J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*

PROCESS CONTROL

Process control was initially developed from the practices of process industries such as paper, steel, and chemical manufacturing. Beginning in the early part of the century, it became more widely practiced starting in the 1940s. Applications now include processes in microelectronics manufacturing, such as bulk crystal growth, chemical vapor deposition, and etching. Process control is important because it provides a means for improving process yield and reducing variation, and also enables new processing capabilities to manufacture unique engineered materials and structures.

Process control provides an approach to developing an automatic control system for complex, nonlinear, distributed parameter systems involving mechanical, transport, electrical, material, and chemical attributes. The term *process control* is used broadly to cover a variety of functions including supervisory control, failure detection, and the lower level control that determines the required inputs to achieve a desired objective. It is applicable to large plants incorporating many subprocesses and for single processes. This article primarily focuses on the issues associated with developing the lower level of control for a single process.

A major goal of process control is to automatically determine the process input settings to achieve the desired output condition (referred to as the command signal) while minimizing the variation of the output from the desired level (e.g., error). Variations are caused by external inputs called disturbances and variation of the plant, such as aging. Performance is evaluated in terms of dynamic aspects (how fast the system needs to respond) and magnitude of the error. A powerful conceptual approach used in control system analysis is to evaluate the output error as a function of command and disturbance signals, which provides a useful basis for designing compensators to meet performance specifications.

Taguchi's concept of the process variability loss function provides an important basis for evaluating the value of process control (1). Given target quality objectives, deviation from that target should be characterized by some loss function. Thus, for a quadratic loss function, the narrower the distribution, the greater value the process provides. One way manufacturers deal with process variations is to divide (or grade) production output into various quality classes, a practice common throughout semiconductor manufacturing. This approach, however, actually adds cost because sorting provides no additional value. In addition, there is the loss associated with the value that could have been attained if the production had met a narrower distribution. In this context, process control is valuable if used to reduce variation, but the cost of the solution must be balanced against the return.

There are two major control options: feedforward and feedback. Feedback utilizes an error signal formed by comparing the actual output to the desired value, which is then used to determine a corrective adjustment to the plant's inputs. Feedback corrects for unknown variations, but it is reactive, acting only after an error has occurred. Feedforward control utilizes process knowledge to determine the input values required to achieve the desired output without process measurements. Such process knowledge can be informal, such as that of the equipment operator, or expressed formally in a mathematical model. Feedforward is anticipatory for known conditions, but does not compensate for unknown factors, such as disturbances or plant variations. Feedforward control is typically used in cases where a cost-effective sensor is not available and a robust model is available.

Such an approach has been used to control wafer temperature distribution in tube furnaces where it is not practical to measure the wafer's temperature distribution.

Feedback systems are implemented either utilizing measurements in "real-time" or after a process is completed. Real time implies that the output (or related process states) are measured as the process is running and the inputs are adjusted on the fly to compensate for errors. If the process variations or commands are slow relative to processing time (i.e., the batch time), then real-time control is not necessarily needed. In this case, run-to-run control is used where the results are measured after the process is completed. If the disturbances have long-term behavior relative to process batch time, run-to-run control works well assuming that there is only one independent disturbance. In cases where variations are fast relative to process timescales, closed-loop control provides significant performance advantages. For example, in spot welding, the thermal dynamics of the process vary significantly throughout the welding cycle, so that development of a dynamic measurement controller provides significant performance improvements (2). Similarly, real-time control also works well to eliminate time-varying disturbances. It also is advantageous if one wants to build in process flexibility (i.e., the ability to change process output levels), even for slower processes, when process modeling is not accurate enough to meet the performance objectives.

It is not always possible to directly measure the variable related to the process objective. An alternative strategy is to identify and control an intermediate or secondary process variable related to the primary variable. In crystal growth, for example, dislocation density is an important material objective for optoelectronic devices which is not feasible to measure in realtime. Even if it were measurable, it might not be desirable because, once a dislocation has been introduced in the crystal matrix, it cannot be removed. However, by controlling the crystal's thermal gradients, dislocations are prevented from forming in the first place. A related strategy is to use cascade control structures wherein disturbance paths are identified and local loops eliminate their effect before they propagate to the primary control objectives.

Other approaches to process control include minimizing input variations [the objective of statistical process control (*SPC*)], thereby eliminating external disturbances, or selecting the operating regime so that the output is insensitive to the disturbance (the later approach is Taguchi's design of experiments technique). The disadvantage of SPC is that it requires continued effort to track down all disturbances, whereas feedback control operates automatically. The disadvantage of the design of experiments is that it requires an extensive set of experiments on the production equipment, which could significantly affect production. Taguchi's technique, however, utilizes existing equipment without additional capital or development costs. In some cases, it is beneficial to combine these techniques. For example, design of experiments is used to determine the desired operating point when number of input parameters are to be set, and local feedback is used to eliminate disturbances and plant variation. Feedback control, however, can be implemented only if there is a manipulable input that compensates for the specific disturbance.

Developing a process control system requires addressing a number of issues beyond the design of the control algorithm (i.e., the equations that map measurement information to the input levels). Important issues include specification of the relevant process objectives (both technical and economic), identifying the constraints posed by the equipment and process physics, determining which control structure best achieves the desired performance objectives, and considering opportunities for improving the process capabilities by changing the system design. Answering these questions typically requires a good understanding of the relevant process physics in addition to the relevant control principles. Modeling is useful in answering these questions, but should be performed from a control perspective. Such a perspective seeks to understand the system by identifying its fundamental dynamics and the limitations they pose, its disturbances, objectives, and available manipulable inputs.

One major control structure issue stems from the fact that more than one input typically affects the desired output(s). Thus, it is important to understand which inputs can achieve the best performance. One way to evaluate the performance of alternatives is utilizing frequency-domain techniques to quantify gain and bandwidth for alternative inputs. Additionally, one should evaluate if some outputs are more difficult to control independently. For example, in a welding process, it is difficult to independently control the width of the heat affected zone and the size of the weld nugget. These problems are related to the inherent coupling of a process, and it may not be physically possible to achieve all the objectives. Thus, an important objective of the analysis is to determine such inherent conflicts.

A control perspective is also useful in selection of the operating regime, i.e., the specification of the nominal operating conditions for all inputs and process settings or trajectories. Important factors include how fast the process proceeds, the significance of different disturbances, and which aspects of the process physics dominate. The operating regime is selected from a basic understanding of the dominant process physics and/or from formal optimization techniques, such as design of experiments.

Consideration should also be given to redesign of the process and/or alternative control structures (i.e., combinations of feedback, feedforward, or cascade) to improve performance. Important examples of the benefit of integrating system and control design include eliminating known disturbances, adding new actuators to achieve independent control of an additional output, and sizing actuators to meet the performance specification.

Development of the control system requires identification of the fundamental dynamics of the process. Important issues to address are factors that fundamentally limit control performance or that need to be compensated for through system or control design. Some of these features might suggest changes in the operating regime selected to avoid the problem, such as if there are poorly damped modes. Other features, such as nonlinearities, need to be identified and considered in designing the control algorithm.

