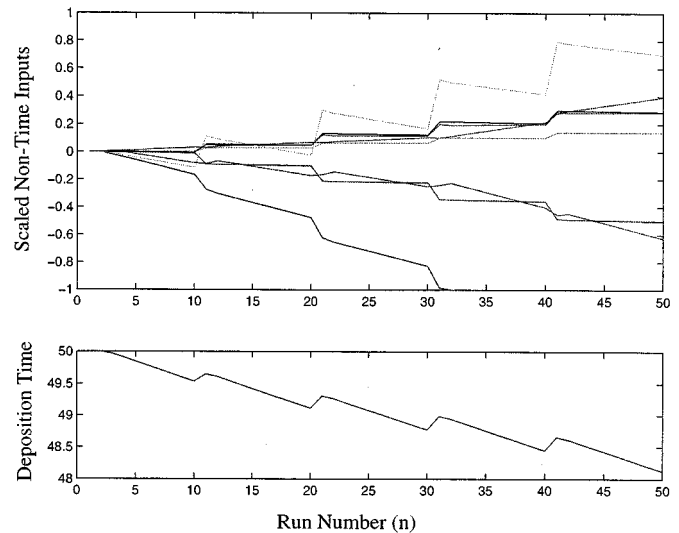


Fig. 17. Noiseless continuous system: controlled input trajectories.

close to target. While it is essentially impossible to see from these plots, the system accurately simulates the combined use of lot-to-lot and run-to-run thickness measurements. Once every 10 runs (one “lot”), the controller is presented with the full set of thickness measurement sites. For the other nine runs in each lot, the error extrapolation joint sensor measurement technique described in Section IV-C is used to fill in sites that are not sampled. This technique works well for the noiseless simulations in part because the uncontrolled thickness maintains a linear profile as the drift progresses.

Note the “bumpiness” of the controlled resistivity in Fig. 16 versus the smooth surface of the controlled thickness. This part of the simulation demonstrates some effects of using lot-to-lot EWMA control for a rapidly drifting process. As described above, measured thicknesses are presented to the controller after every wafer, simulating true run-to-run control. Measured resistivities are only presented to the controller once every ten runs, simulating lot-to-lot control. The controller cannot react to drifts between measured wafers, thus the process continues to drift during the processing of a lot. This type of drifting resistivity is not expected to occur in the actual equipment, so the results are primarily for demonstration purposes. If such persistent and rapid drifts are found to occur, a double EWMA

(or PCC) controller can be implemented and applied to the problem [1].

Fig. 17 shows the input trajectories for the time and nontime inputs. The simulation requires changes in the nontime inputs that are comparable to their bounded ranges (± 1). Also, the lot-to-lot updates result in noticeable discontinuities (bumps) in the input trajectories, much like those seen in the controlled resistivity plot.

The Noiseless Discrete System: The drifting processes shown in Figs. 13 and 14 are employed once again, but this time the controller is required to select discretized process settings. Table IV specifies the discretization levels available for the processing equipment. Figs. 18 and 19 show the resulting controlled thickness and resistivity profiles. The system’s controllability is hardly affected by the added limitations on the controller. We will see in the following sections that process noise provides considerably more variation than discretization effects.

The Noisy Continuous System: While the scenarios described above provided some basic benchmarks for the epi control system, a more realistic simulation is demonstrated

TABLE IV
DOE INPUT DISCRETIZATION LEVELS

Factor	Discretization level
Deposit time	0.1 sec
Deposit temperature	1.0 °C
Dopant mixing ratio	0.1 %
Dopant main flow	0.1 sccm
% Lower power	1.0 %
% Inner power	1.0 %
Dilutant (H ₂) flow	0.1 slm
Trichlorosilane (TCS) flow	0.1 slm
Center gas flow valve (Accusett inner setpoint)	1 "unit"
Outer gas flow valve (Accusett outer setpoint)	1 "unit"

