15th Triennial World Congress, Barcelona, Spain

RUN-TO-RUN CONTROL AND PERFORMANCE MONITORING OF OVERLAY IN SEMICONDUCTOR MANUFACTURING

C.A. Bode[1], B.S. Ko[2], and T.F. Edgar[3]

[1]Advanced Micro Devices
Austin, TX 78741
U.S.A.

[2]Exxon-Mobil
Baton Rouge, LA 70806
U.S.A.

[3]Department of Chemical Engineering
University of Texas at Austin
Austin, TX 78712-1062
U.S.A.

Abstract

In the manufacture of semiconductor products, overlay is one of the most critical design specifications. Overlay is the position of a pattern relative to underlying layers, and overlay control largely determines the minimum feature size that may be incorporated into semiconductor device designs. Overlay control must be performed on a run-to-run basis, i.e. at the end of a run when product characteristics are available because they cannot be directly measured during a run. In this research a process model and a run-to-run control scheme was developed for overlay control, based on linear model predictive control (LMPC), and successfully implemented in a commercial facility. Performance monitoring of the closed-loop process was also carried out.

Keywords: Model predictive control, overlay, lithography, microelectronics manufacturing, controller performance monitoring

1. INTRODUCTION

The manufacture of semiconductor products requires that the devices in question be processed through a number of unit operations, such as lithography, diffusion, etch, implant, etc. In each operation, a single layer of the device is created or altered to form a specific piece of the circuitry. Lithography is the process used to isolate areas of the silicon wafer. It involves transferring a masking pattern from a template, called a reticle, onto the surface of a silicon wafer. This is achieved primarily by coating the wafer with a photosensitive polymers whose chemical properties are altered when exposed to a specific wavelength of radiation. After exposure of the resist and removal by a solvent, a patterned film is left that isolates areas of the wafer. It can be argued that lithography is the most critical process in semiconductor manufacturing and is largely responsible for driving improvements in the design of device circuitry.

The two most important aspects of the lithography sequence are the size and the position of the resist pattern relative to the substrate pattern. The size of the pattern, most commonly referred to as the critical dimension (CD), is a measure of the width of a particular feature within a given pattern. The position of the resist pattern relative to underlying layers is known within the industry as overlay. It is these two metrics of the patterning process which largely determine the health of the patterning process, and are therefore the most closely controlled aspects (Levinson, 1999).

Overlay is defined as the relative alignment of successive masked layers within the device. At each point on the wafer, overlay is the difference, *O*, between the vector position of the substrate geometry P_1, and the corresponding point of the next mask pattern, P_2, as shown in Equation 1.

$$O = P_1 - P_2 \tag{1}$$

The tolerance for overlay error is that the maximum positional deviation of one point in the substrate pattern from the corresponding point in the resist pattern must be no greater than one-third of the minimum spacing between features in the device mask. This ensures sufficient overlap between all of the circuits within the device, such that they are functional and reliable. The minimum feature size depends upon the minimization of the maximum overlay error that can be produced by a given lithography tool, and the speed, circuit density, size and other characteristics of semiconductor devices are related to this minimum feature size. Failure to meet these specifications force the wafers to be reworked through the masking step if detected, and may cause reliability issues or yield losses in the final device if not detected (Booth et al., 1992). The overlay tolerance is projected to decrease by

40% over the next 5 years, as part of the technology roadmap for semiconductors.

2. RUN-TO-RUN CONTROL

Run-to-run control may be viewed as a supervisory controller which manipulates the setpoints of underlying tool controllers. By analyzing the results of previous batches, the run-to-run controllers should be able to manipulate the batch recipe in order to reduce output variability. The main motivation for run-to-run control is a lack of in-situ measurements of the product qualities of interest. Most semiconductor products must be moved from the processing chamber to a metrology tool before an accurate measurement of the controlled variable value can be taken. The job of the supervisory, run-to-run controller is to adjust the recipes to reduce variability in the output product.

Run-to-run control is quite different from statistical process control (SPC), in that it actively attempts to compensate for error in the output of the process on a run-to-run basis. Rather than assume a stationary process, run-to-run control assumes that there may be slow drift or abrupt and persistent changes in the process. Through a process model, run-to-run control actively calculates an estimate of process state, and calculates the recipe changes necessary to keep the process on target given that estimate. Should the process begin to drift away from target, the controller compensates for this through modification of the recipe. The same is true in the case of abrupt and persistent changes in the process, such as step disturbances.