Process Control Analysis and Design Issues

This section provides a more formal statement of the analysis and design options for control. While this analysis is presented in terms of linear systems, the analytical framework of control theory provides guidance and a formal basis for developing an appropriate process control system.

Control Structures and Performance Analysis. Feedforward and feedback control is represented by the block diagrams in Fig. 1. Each block represents the mapping of an input u to an output y. The plant (e.g., process) is represented by g and the controller by k. Lower case letters represent single-input, single-output (*SISO*) systems whereas upper case letters represent multiple-input, multiple-output (*MIMO*) systems. Thus g is a scalar operator mapping u to y ($y = gu$) whereas G is a matrix operator mapping the vector of inputs to the vector of outputs ($y = G\,u$). The dynamics of linear constant coefficient systems is analyzed in the Laplace domain with complex transform variable s in terms of transfer functions such as $g(s)$ and $k(s)$. (The functional dependence on s will not be explicitly indicated henceforth.) Time-varying inputs for r, n, and d are readily analyzed in terms of a family of inputs (impulse, step, and ramp) and in the frequency domain (see the following section).

The input-output relationships for a SISO system are as follows (3,4):

$$\text{feedforward} \qquad y = gk_{ff}r + d_o + gd_i \tag{1}$$

$$\text{feedback} \qquad y = \frac{gk}{1+gk}r + \frac{1}{1+gk}d_o + \frac{g}{1+gk}d_i - \frac{gk}{1+gk}n \tag{2}$$

where d_o is an unknown disturbance that acts on the plant and is reflected at the output, d_i is a disturbance that affects a process input, and n is a noise signal that corrupts the measurement of the output. (These

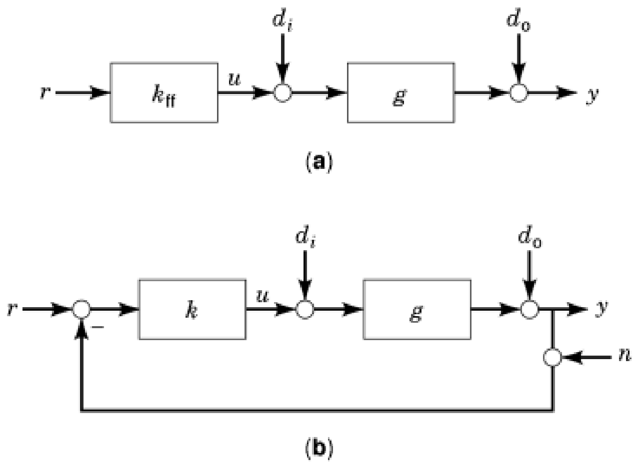

Fig. 1. (a) Feedforward and (b) feedback control structures.

relationships are generalized for MIMO systems by matrix algebra.) The output disturbance d_o indicates that there is an effect on the output but the functional form is not known, whereas an input disturbance d_i represents variations of the manipulated input or additional external inputs whose functional input/output mapping is known.

The achievable control performance is analyzed in terms of the error where $e = r - y$. The performance of the two structures is expressed as

$$\text{feedforward} \qquad e = d_o + g d_i \tag{3}$$

$$\begin{aligned}\text{feedback} \qquad e &= \frac{1}{1+gk}(r - d_o) - \frac{g}{1+gk} d_i + \frac{gk}{1+gk} n \\ &= S(r - d_0) + Tn - \frac{g}{1+gk} d_i\end{aligned} \tag{4}$$

Equation (4) reveals that feedback control reduces process variation, that is $|e|$ can be made small. In contrast, process variation for feedforward control is reduced only by eliminating disturbances and plant variation (see Eqs. (3) and (5)).

Analysis of the steady-state performance to step inputs provides insight into feedback control performance. This can be done by utilizing the final value theorem, which for step inputs corresponds to evaluating the magnitude of the corresponding closed-loop transfer functions [$T = gk/(1 + gk)$, $S = 1/(1 + gk)$] at steady state (i.e., $s = 0$.). In general, making $|gk| \gg 1$ achieves good performance for following commands and rejecting disturbances because this makes $|S| \ll 1$. Performance with respect to insensitivity to noise, however, is poor because $|T| \sim 1$. Note that several dynamic factors limit the maximum gain that can be used, thereby limiting achievable performance (see the following section).

One can consider the feedback controller k as an amplifier because $u = ke_m$ (where e_m is the measured error) and u is a power-level signal. Viewing k as an amplifier is consistent with requiring $|gk| \gg 1$ to achieve good performance. A simple version of such a controller is a proportional gain (i.e., P-type controller). More complex forms with varying gains for different frequency ranges can achieve better performance. In cases where

u is not an electrical signal, the controller can consist of two parts: the control algorithm and the actuator that manipulates the physical input u.

The achievable performance relative to variation of the process and/or model error, δg, is given by

$$\text{feedforward} \qquad \frac{\delta y}{r} = \frac{\delta g}{m} \tag{5}$$

$$\text{feedback} \qquad \frac{\delta y}{r} = S\frac{\delta g}{m} \tag{6}$$

where $g = m + \delta g$ and m is the nominal process or model. Thus, feedforward is sensitive to the relative model error, but feedback control compensates for such error or process variation, provided S is made small.

The different approaches to reducing process variation are analyzed by linearizing the output about the nominal operating point (W_o, y_o) and expressing the output variation as

$$\delta y = \frac{\partial f}{\partial w}\delta w + \frac{\partial f}{\partial u}\delta u \tag{7}$$

where the nonlinear mapping of the material inputs w and manipulable inputs u is given by $y = f(w, u)$. Classical SPC and Taguchi are considered feedforward approaches wherein SPC seeks to identify the assignable causes of the variation and to eliminate them, thus making $|\delta w|$ small. Taguchi seeks to identify the operating regime (i.e., values of w_o, u_o) whereby the output is insensitive to the variations, that is $|\delta f/\delta w| \ll 1$. The merit of this method is that it requires no additional capital expense. However, because it is based on experimentally mapping the input/output space, it might be quite expensive to run the experiments on the production line. The required condition for implementing a feedback solution is that the plant gain $g = \delta f/\delta u \neq 0$. Thus, in some cases there are disturbances or operating conditions that cannot be corrected for by feedback.