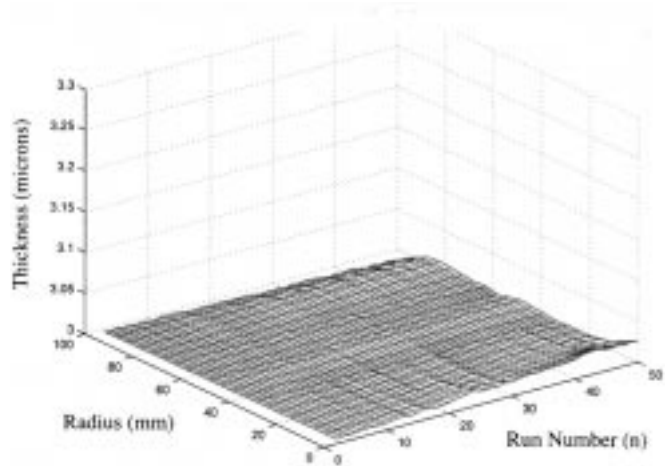


Fig. 18. Noiseless discrete system: Controlled thickness.

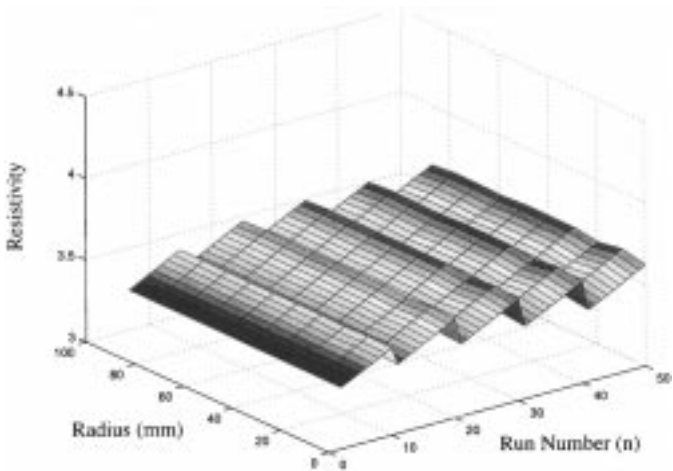


Fig. 19. Noiseless discrete system: Controlled resistivity.

by including random process noise. Figs. 20 and 21 show the uncontrolled thickness and resistivity profiles for a noisy, drifting process. The deterministic drifting components are identical to those found in Figs. 13 and 14. The random noise components are generated using the replicate-based mean shift noise and site noise statistics found in Tables II and III.

As shown by Figs. 22 and 23, the controller is able to maintain fairly good thickness and resistivity control despite the ad-

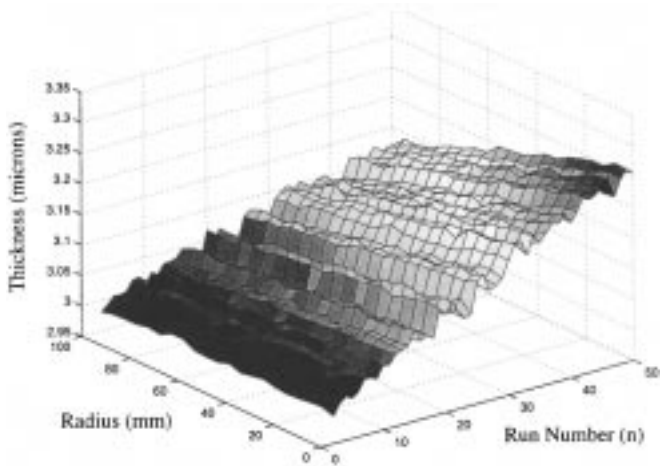


Fig. 20. Noisy system: Uncontrolled thickness.