Development of run-to-run control strategy first begins by formulating a model of the process. Within the semiconductor industry these models are generally linear, or are linearized about the operating point of the process. Recent applications of run-to-run control by Bode (2001), Campbell (1999), and Edgar et al. (1999) have shown that multivariable control that allows for constraints offers definite benefits over conventional control strategies.

3.1. *Run-to-Run Linear Model Predictive Control*

Linear Model Predictive Control, or LMPC, refers to control algorithms that use a linear process model and a linear or quadratic open-loop objective function and linear constraints to compute the requisite manipulated variables over a future time horizon. LMPC uses a linear state-space model of the process to be controlled,

$$x_{k+1} = Ax_k + Bu_k \tag{2}$$

$$y_k = Cx_k \tag{3}$$

In these equations x_k represents the vector of states, u_k is the vector of inputs, and y_k is the output vector, all at run number k. The matrices, *A*, *B*, and *C* have constant coefficients. LMPC utilizes the state-space model to control the plant with the objective function

$$\min_{u^N} \sum_{j=0}^{\infty} (y_{k+j}^T Q y_{k+j} + u_{k+j}^T R u_{k+j} + \Delta u_{k+j}^T S \Delta u_{k+j}) \tag{4}$$

The flexibility afforded by the objective function allows better tuning of the process to provide a more robust control solution.

Lithography overlay control is performed by modifying adjustable exposure system controls to align each successive pattern in a device. These controllers typically have a unity gain, so that the control action and resultant change in overlay error are equivalent in most lithographic systems. Differences between nominal stage positions in different tools make the axis of alignment non-zero in most cases. This relationship can be described with the simple linear model in equation (5). The overlay model is simply represented as follows,

$$y_{k+1} = u_k + c_k \tag{5}$$

where y_{k+1} is the predicted overlay error, u_k is the manipulated variable associated with the output, and c_k is the model intercept. The model intercept represents the nominal position of the resist pattern. It is the estimate of the overlay error that would result if the manipulated variables were set to zero. Development of the state-space model for LMPC requires a modification of the standard linear model to include an intercept term, p_k. This can be accomplished by using the output disturbance model form of LMPC. A vector p_k is added to the model as a constant disturbance term in the output as shown below.

$$p_{k+1} = p_k \tag{6}$$

Though the disturbance state, p_k, largely incorporates effects from upstream processing, it is

manifested as a shift in the overlay performance of the stepper independent of the inputs. This assumption more closely follows the form of an output step disturbance.
Equations (7) and (8) show the complete state-space model for the overlay controller.

$$\begin{bmatrix} \hat{x}_{k+1} \\ \hat{p}_{k+1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{x}_k \\ \hat{p}_k \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u_k \qquad (7)$$

$$y_k = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{x}_k \\ \hat{p}_k \end{bmatrix} \qquad (8)$$

The complete augmented state vector, which includes x_k and p_k, is a 16 x 1 vector. The disturbance states, p_k, represent the intercept of the original model, and are estimates of the nominal position of the resist mask generated by the exposure tool. Alternatively, this can be viewed as the overlay error that would result if the manipulated variables within the recipe were set to zero. The input vector, u_k, is an 8 x 1 vector of the overlay settings within the exposure recipe. Each of these eight parameters is directly associated with one of the modes of overlay error. The elements of the 8 x 1 output vector, y_k, are the lot-mean overlay error parameters calculated by the metrology tools.

The filter equation must also be updated to include the augmented state vector. The following is the form chosen for overlay control.

$$\begin{bmatrix} \hat{x}_{k+1|k} \\ \hat{p}_{k+1|k} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{x}_{k|k-1} \\ \hat{p}_{k|k-1} \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u_k + \begin{bmatrix} 0 \\ L \end{bmatrix} (y_k - C\hat{x}_{k|k-1} - \hat{p}_{k|k-1}) \qquad (9)$$

This form gives open-loop estimation of the original state vector and linear filtering of the disturbance vector. The number of states that can be estimated from a single metrology event is necessarily less than or equal to the number of measurements from that metrology. As there are an equal number of states and measurements, only one of the states can be estimated at a time. The state in the overlay model is assumed to have no physical significance, and takes the value of the previous input at all *k* indices. This is achieved by setting *A* to zero, which is appropriate for overlay control in that successive runs are not correlated. All other matrices in the model are equal to the identity matrix, except for the configurable Kalman filter gain *L*.