Closed-Loop Frequency and Dynamic Analysis. Frequency analysis yields important insight into designing an appropriate process control system and into factors that pose important performance limitations. It is also used to analyze the performance of more complex inputs and nonlinear systems linearized about a nominal operating point. The output response of a linear system $g(s)$ to a sinusoidal input $u(t) = \sin \omega t$ is given by

$$y(t) = |g(j\omega)| \sin(\omega t + \phi) \tag{8}$$

where $|g(j\omega)|$ is the magnitude of the complex transfer function evaluated at frequency ω and ϕ is a phase shift. The frequency characteristics of a system can be visualized by plotting the magnitude of $g(s)$ as a function of frequency ($s = j\omega$) on a log-log plot, known as a Bode plot (Fig. 2). An important characteristic is the range of frequencies with uniform amplification called the bandwidth ω_{BW}. The Laplace domain provides a basis for relating the characteristics of the Bode plot to the time-domain characteristics of the process. For example, the open-loop bandwidth approximates the open-loop dynamic response as $\omega_{\mathrm{BW}} \sim \lambda_{\mathrm{dominant}} \sim \tau^{-1}{}_{\mathrm{dominant}}$ where $\lambda_{\mathrm{dominant}}$ is the dominant pole of the system, and τ_{dominant} is the dominant time constant of the system (3).

Frequency analysis is useful for analyzing the error magnitude and dynamics for closed-loop systems and for compensator design. Important design insight is obtained by comparing the magnitude plots of $|gk|$, sensitivity $S(s)$, and closed-loop $T(s)$ transfer functions (Fig. 2). The closed-loop bandwidth of T ($\omega_{\mathrm{CL_{BW}}}$) is bounded by the frequency that $|gk| \sim 1$. Since the closed-loop bandwidth characterizes the range of frequencies

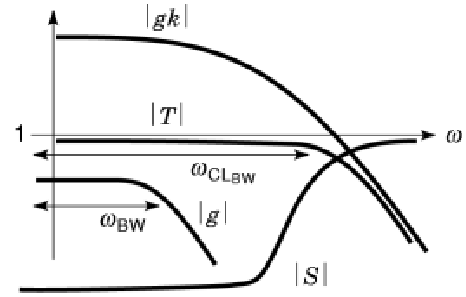

Fig. 2. Bode magnitude plot of open- and closed-loop transfer functions.

for good command following, disturbance rejection, insensitivity to plant variations, and speed of response, it is used as a primary design variable. Thus $K(s)$ is selected to achieve the desired crossover frequency for gk. The desired performance of the system is expressed in terms of the range of frequencies where r, d, and n have significant energy (i.e., large magnitude).

Noise and stability pose important performance limitations because S and T cannot be made small in the same frequency range. Thus, good performance cannot be achieved if the power spectrums for desired command following and disturbances overlap the noise spectrum. In addition, the Nyquist stability criterion poses a conservative magnitude bound on $|gk|$ relative to the uncertainty of the open-loop process dynamics as $|T| < |\delta g/m|$. Thus, $\omega_{CL_{BW}}$ is limited by model uncertainty, and an accurate process model is needed through the desired closed-loop bandwidth. (Similar logic can also be used for determining the model accuracy required to achieve performance objectives for a model-based feedforward control approach.)

The frequency-domain analysis also provides a context for selecting appropriate actuators. Each process typically has a variety of inputs that might be used as the manipulable variable for real-time control. From a general nonlinear representation of the process, $y = f(u_1, u_2, \ldots)$ where u_i is the ith input of l different inputs, one can linearize the system about the nominal operating point y^o, $u^o{}_i$. Because at least p inputs are needed to independently control p outputs, the first design decision is which p inputs achieve the best performance. For single-input, single-output (SISO) systems, the question reduces to Which of the actuators achieve the best performance? The comparison is made in terms of the open-loop bandwidth, reflecting how fast that input affects output, and the open-loop gain, which indicates how large an impact input has on output.

Inputs with the greatest bandwidth and gain typically result in the best performance. However it is not necessarily true that all high-bandwidth systems have large gain, thus possibly requiring a tradeoff analysis. High-bandwidth systems require less gain to meet the desired closed-loop bandwidth than slower systems, thus typically they have a larger robustness margin. In addition, they may have less model error over the same frequency range because the dominant pole is at a higher frequency. Similarly, high-gain systems require less controller gain to achieve a desired error performance level, thereby improving the robustness bounds. Alternative actuators are compared by plotting the Bode plots for each input. Because the inputs are of different units, the transfer functions should be scaled so they can be compared on a consistent basis. For example, normalizing by the nominal value of each input expresses each input as a fractional variation of the nominal input.

In selecting actuators, one should also evaluate limitations posed by dynamic issues such as nonminimum phase characteristics including time delays, open-loop instabilities, and right-half-plane (*RHP*) zeros (4). This factors fundamentally limit achievable closed-loop performance. For example, a RHP zero causes defective transient responses in that the initial response is in the direction opposite to the final steady-state. In practice, the closed-loop bandwidth is limited to the bandwidth of the RHP zero.

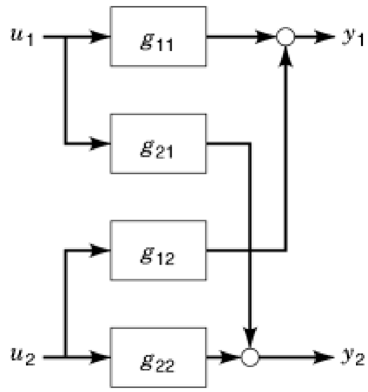

Fig. 3. MIMO system block diagram.

MIMO Systems. Additional issues arise for MIMO systems because of cross-coupling between inputs and outputs. This coupling causes more than one output when a single input is varied (Fig. 3). MIMO systems are represented by $y = G\,u$, where G is a $p \times l$ matrix mapping l inputs to p outputs. The frequency analysis is extended to MIMO systems by utilizing the spectral norms and principal gains of the transfer matrix which are calculated by a singular value decomposition. For a $p \times p$ matrix, there are p singular values, σ_i, which are positive scalars (3,4). Thus, for $|u| = 1$, $\underline{\sigma}(G) \leq |y| \leq \bar{\sigma}\ (G)$ where $\underline{\sigma}$, $\bar{\sigma}$ are the maximum and minimum singular values of $G(s)$. $\underline{\sigma}(G) \leq |y| \leq \bar{\sigma}\ (G)\underline{\sigma}(G) \leq |y| \leq \bar{\sigma}\ (G)$

There are a variety of control design algorithms for MIMO systems that compensate for cross-coupling. However, there are some systems, where the coupling makes it difficult to independently control all outputs. This problem is characterized a large condition numbers for G, that is, $\kappa_2(G) = \bar{\sigma}/\underline{\sigma} \gg 1$. Mathematically, a large condition number indicates that the matrix is close to loosing rank, that there are output directions which entail large control efforts. Such large inputs might saturate the actuators, which could result in loss of stability or inability to achieve the desired output value. On a practical basis, trying to design a system to operate under these conditions requires actuators which, for the most part are not used significantly, except to reach sensitive directions. Another analysis method that provide similar insight is the relative gain array (*RGA*) (3).

An important solution for poorly conditioned systems is to seek alternative system designs that improve the ability to control the process. Determining the reason for the poor conditioning is helpful and is obtained from the singular vectors of the process or by decomposing the transfer matrix into column vectors (15). Poor conditioning results from several actuators with similar effects on the outputs or an input that does not have the same magnitude of impact as the others. Alternative designs can be proposed once the nature of the limitation is determined.