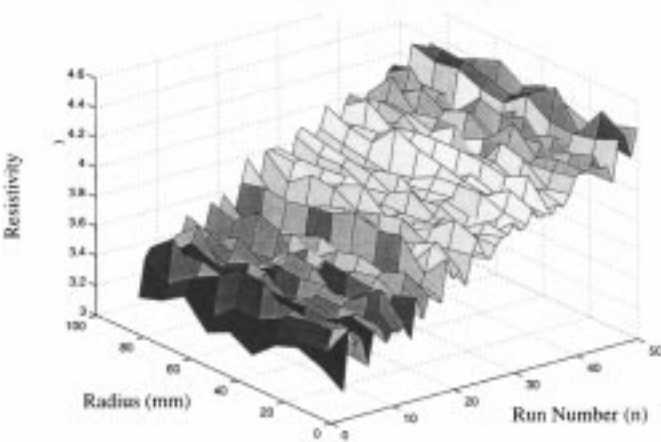


Fig. 21. Noisy system: Uncontrolled resistivity.

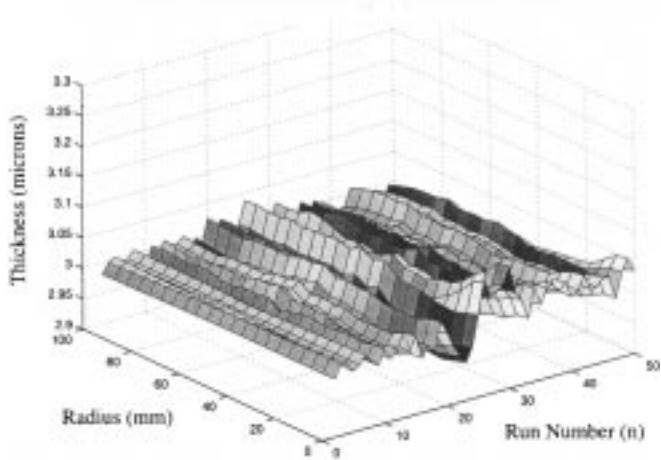


Fig. 22. Noisy continuous system: Controlled thickness.

dition of process noise. Variation seen in the controlled profiles appears to be similar to, but slightly larger than, the underlying process noise. For thickness control, additional noise comes from the drift components and the error extrapolation strategy. Additional noise in the controlled resistivity primarily comes from process drift combined with lot-to-lot model updates. Notice that the obvious “bumps” in controlled resistivity

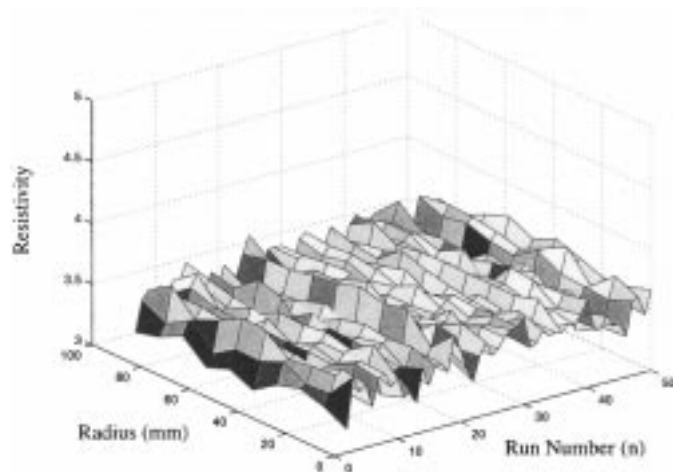


Fig. 23. Noisy continuous system: Controlled resistivity.