The weighting parameters within the objective function are chosen to achieve the desired response of the control system. The input penalty term, *R*, is excluded from the overlay objective as it has no relevance to control of this process. The inputs have neither cost nor bounds within the region of control required to bring overlay to target. This leaves *Q* and *S* as the two tuning matrices within the equation. Since their relative weight rather than their absolute value determines performance, Q is chosen to equal the identify matrix and S is varied over a range of values to determine the desired weight.

Constraints may be added to the solution of the objective function by defining maximum allowable values of the input, output, and the input rate of change for each of the overlay variables. The actual solution of the objective function, both in simulation and production implementation, was calculated using the MATLAB © software program.

Prior to implementation, simulation studies of the algorithm were carried out to address various types of disturbances (drift, step, impulse, noise). The controller must be able to handle each of these conditions as desired from a process control standpoint. The impact of both high-frequency noise and impulse disturbances on state estimation must be minimized while the drift and step disturbances must be rejected by the controller to ensure that a minimum of product is subject to persistent bias. The LMPC objective function has a input rate-of-change penalty, and the controller employed a Kalman filter to reject the high-frequency noise. While no controller can prevent impulse disturbances, these two aspects of the controller minimized the impact of a single, aberrant data point on controller performance. Steady drift was tracked by the controller using feedback control. Step disturbances may be rejected most quickly by LMPC control by imposing a constraint on the estimated output.

As an example of a simulated test of the controller, LMPC was applied to actual manufacturing data collected from the Advanced Micro Devices (AMD) 0.18 micron Fab 25 facility. The data trend shown in Figure 1 is that of magnification, which is one of the more unstable overlay parameters in the exposure process. The input variables used for each run were subtracted from the measured overlay error to estimate the model intercept for each lot within the data. This intercept represents the disturbance state, p_k, that is tracked by the LMPC observer. Each data point is

then a representation of the disturbance state within the LMPC model. Noise is the dominant source of variation within the signal, although three large step disturbances are apparent, due to tool maintenance and stepper matching qualifications. Lots whose incoming magnification properties are unlike those lots that surround them, which gives the appearance of an impulse disturbance, are distributed throughout the data. The tail end of the signal shows a bimodal distribution in the magnification state of the lots, due to performance differences between various products manufactured at the site. The intercept for each lot was calculated as the difference between the input and output of the process, which was then normalized as a percentage of the maximum value of the intercept terms in the data set.

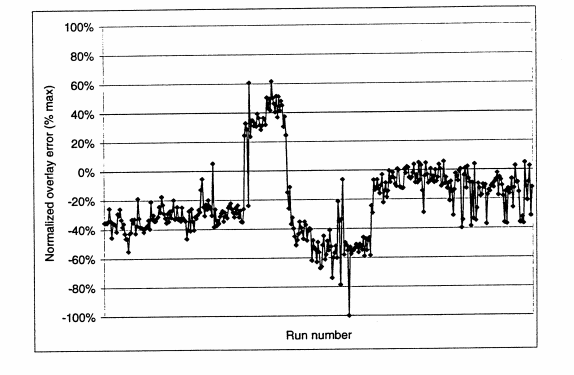

Figure 1: Magnification state trajectory calculated from AMD's Fab 25 production data which simulates an uncontrolled process.

LMPC was applied to this error trend with varying values of the adjustable control parameters by tuning the weighting matrices in equation (4). LMPC produced a clear improvement in overall reduction in variation, as shown in Figure 2. Recovery from the step disturbances was accomplished in about five process runs, which is excellent considering the magnitude of the disturbances. The simulation assumed that there was no process lag, meaning that metrology data were available from each lot before the next lot was processed. Although this is generally not true for processes in high-volume manufacturing, degradation in the performance of any type of control would be expected.

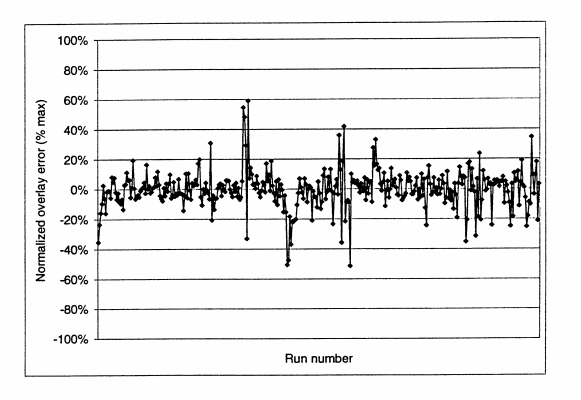

Figure 2: Simulated optimal LMPC control performance based upon the estimated Fab 25 magnification state trend.