Because both σ_i and κ_2 depend on the input/output scaling, all variables should be scaled consistently. Use of dimensionless variables compensate for the different units used for each input and output. Different normalization methods are used such as scaling about the nominal operating point, or defining perturbations that reflect relevant engineering scales, such as tolerances, saturation limits, or error limits. For example, chemical engineers normalize outputs by "transmitter spans" and inputs "by appropriate valve gains" (3).

Other Control Structures. Two important variations of control structure use alternative measurements of the primary process output. In processes where it is not practical or possible to directly measure the process output, a secondary variable related to the primary objective is used for feedback. For example, in bulk crystal growth, dislocation density is important because it affects the electro-optical characteristics of the material, but it cannot be measured in real time. Because dislocations are related to temperature gradients, one can instead control the temperature gradients to prevent the dislocation from being introduced into the

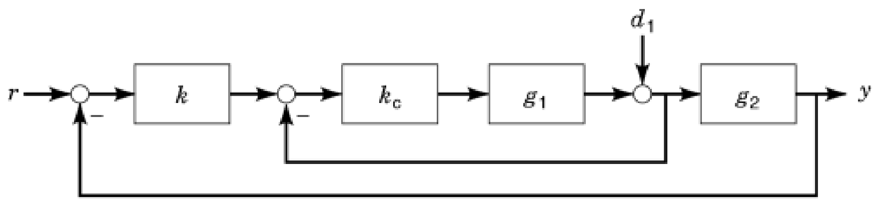

Fig. 4. Cascade control structure.

crystal matrix. In this example, there is an added benefit because a process state upstream of the output is controlled by feedback. Because feedback control is reactive, it only takes control action after an error has occurred. In the case of dislocations, however, once a dislocation has been introduced into the crystal matrix, it cannot be eliminated. Thus by controlling the upstream variable (e.g., the temperature gradients), one prevents the defect from occurring.

A related option is to utilize a cascade control structure (Fig. 4) (3). If a disturbance or a known process variation can be identified and measured, it is possible to close the loop around this variable to ensure that the variation does not propagate downstream to affect the primary objective. The benefit of this practice is that it results in better overall performance for a level of desired robustness because it significantly reduces the action that the primary control loop needs to perform.

Control Algorithms. There are a variety of methods for designing the control algorithm, and the reader is referred to the other related articles for specific details. For linear systems, important design methods include classical design methods such as *PID* for *SISO* systems, and more advanced techniques such as optimal control, robust control, H infinity, and model predictive (3) which can be used for both SISO and MIMO systems. These methods require a dynamic process model which can be either derived from first principles or experimentally using system identification techniques. Alternatively, adaptive control techniques do not require a system model.

For systems that vary in time or with a changing parameter, gain scheduling is used to interpolate between control designs developed for different conditions. Where there are significant distributions of disturbances or noise, stochastic control design can be used. Some systems have characteristics that require special approaches. These include nonlinearities and distributed parameter systems (which entail spatial variations). Other control approaches provide important advantages such as neural networks for nonlinear systems, fuzzy logic that responds to different conditions, and non-real-time techniques based on statistics and designed experiments (4,5).

Application Examples: Electronic Materials Processing

Furnaces. Control of temperature is critical in many processes because it determines process uniformity, yield, material characteristics, and production rate. Two types of thermal systems used in the electronic fabrication industry that illustrate interesting control issues are tube furnaces to heat many wafers at the same time and rapid thermal processing (*RTP*) systems for single wafers.

Control of Tube Furnaces. Tube furnaces applications include dopant diffusion and oxidation, which are highly sensitive to temperature variations across the wafer. In addition, temperature differences also induce thermal stress that damage the wafer. Because wafers are loaded into a quartz boat and heated from the edge, the rate at which the wafers are brought up to the processing temperature is limited by the radial conduction of heat and the thermal dynamics of the furnace and load. Typically, independently actuated, multiply segmented heaters are used to compensate for spatial thermal variations. These variations occur because of different loading (i.e., the number of wafers processed each in run), the spatial characteristics of heat transfer modes (such as natural convection effects that differ significantly for vertical and horizontal furnaces), and end effects. Important process objectives include the rate of heat-up/cool down (which affect production rate without providing value), maintaining thermal uniformity, and achieving the process set-point temperature. Because of sensor limitations, it is difficult to directly measure the wafer temperature. Thermocouples used in the furnace include spike thermocouples, located near the heater elements, and profile thermocouples, located near the wafer edges. Thus, tube-furnaces are controlled by secondary measurements.

The process is a coupled MIMO problem because each furnace segment affects its neighbors. Traditionally, decoupled proportional-integral-differential (*PID*) loops control each zone, which does not take into account the cross talk of neighboring elements. This coupling limits how tightly each loop is tuned, because unmodeled dynamics cause it to become unstable or introduce a disturbance into the neighboring region.

Several model-based control schemes have been developed to overcome the measurement limitations. These schemes are essentially a hybrid of feedback and feedforward, where the model is infers the wafer thermal distribution based on the process model and measurement data and designs a MIMO controller. Because many of the primary heat transfer coefficients dependent on operating temperature, wafer loading, geometry, and material properties that are difficult to measure directly and/or highly variable, the model has a number of coefficients that must be empirically determined with instrumented dummy wafers.

A MIMO control approach enables more aggressive control action because the otherwise ignored interactions between zones is now taken into accout. Development and application of such a process has been undertaken by a team combining the manufacturer (Motorola), the furnace vendor (Silicon Valley Group), and the control company (Voyan) (5). Implementation has reduced process cycle time by 18% and the fractional variation in standard deviation (defined as the standard deviation normalized by the set-point temperature) was reduced from 1.67 to 0.77% at high temperatures (950 °C).

Rapid Thermal Processing. Rapid thermal processing (*RTP*) technology enables fast processing for a variety of semiconductor manufacturing applications including diffusion, oxidation, chemical vapor deposition, and nitridation. Tungsten-halogen lamps heat the entire wafer surface, thereby minimizing processing times and achieving novel structures by avoiding the slow thermal relaxation that occurs with slow ramp rates. RTP's commercial prospects are in low-volume production of application-specific integrated circuits (*ASIC*), shortening the development time for new products, and possibly competing with large conventional fabricating processes because of the reduced processing time. Typical operating characteristics include ramp rates on the order of 50 °C/s from ambient up to 1100 °C, constant temperature for 1–5 minutes, followed by a rapid cool down rate. Temperature uniformity across the wafer is a critical performance objective. For example, the sensitivity of coating thickness variations to temperature variation for oxidation or CVD deposition of films can be calculated from the deposition rate relation $R_{dep} = k \exp(-E/kT)$ because these processes are typically thermally activated. Normalized sensitivities for the growth of polysilicon layers are 2.5%/°C of variation.