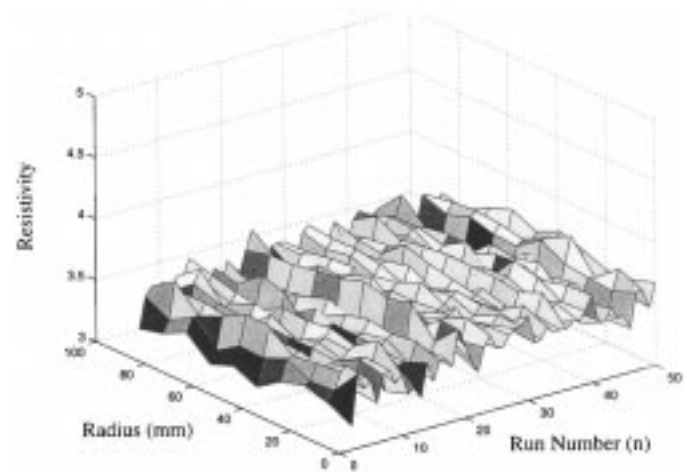


Fig. 26. Noisy discrete system: Controlled resistivity.

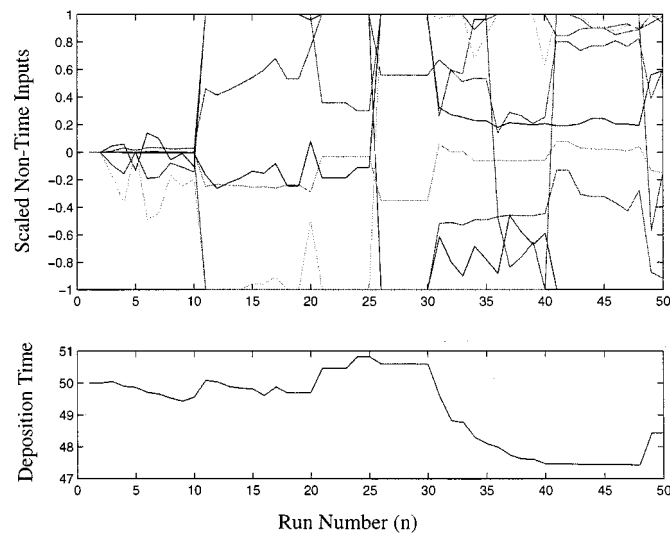


Fig. 24. Noisy continuous system: controlled input trajectories.

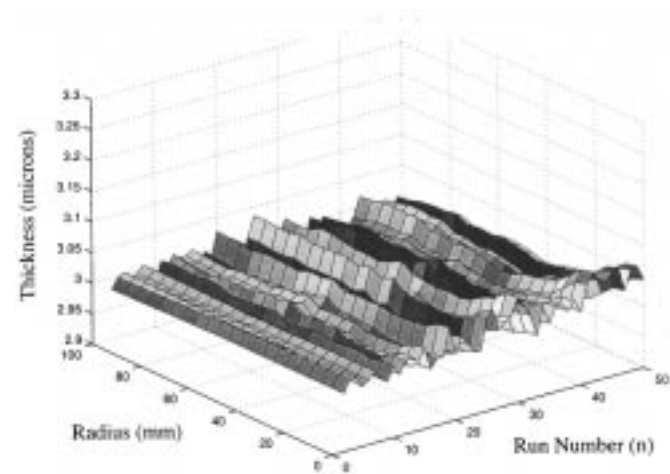


Fig. 25. Noisy discrete system: Controlled thickness.

seen with control of a noiseless system are almost completely hidden by the addition of process noise.

Fig. 24 shows the input trajectories for the time and nontime inputs. Notice that the trajectories for the noisy system swing much more wildly than those of the noiseless system. If this appears to degrade the performance of a real system, it can be addressed by lowering the EWMA weights and/or penalizing input changes. The use of additional input weighting terms appears to give the best results, as shown in [5].

The Noisy Discrete System: For completeness, the noisy drifting system of Figs. 20 and 21 is controlled using discretized inputs, as was done with the noiseless system. Figs. 25 and 26 show the controlled thickness and resistivities, respectively. The results do not change significantly by constraining the controller in this manner.

VI. CONCLUSION

This paper has laid out a complete background and development strategy for combined lot-to-lot and run-to-run uniformity control of an epi deposition process. Process modeling has been addressed through the construction and execution of a DOE. The deposition rate and resistivity models provide promising results, indicating that our MRS, affine, time-based, EWMA controller should provide effective uniformity control for the epi deposition process. Run-to-run control experiments in an industrial setting are needed for final verification of the new controller.