3. IMPLEMENTATION OF OVERLAY LMPC

At the outset of this project in 1998, photolithography overlay control was deemed an important undertaking for AMD in its Fab 25 fabrication facility. At that time, the management of overlay control required significant engineering support. Tool matching events were performed periodically in an attempt to minimize differences in performance between tools. In addition, the manipulated variables responsible for overlay performance were included in each of the recipes used to process lots through the lithography tracks, and each required constant supervision to maintain control. This was largely a manual task, requiring numerous engineering hours each per week to analyze overlay performance, identify the recipes that needed modification, and perform the updates to those recipes. This high cost of maintaining the process prompted the search for a better control solution.

LMPC was applied to the control of lithography overlay in both Fab 25 and Fab 30, which are state-of-the-art fabrication facilities at AMD. Each of these fabs is a high-volume production facility utilizing a large number of exposure tools. In order to support the production line, the lithography modules operate in mix-and-match production environments. This means that one of a number of steppers may be chosen to process a lot at a given operation. Mix-and-match production is the most challenging environment in which to control overlay, as this maximizes the number of potential sources of disturbance to the control state. Successful control of such a facility, however, also maximizes its production capacity.

As a first pass at removing some of the variation from the overlay control signal which is

subject to a significant amount of noise, outlier rejection was used to cull significantly aberrant data from the process. It was generally clear from operating experience when a lot is a outlier by the magnitude of the error generated from metrology. Simple limits on the allowable measured error can successfully identify those lots which have overlay performance that significantly departs from the rest of the line. One may also set a limit on the amount of residual error in the fitted model, though this only captures those cases when the metrology results are erroneous.

The deployment of the run-to-run controller eliminated the need for engineering intervention to maintain and distribute overlay recipe settings to the exposure tools, thereby increasing the uptime for the tools and the amount of engineering resources that can be applied to other tasks within the module. The control state is updated each time new metrology data are made available, producing control settings that are based on all available process information. Once the initial deployment was completed, little incremental effort was required to deploy control to additional tools or layers. Improvements to the control method, when necessary, were implemented through the centralized control code and distributed to all tools and operations immediately and uniformly. In short, the task of overlay control was greatly simplified through the implementation of run-to-run control.

Automated overlay control deployed to Fab 25 was able to reduce the maximum site-level error, averaged over all controlled masking operations, by 43% over manual methods. The average maximum error at the beginning of the project was 90% of the allowable overlay error. As the controller was deployed to more masking layers and refined in configuration, it was able to reduce the overall error over a two-year period to stable operation at roughly 51% of the average specification limit.

The first phase of deployment of run-to-run control was a standard EWMA controller, lasting 23 months. LMPC was deployed to the fabrication facility in favor of the EWMA controller and has been used successfully for over one year. The LMPC method, along with the other improvements detailed within this work, was able to realize a 9% improvement to the average overlay error over the EWMA controller.

In addition to the improved control, the deployment of the LMPC method yielded other manufacturing benefits. Test wafers, used widely within semiconductor manufacturing, are non-product wafers or small production wafers lots which are run through a process to assess its performance. As test wafers add to the cost of running the process, both in material costs and tool time taken away from normal production, reduction of test wafer utilization is desirable. The LMPC control method facilitated a virtual elimination of test wafers for the purpose of overlay control. It also automated recipe management, significantly reducing the amount of engineering time required to maintain the process as well as eliminating human error. These benefits, along with the improved control, increased tool availability and production capacity of the lithography module.

3.1 *Performance Monitoring*

In order to further characterize the capability of the LMPC method, its closed-loop control performance was compared to a minimum variance control benchmark. Minimum variance control is a method that may be used to determine the theoretical limit of control performance for a given process, and can be obtained by weighting only output changes in LMPC. This is achieved by determining the amount of variance within the controlled variables of the process that is invariant to the control method employed. This portion of the variance is referred to as the minimum variance of the process, and represents the theoretical limit of performance that may be reached through feedback control without any restrictions on the control moves (Harris, 1989). Ko and Edgar (2001) have developed pertinent equations for performance monitoring of constrained MPC.