Real-time control is critical in several aspects of RTP. Although the ramp rate is determined by the lamp power flux, achieving uniform thermal distribution in both steady-state and during the ramps is difficult because of heat transfer variations across the wafer. The different view factors, wafer fixturing, and the variations in radiation and convective heat transfer over the range of temperatures and pressures prevent use of a single open-loop thermal actuator design to achieve uniform temperature for all conditions. Even a single closed-loop controller will not work. Process nonlinearity is characterized by gain values and time constants that vary by an order of magnitude over the operating regime (7).

System design is important in achieving good control performance for RTP. Independently driven thermal actuators are desirable to compensate for heat transfer variations across the wafer. In a joint TI–Stanford design, three independently controllable concentric rings of actuators were used (6). The maximum input power was 2 kW, 12 kW, and 24 kW (starting with the inner ring), consistent with the increased loss toward the outer portion of the wafer. Even though reflectors were used, there is some actuator overlap because each portion of the wafer "sees" more than one bank of lamps. Thus, the MIMO aspects of the problem must be

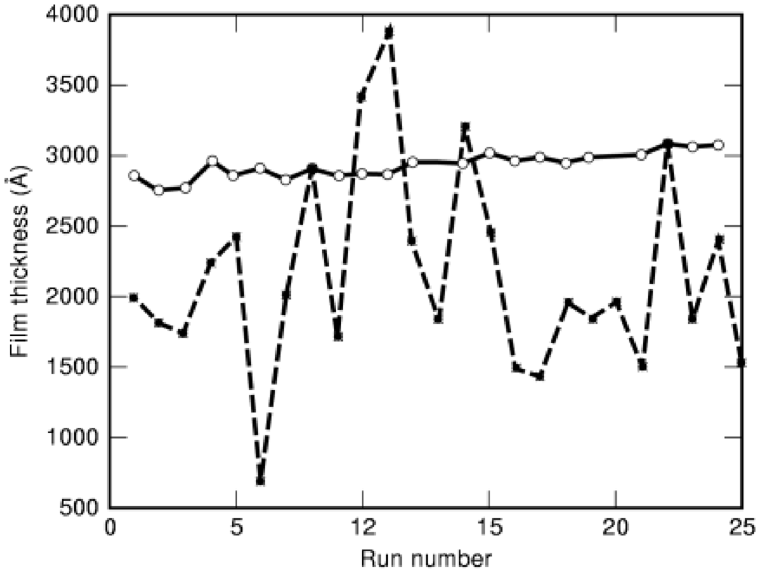

Fig. 5. RTP comparison of open-loop (dashed line) and closed-loop (solid line) operation for two 24 wafer lots (7).

considered at both the system design and control design levels. At the system design level, it is appropriate to evaluate the condition number of the system, that is, κ_2 for the DC transfer matrix. In the first design, $\kappa_2 \sim 110$, indicating that there are some output directions that would result in large actions in one or more actuators ($\sigma_1 = 12.2$, $\sigma_2 = 1.1$, $\sigma_3 = 0.11$). This might not be a significant problem because the maximum singular value corresponds to a condition of uniform temperature, which is the desired command signal. It does indicate that the lamps could be wired in two independent banks. However, it is difficult to compensate for a disturbance corresponding to the worst condition. To solve this problem, a baffle was designed to change the heat fluxes of the actuators to the wafer, thereby reducing the condition number to 23. Note that a low condition number does not necessarily mean that the actuators are decoupled.

Making accurate measurements is a significant challenge for RTP control. Thermocouples are not desirable because they require contact with the wafer and entail some amount of lag. Pyrometers are noncontact, but are sensitive to variations in surface emissivity related to temperature and film characteristics. Acoustic sensors, based on the temperature sensitivity of the wafer's elasticity, are being developed that avoid these problems.

One solution for optical pyrometer drift is to use run-to-run control to estimate the sensor error (7). Figure 5 compares open-loop RTP control with closed-loop. The open-loop varies widely, indicating that conventional run-to-run control does not perform well because there are no long-term trends. The closed-loop control, in contrast, has a ramp component, indicative of a systematic sensor error. The run-to-run measurement to detect this drift would be film thickness, because it is strongly temperature-dependent.

A number of different design approaches have been used to develop the control algorithm (6). Utilization of independent PID loops does not result in good performance because of the cross talk between inputs. The large range of operating temperatures requires a design technique to handle the related nonlinearities. The control design must also deal with the MIMO nature of the process and the significant modeling uncertainties. One approach is to incorporate both a feedforward and feedback approach implemented with internal model control (*IMC*). The process model is obtained from first principles or by system identification. To compensate for nonlinear variations, gain scheduling is implemented by parameterizing the model coefficients as functions of temperature. A linear quadratic Gaussian (*LQG*) design was also developed on the basis of linear models obtained from system identification. LQG easily accommodates the MIMO nature of the design problem and can also be augmented to include integrated error states ensuring zero steady-state error. Results achieved are control of ramp temperature distribution to $\pm 5^\circ$C and $\pm 0.5^\circ$C during hold.

Etching. Etching is a process that transfers patterns to a wafer by removing material not covered by a mask. Researchers at the University of Michigan have applied real-time control to reactive ion etching (*RIE*), which utilizes plasma ions to assist the etching process (8,9). A bias voltage accelerates the ions toward the wafer surface, enhancing etching normal to the surface. The process chemistry is quite complex including both gas phase and surface reactions. Etching process objectives include selectivity, the ability to etch only the desired layer and not other material; uniformity of etch rate over the wafer; anisotropy, the ability to etch the layer primarily in the vertical direction to prevent overhangs under the mask; and the etch depth.

The process can be divided into two coupled parts: plasma generation and etching. Inputs to the process include power, gas flow rates, and pressure set point where the pressure is regulated by varying a throttle valve upstream of a vacuum pump. Fluorine concentration [F] and bias voltage are outputs of the plasma process and inputs to the etch process. The conventional control practice is to set the power input and flow rate levels and use closed loop control to maintain chamber pressure. These variables are "far away" from the physics that determine the etch characteristics, suggesting that regulating intermediate plasma variables might yield better performance. A cascaded control structure is used with an inner loop around the plasma generator and the outer loop using the etch rate to determine the set points of the inner loop. If no real-time etch rate measurements are available, run-to-run control is used to adjust the set-point values of the intermediate variables. The function of the closed-loop control of plasma variables is to reject plasma generation process disturbances. Such disturbances include variations in the RF power and matching network, variations in the performance of the mass flow controllers, aging of the chamber, due to polymers deposited on the inside surfaces, and loading variations that occur when the amount of surface area varies (because of the number of wafers or mask variation).

Experiments indicate that etching has important dynamic characteristics that could also benefit from closed loop control. At the end of the etch, less material is left so that the etch rate increases. In some cases, this results in etching the sidewalls, resulting in undercutting. Thus, whereas the conventional practice is to utilize simple end-point detection by monitoring when new species are introduced into the gas phase, additional benefits can be achieved by etch rate control. There can also be initial transients due to contaminant variations, which change total etch time. Having dynamic control capability enables compensation for both of these factors.