ACKNOWLEDGMENT

The authors would like to thank Applied Materials for supporting this work and providing the processing equipment and expertise for this project. We would also like to thank the software development team at On-Line Technologies MKS Instruments (On-Line Products Group) for integrating the various equipment control, communication, and data processing capabilities.

REFERENCES

- [1] T. Smith, "Novel techniques for the run by run process control of chemical-mechanical polishing," S.M. thesis, MIT, Dept. EECS, 1996.
- [2] A. Gower, "An architecture for flexible distributed experimentation and control with an AME 5000 plasma etcher," S.M. thesis, MIT, Dept. EECS, 1996.

- [3] P. Rosenthal, P. R. Solomon, S. Charpenay, V. Yakovlev, W. Zhang, A. Bonanno, M. Spartz, J. Xu, A. Gower, T. Smith, D. Boning, A. Waldhauer, and W. Aarts, "Infrared spectroscopy for control and fault detection of advanced semiconductor processes," presented at the 1998 Int. Conf. Characterization and Metrology for ULSI Technology, Gaithersburg, MD, Mar. 1998.
- [4] A. Gower, D. Boning, P. Rosenthal, and A. Waldhauer, "Advanced multi-objective control for epitaxial silicon deposition," presented at the Advanced Semiconductor Manufacturing Conf., Sept. 2000.
- [5] A. Gower-Hall, "Integrated model-based run-to-run uniformity control for epitaxial silicon deposition," Ph.D. dissertation, MIT, Dept. EECS, 2001.
- [6] E. Sachs, R. Guo, S. Ha, and A. Hu, "Process control system for VLSI fabrication," *IEEE Trans. Semiconduct. Manufact.*, vol. 4, pp. 134–144, May 1991.
- [7] S. Ha and E. Sachs, "On-line control of process uniformity in single wafer processes," *IEEE Trans. Semiconduct. Manufact.*, vol. 12, pp. 200–213, May 1999.
- [8] W. Moyne, "Run-by-run control: Interfaces, implementation, and integration," S.M. thesis, MIT, Dept. EECS, 1995.
- [9] E. Sachs, A. Hu, and A. Ingolfsson, "Run by run process control: Combining SPC and feedback control," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, pp. 26–43, Feb. 1995.
- [10] S. Adivikolanu and E. Zafiriou, "Internal model control approach to run-to-run control for semiconductor manufacturing," in *Proc. American Control Conf.*, June 1997, pp. 145–149.
- [11] —, "Robust run-to-run control for semiconductor manufacturing: An internal model control approach," in *Proc. American Control Conf.*, June 1998, pp. 3687–3691.
- [12] J. Baras and N. Patel, "Designing response surface model based run-by-run controllers: A new approach," in *Proc. IEEE/CPMT Int. Elec. Manufacturing Technology (IEMT) Symp.*, 1995, pp. 210–217.
- [13] —, "A framework for robust run by run control with lot delayed measurements," *IEEE Trans. Semiconduct. Manufact.*, vol. 10, pp. 75–83, Feb. 1997.
- [14] T. Smith and D. Boning, "Artificial neural network exponentially weighted moving average controller for semiconductor processes," *J. Vac. Sci. Technol. A*, vol. 15, no. 3, pp. 1377–1384, May 1997.
- [15] M. Le, T. Smith, D. Boning, and H. Sawin, "Run to run model based process control on a dual coil transformer coupled plasma etcher," in *191st Meeting Electrochem. Soc.*, May 1997, p. 332.
- [16] E. Hamby, T. Kabamba, and P. Khargonekar, "A probabilistic approach to run-to-run control," *IEEE Trans. Semiconduct. Manufact.*, vol. 11, pp. 654–669, Nov. 1998.
- [17] T. Smith and D. Boning, "A self-tuning EWMA controller utilizing artificial neural network function approximation techniques," *IEEE Trans. Comp., Packag., Manufact. Technol. C*, vol. 20, pp. 121–132, 1997.
- [18] N. Patel and S. Jenkins, "Adaptive optimization of run-to-run controllers: The EWMA example," *IEEE Trans. Semiconduct. Manufact.*, vol. 13, pp. 97–107, Feb. 2000.
- [19] J. Davis, R. Gyurcsik, J. Lu, and J. Hughes-Oliver, "A robust metric for measuring within-wafer uniformity," *IEEE Trans. Comp., Packag., Manufact. Technol.*, vol. 19, no. 4, pp. 283–289, Oct. 1996.
- [20] T. Smith, B. Goodlin, D. Boning, and H. Sawin, "A statistical analysis of single and multiple response surface modeling," *IEEE Trans. Semiconduct. Manufact.*, vol. 12, pp. 419–430, Nov. 1999.