A sample of production data was taken from a lithography process in Fab 25 in order to compare it to the minimum variance benchmark. This sample was broken into four distinct data sets to facilitate several measurements of closed-loop performance. The first data set was further divided into three segments with 500 samples each. Table 1 shows the results of the performance assessment in terms of performance index defined in equation 13.

$$\text{Performance Index} = \frac{\text{Current output variance}}{\text{Minimum achieveable variance}} \qquad (10)$$

The remaining data sets were analyzed as one segment for each data set, and the performance index assessment results are summarized in Table 2. The index listed in bold only occurs for one data set and one controlled variable.

	Segment I	Segment II	Segment III
X translation	1.095	1.151	1.106
Y translation	1.143	1.115	**1.236**
Wafer rotation	1.073	1.162	**1.294**
X expansion	1.130	1.068	1.144
Y expansion	1.091	1.060	**1.289**
Magnification	**1.506**	1.138	1.120
Reticle rotation	1.165	1.093	1.129
Non-orthogonality	1.006	1.180	**1.204**

Table 1: Minimum variance performance index for the first data set. Data listed in bold have variance of more than 20% vs. the minimum variance.

	Data set 2	Data set 3	Data set 4
X translation	1.047	1.093	1.057
Y translation	1.071	1.077	1.114
Wafer rotation	1.042	1.008	**1.243**
X expansion	1.057	1.116	1.104
Y expansion	1.056	1.021	1.133
Magnification	1.112	1.056	1.044
Reticle rotation	1.065	1.076	1.061
Non-orthogonality	1.059	1.005	1.044

Table 2: Minimum variance performance index for later data sets.

These analyses indicate that the installed controller (LMPC) was achieving performance close to that of minimum variance control for each controlled variable. A few exceptions, noted by bold numerals, show that there are situations present in the operation of the exposure tools that can perturb the LMPC method and cause a degradation in performance. The metrology rate employed is Fab 25 is significantly less that one hundred percent, which increases the average metrology time lag and creates situations where the process has drifted significantly since the last metrology event. It is likely that Segment III in Table 1 also experienced more disturbances and involved a larger number of tools. The LMPC controller, unlike the minimum variance controller, employs a constraint on the input rate of change that causes a slower rejection of large disturbances. Equipment issues, both in the process and metrology tools, led to situations where there was apparent cross-correlation between the controlled variables which may or may not have been actual interdependency. Overall, however, the LMPC method achieved control near to the theoretical limit of performance, demonstrating that the controller is well suited for overlay control.

5. CONCLUSIONS

Run-to-run control was implemented in overlay control in lithography in commercial microelectronics manufacturing facility at AMD. The run-to-run controller was based on linear model predictive control (LMPC) and Kalman filtering, and the multivariable model of overlay included eight controlled variables and eight manipulated variables, albeit with negligible dynamics from run-to-run. LMPC was able to realize a reduction in variance of the controlled variables over the EWMA control previously used, and it has been used successfully for many lithography tools at AMD. Performance monitoring also indicates that the control system is able to keep the system close to the minimum variance benchmark, taking into account the effect of constraints.

REFERENCES

Bode, C.A., (2001), *Run-to-Run Control of Overlay and Linewidth in Semiconductor manufacturing*, Ph.D. dissertation, University of Texas, Austin, TX.

Booth, R.M., K.A. Tallman, T.J. Wiltshire, and P.L. Lee (1992). A Statistical Approach to Quality Control of Non-normal Lithographical Overlay Distributions. *IBM J. Res. Dev.* **36**(5),835-843.

Campbell, J. (1999), *Model Predictive Run-to-Run Control of Chemical Mechanical Planarization*, Ph.D. dissertation, University of Texas, Austin, TX.

Edgar, T.F., W.J. Campbell, and C.A. Bode (1999). Model-based Control in Microelectronics Manufacturing. *In Proc. 38th Conference on Decision and Control*, Volume 4. IEEE Control Systems Society.

Harris, T.J. (1989). Assessment of Control Loop Performance. *Canad. J. Chem. Engr.* **67**,856-861.

Ko, B.S., and T.F. Edgar (2001), Performance Assessment of Multivariable Feedback Control Systems, *AIChE J.*, **47,**1363-1371.

Levinson, H.J. (1999), *Lithography Process Control*, SPIE Optical Engineering Press, Bellingham, WA.