Because the plasma generator is nonlinear, either local linear controllers can be developed to operate at a set point or a strategy to compensate for the nonlinearities can be applied. Important design limitations are imposed by both sensor and actuator considerations. Because of the poor signal-to-noise ratio of the fluorine concentration sensor, the bandwidth is limited to less than 1 radian/second. Some of the actuators, such as the throttle valve, are highly nonlinear, so that the regulation range is restricted for use in linear local controllers. To minimize steady-state error, free integrator dynamics were added to the controller. Both PID and LQG/LTR methods are used. The resulting performance of real-time control for maintaining plasma states is shown in Fig. 6, which indicates superior rejection of load variability and a more uniform etch rate throughout the process (8).

It is also desirable to control etch uniformity across the wafer surface, requiring a multivariable control approach. If real-time measurement of etch rate and uniformity output is not practical, a postprocess measurement run-to-run controller is used, built from a response surface mapping of the local plasma states to the etch outputs (9). Development of the feedback control algorithm requires a more general approach because the MIMO controller entails significant change in the operating regime. One approach is to consider that the process dynamics are linear and are coupled to a nonlinear static map of the inputs to output magnitudes (e.g., a Hammerstein model 9a). One can design a linear dynamic controller that operates in conjunction with a static controller to compensate for the nonlinearity.

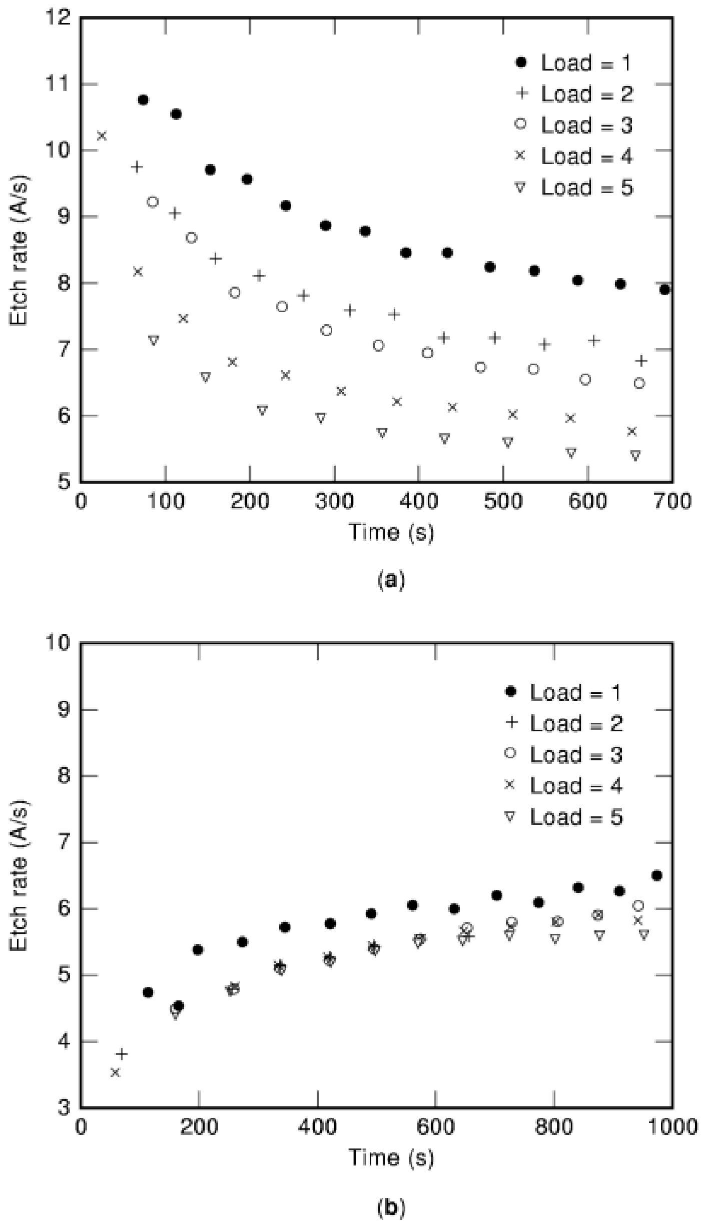

Fig. 6. RIE etch rate responses for different loading conditions: (a) open-loop, (b) closed-loop (8).

Experimental evaluation reveals that the pure run-to-run approach has results nearly comparable to the cascade approach for uniformity. This suggests that there may be a lack of good control authority in the selected operating regime and/or that the disturbances are more related to the etch process and, therefore, are not compensated for by the plasma real-time controller.

Chemical Vapor Deposition. Chemical vapor deposition (*CVD*) enables the manufacture of coatings with a variety of properties used in semiconductor production. Different coating structures and compositions are achieved by proper choice of precursors and deposition conditions which to date are primarily determined empirically. There are numerous CVD variations in terms of different precursors, energy sources, and reactor design. Important processing objectives include composition, structure, thickness, and uniformity of the deposited coating. Process examples used to provide insight into the application of control include control of metallorganic CVD (*MOCVD*) and deposition of TiN.

MOCVD uses liquid precursors in a bubbler to bring a gas phase of the metallorganic molecules into the reactor. Both III-V and II-VI compounds have been grown for applications, such as light emitting diodes, laser structures, and photodiodes. Such structures need precise tailoring of the spatial composition of the coating, which must be repeatable run-to-run. Two approaches employed to control coating composition include (1) control of the actuators and (2) feedback control from a measurement of composition. The first approach will work if the system is operating in a regime where mass transport limits growth but fails if chemical reactions interfere.

Gaffney et al. (10) focus on improving the performance of the input actuators, relying on a feedforward control approach to meet the composition objectives. This approach works if there are no process disturbances (other than inlet material variations) and there are no significant dynamics to the process relative to the desired deposition structures. Utilization of measurements related to the state of the bubbler enable direct feedback control of this actuator with a simple control law. Evaluation of the performance is established by measuring the optoelectronic characteristics of superlattices grown with five alternating layers of 3000 Å GaInAs with 200 Å layers of InP.

Warnick's et al. (11) approach is to measure the composition and thickness of the coating in real time with an ellipsometer. Because of the range of compositions desired to grow Al_xGa_{1-x}, the input-output mapping is nonlinear. The dopant rate dynamics are approximated as a simple integration of feed material inputs with a time delay to represent the transport delay, without detailed representation of process states and/or coupling. A Hammerstein model is used because the linear dynamics are assumed to be invariant to the nonlinearities. This could be valid for a specific operating point, for example, if the bulk pressure and temperature of the process are held constant but there are variations in composition. A nonlinear mapping is proposed that decouples the two outputs (composition and thickness) and linearizes the input-output mapping. Thus, linear control techniques are used for the two resulting decoupled linear systems. Free integrators are incorporated to insure asymptotic performance of the closed-loop system, in spite of the significant parametric modeling error that might exist. The bound on transient errors is established by the achievable closed-loop bandwidth. The most significant limitation is that the accuracy of the measurements is not clearly established, thereby introducing a possible DC error in all control efforts. The transient performance characteristics of the closed-loop system are also limited by the dead time. Because the processes are operated in the mass transport regime, the time delay also corresponds to the rate of growth, and limits how sharp a transition is made. In contrast, the feedforward approach of actuator control does not have such a limitation.