Aaron E. Gower-Hall received the B.S. degree from Worcester Polytechnic Institute, Worcester, MA, in 1994. He received the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1996 and 2001, respectively.

His principle research interests involve modeling, optimization, and distributed architectures for semiconductor equipment process control, including run-to-run and statistical process control as well as fault detection and diagnosis.

Dr. Gower-Hall is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.



Duane S. Boning (S'83–M'89) received degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in electrical engineering and computer science, including the Ph.D. degree in 1991.

He is Associate Professor of Electrical Engineering and Computer Science at MIT, and is Associate Director of the MIT Microsystems Technology Laboratories. From 1991 to 1993, he was a Member of the Technical Staff at Texas Instruments in Dallas, where he worked on process/device simulation tool integration, semiconductor process representation, and statistical modeling and optimization. At MIT, his research focuses on variation modeling, interconnect technology, and control in semiconductor processes, with special emphasis on chemical mechanical polishing and plasma etch. Additional interests include CAD tools for statistical process, device, and circuit design, network technology for distributed design and fabrication, and computer integrated manufacturing. He has over 120 papers and conference presentations in these areas of research.

Dr. Boning is Editor for the IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, and is a member of the ECS, MRS, and ACM.

Peter Rosenthal received the bachelor's degree in physics from Yale University, New Haven, CT, studied surface science at AT&T Bell Laboratories from 1982 to 1985, received the Ph.D. degree in applied physics from Stanford University, Stanford, CA, in 1991, and was a NIST N.R.C. Postdoctoral Fellow studying the fabrication and far infrared properties of high temperature superconducting devices from 1991 to 1993.

He is the General Manager of the On-Line Products Group at MKS Instruments, Inc. has been investigating the science and technology of electronics materials and process control through his career. As a Member of Technical Staff at Advanced Fuel Research (AFR) from 1993 to 1995 and at On-Line Technologies from 1996 to 2001, he has managed a technical and product development program to commercialize the application of Fourier transform infrared (FTIR) spectroscopy for leading edge semiconductor metrology, gas analysis, and process control. Following the acquisition of On-Line Technologies by MKS Instruments, he is currently managing programs to apply FTIR technology to the problems of safety, environmental monitoring and quality control in manufacturing advanced semiconductors, dielectrics, and ultra-pure gas production.

Ann Waldhauer received the B.S. degree from Cornell University, Ithaca, NY, the M.S. degree from the University of California, Berkeley, and the Ph.D. degree from the Université Pierre et Marie Curie (Paris VI).

She joined Applied Materials, Santa Clara, CA, in 1995, where she is responsible for analytical and development issues in the Epi Technology Development group. In this context, she has worked on annealing and epitaxial deposition processes, wafer surface quality, microcontamination reduction, and materials quality, as well as in-line metrology. Prior to joining Applied Materials, she worked at Stanford University, Hewlett-Packard, Bell Laboratories, and LURE Laboratory in France, on instrumentation and materials science research.