Major problems with each of these approaches suggest that a combination of methods could achieve better performance. The feedforward aspect would overcome the limitations posed by the time delay, whereas feedback would compensate for parameter uncertainties and unmodeled process physics. The combined control system could result in tighter control of spatial composition, thereby enabling faster growth rates with acceptable error.

To generalize control of CVD, one should consider the control structure design question. Gevelber et al. (12,13) evaluated alternative actuator performance capabilities for CVD, where, for many applications, control of grain size in addition to uniformity is critical. A mechanistic model of nucleation and growth suggests that independent control of temperature and deposition rate are required to control grain size. To determine which of the available actuators should be utilized, both the steady-state gain (Fig. 7) and dynamic characteristics (Fig. 8) of each actuator are evaluated in terms of controlling deposition rate (12,13). This analysis is conducted as a function of temperature because it strongly parameterizes changes in the dominant physics. Several input variables are ruled out because they have poor features like sudden reduction of actuator effectiveness

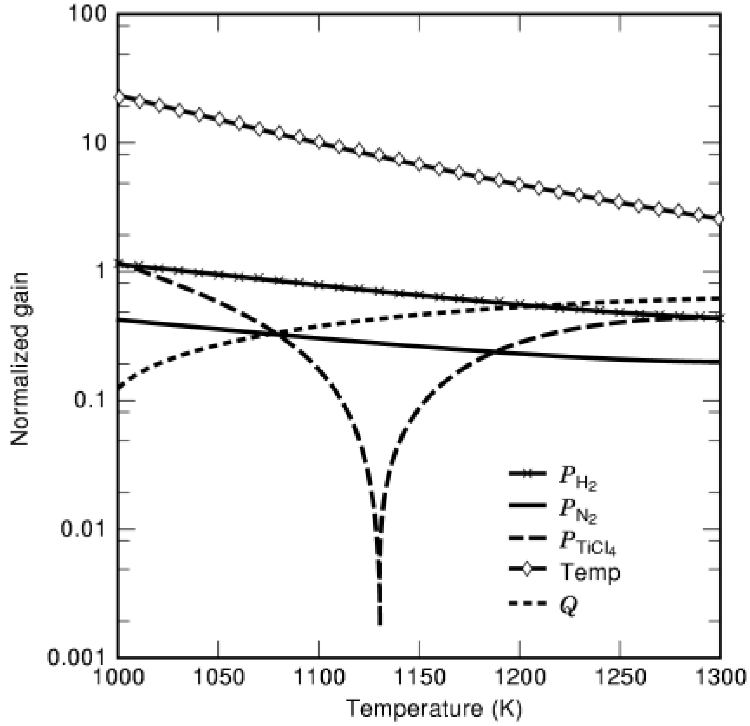

Fig. 7. Steady-state normalized gains of R_{dep} for various inputs for TiN CVD (13).

(partial pressure of $TiCl_4$). Analysis of the dynamic features are obtained from an analysis of the linearized system. One of the actuators has right-half-plane zero dynamics (pressure) that would limit the achievable control performance. In the case presented, a clear choice is given solely in terms of gain because the open-loop dynamics are similar. Note, however, that the dynamics vary significantly with temperature.

Selection of the appropriate operating regime must account for the variety of conflicting performance objectives, such as maximizing growth rate while minimizing nonuniformities, while considering the variations in the dominant process physics. Modeling analysis provides useful insight into answering these questions and should be directed to obtaining information about the dominant phenomena scale. Thus, for this example, because the dominant time delay scales with the reactor volume $\forall$ and flow rate Q, as $T_D \sim \forall/Q$, one can adjust the operating point (here the total flow rate Q), to minimize the fundamental limitation posed by the time delay (12). In hot-wall batch reactors, the reactant concentration varies throughout the reactor, resulting in varying growth rate from point to point. A ratio indicating the sensitivity is developed to determine which operating regime enables one to meet uniformity requirements while maximizing growth rates.

Crystal Growth. Growth of bulk crystals by the Czochralski process is one of the first steps in wafer fabrication. The process begins by bringing a seed crystal to thermal equilibrium with the molten semiconductor, and "pulling" the crystal from the melt. Heat transfer is a major phenomenon that determines growth characteristics. Important material properties that determine the crystal quality for different applications include dislocation density and dopant distribution. Si crystals are currently grown dislocation-free for commercial use with diameters of 200 mm. Some crystals, however, are more difficult to grow. For example, GaAs has a lower thermal conductivity than Si and a lower critically resolved shear stress (the stress at which a dislocation is

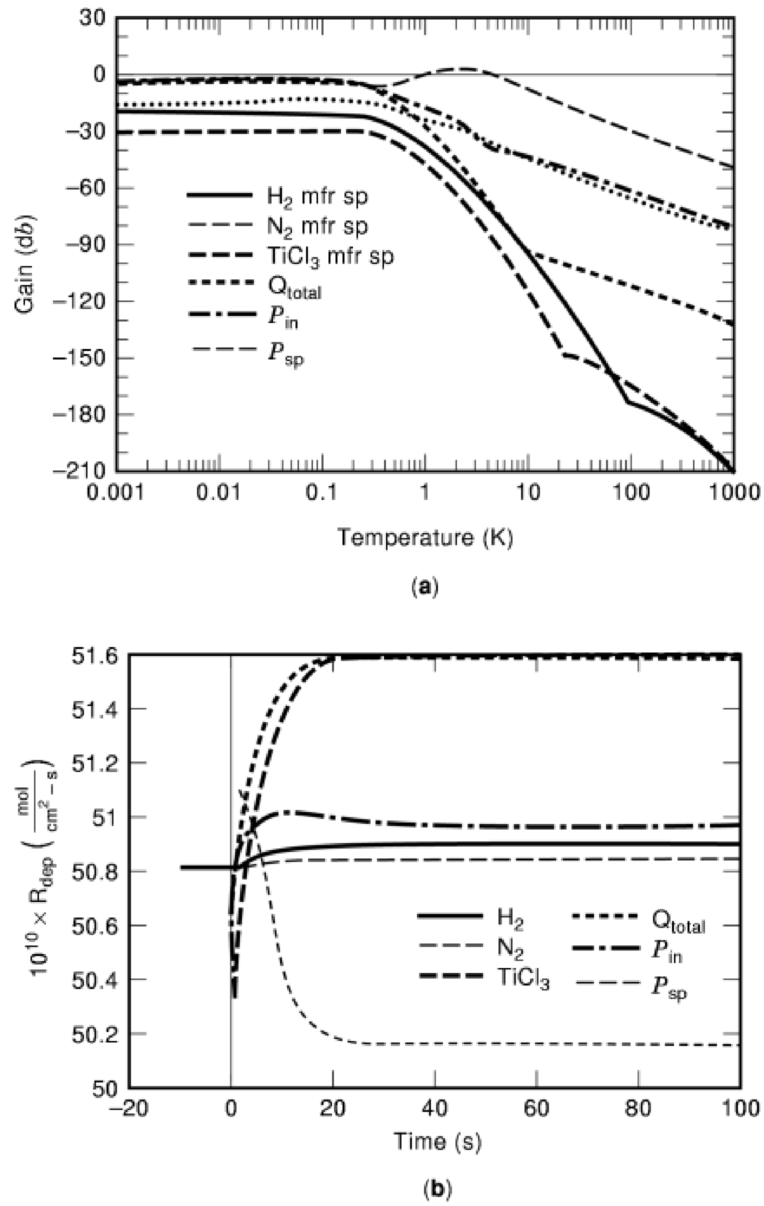

Fig. 8. (a) Bode and (b) 2% step response of R_{dep} for various inputs for TiN CVD, T = 1300K (12).

formed). Thus temperature gradients are higher, resulting in greater thermal stress and higher dislocation densities.

Dislocations are formed at the melt-crystal interface and also formed and multiplied in the crystal. The diameter is determined at the interface between the crystal and melt, whereas the dopant distribution is determined by the melt flow and the interface. Because the nature of this coupling, it cannot be guaranteed that any arbitrary set of material objectives can be achieved. For example, the interfacial shape that minimizes dislocation formation might not be optimal for other defects and dopant distribution, suggesting that a tradeoff analysis is required. To the extent that the detailed physics can be modeled, the analysis can be formalized. However, because much of the important physics is not analytically tractable, important insight is obtained from understanding the physics and experimental investigations.

Feedback systems were developed in the 1960s for automatic diameter control. Several different diameter measuring schemes are used including optical and weight measurement. Although there are a number of actuators for the process including heater power, pulling rates, rotation rates of the crystal and crucible position, and possibly an external magnetic field applied to minimize melt turbulence, typically only heater power and pulling rate are used for diameter control. However, because the other material objectives are coupled to the interface, they could adversely be affected by the action of the diameter servo loop.

Examination of the process indicates that several disturbances act on the growth interface (14). Because most pullers are run with a fixed charge, the reduction in melt level changes the heat flux and, therefore, the growth dynamics. In addition, as the crystal grows longer, it experiences a changing thermal environment. Thus, a cascaded control structure with an inner melt temperature loop can eliminate the melt disturbance. Direct compensation for the changing thermal environment is more difficult unless one considers designing an independent thermal actuator. The benefit of such an actuator is even more significant if it is desired to control the diameter and also the interface shape. Gevelber's analysis of the MIMO nature of this problem reveals that the system is poorly conditioned without such an actuator (15).

Modeling the dominant process physics reveals a right-half-plane zero associated with the weight measurement (14). The modeling provides a parametric expression of the zero's location relative to the process poles which is used to determine when there are significant performance limitations. The parametric description of the pole location is also used to understand when the system is underdamped, resulting in growth variations. This explains, for example, why it is more difficult to grow GaAs crystals and is used to help select the operating regime. Modeling also suggests important design consequences. For example, although a bottom heater does not achieve better control authority over the interface shape, it does result in a significantly faster system that may yield performance advantages.

Conventional control loops on pullers have been implemented with PID logic. Rivera and Seider have proposed a model predictive controller (*MPC*) and control structure to help integrate the different process objectives (16). A cascaded control structure accounts for the different timescales. One loop is the bulk controller that manipulates the heater power inputs. The objectives of this loop include pulling velocity, thermal stresses, and the melt dominant distribution. The other loop controls the radius by manipulating the pulling velocity. MPC coupled with models of the process achieves control solutions which directly take into account constraints to meet the other objectives.

BIBLIOGRAPHY

1. M. S. Phadke, *Quality Engineering Using Robust Design*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
2. K. Haefner *et al.*, Real time adaptive spot welding control. In D. E. Hardt (ed.), *Control Methods for Manufacturing Processes.* New York: The Dynamic Systems and Controls Division, ASME, November 27–December 2, 1988, pp. 51–62.
3. W. L. Luyben, *Process Modeling, Simulation and Control for Chemical Engineers*, 2nd ed., New York: McGraw-Hill, 1990.
4. J. M. Maciejowski, *Multivariable Feedback Design*, Reading, MA: Addison-Wesley, 1989.
5. C. Porter *et al.*, Improving furnaces with model-based temperature control. *Solid State Technology*, **39** (11): 119–131, 1996.
6. T. Kailath *et al.*, Control for advanced semiconductor device manufacturing: A case history. In W. S. Levine (ed.), *The Control Handbook*, Boca Raton, FL: CRC Press, 1996.
7. C. Schaper, T. Kailath, Thermal model validation for rapid thermal chemical vapor deposition of polysilicon. *J. Electrochem. Soc.*, **143** (1): 374–381, 1996.

8. O. D. Patterson, P. P. Khargonekar, Reduction of loading effect in reactive ion etching using real time closed loop control, *J. Electrochem. Soc.*, **144** (8): 2866–2871, 1997.
9. M. Hankinson *et al.*, Combined Real-Time and Run-to-Run Control of Etch Depth and Spatial Uniformity in Plasma Etching, *J. Electrochem. Soc.*, 144 (7): 2473–2479, 1997. (a) L. Ljung,, *System Identification*, Englewood Cliffs, NJ: Prentice-Hall, 1987.
10. M. S. Gaffney *et al.*, Control of III-V epitaxy in a metalorganic chemical vapor deposition process: impact of source flow control on composition and thickness, *J. Crystal Growth*, **167** (1–2): 8–16, 1996.
11. S. C. Warnick, M. A. Dahleh, Feedback control of MOCVD growth of submicron compound semiconductor films. *IEEE Trans. Control Syst. Technol.*, **6**: 62–71, 1998.
12. M. Gevelber, M. T. Quiñones, M. L. Bufano, Towards closed-loop control of CVD coating microstructure, *J. Mater. Sci. Eng. A*, **209**: 377–383, 1996.
13. M. A. Gevelber *et al.*, Modelling TiN Deposition for Control of CVD. In T. M. Bessmann et al. (eds.), 96-5: *Int. Conf. CVD XIII*, 189th Meet. Electrochem. Soc., Los Angeles, May 1996, pp. 157–162.
14. M. A. Gevelber, G. Stephanopoulos, Dynamics and control of the Czochralski process, II. Objectives and control structure design, *J. Crystal Growth*, **91**: 199–217, 1988.
15. M. A. Gevelber, G. Stephanopoulos, Control and system design for the Czochralski crystal growth process, *J. Dynamic Syst., Meas., Control*, **115** (1): 115–121, 1993.
16. R. Irizarry–Rivera, W. D. Seider, Model-Predictive Control of the Czochralski Crystallization Process: Part I—Conduction-Dominated Melt, *J. Crystal Growth*, **178**: 593–611, 1997.

MICHAEL A. GEVELBER
Boston University