CONTROL AND MONITORING OF SEMICONDUCTOR MANUFACTURING PROCESSES: CHALLENGES AND OPPORTUNITIES

S. Joe Qin, **Gregory Cherry**, **Richard Good**, **Jin Wang**, and **Christopher A. Harrison**

Department of Chemical Engineering
The University of Texas at Austin
Austin, TX 78712, USA

Abstract:
The semiconductor industry is going through a technology transition from 200mm to 300mm wafers to improve manufacturing efficiency and reduce manufacturing cost per chip. These technological changes present a unique opportunity to optimally design the process control systems for the next generation fabs. In this paper we first propose a hierarchical fab-wide control framework with the integration of 300mm equipment and metrology tools and highly automated material handling system. Relevant existing run-to-run technology is reviewed and analyzed in the fab-wide control context, process and metrology data monitoring are discussed with an example, and missing components are pointed out as opportunities for future research and development. Concluding remarks are given at the end of the paper.

Keywords:
semiconductor manufacturing, fab-wide control, electrical parameter control, run to run control, fault detection and classification, metrology data monitoring

1. INTRODUCTION

The semiconductor industry is going through a technology transition from 200mm to 300mm wafers to improve manufacturing efficiency and reduce manufacturing cost per chip. Along with this transition is the doubling of capital expenditure in a 300mm fab versus a 200mm fab. Other technological changes include:

- Single wafer processing capability;
- Fully automated material handling systems (AMHS) with inter-bay and intra-bay transportation;
- Integrated metrology that allows for timely quality control; and
- Highly automated process control and fault diagnosis

Owing to the capital intensity of the new generation fabs, it is critical to maintain highly efficient operations, minimize downtime of equipment, and optimize the yield of high quality products. The International Technology Roadmap (Semiconductor Industry Association, 2003) clearly identifies that factory information and control systems are a critical enabling technology to reduce cycle-time and improve yield. These technological changes present a unique opportunity to optimally design the process control systems for the new generation fabs.

A persistent challenge in semiconductor manufacturing control is the lack of critical in-situ sensors to provide real time information of the wafer

[1] Corresponding author. e-mail: qin@che.utexas.edu.

status for feedback control and optimization. Fortunately, recent advance in metrology technology provides an opportunity for improving the timeliness and usefulness of the measurement data. Typically a modern fab has the following measurement data available for analysis and control:

(1) Real time trace data at the tool level which reflect the equipment health condition and provide feedback for real time control;
(2) Integrated metrology and in-line metrology data available for geometric dimensions after a major processing step, with small to moderate metrology delay;
(3) Sample and final electrical test (E-test) data available for electrical properties with medium or long time delay, but they have the most important information about the manufacturing effectiveness.

Advanced control and optimization methodology should maximize the use of all the information in an integrated hierarchy for highly efficient manufacturing and tight product quality control.

Monitoring and control of semiconductor manufacturing processes have been investigated at a number of U.S. universities and industrial research laboratories. Representative work includes U.C. Berkeley (Lee and Spanos, 1995; May *et al.*, 1991) on statistical modeling and control of plasma etchers, Michigan on real-time and run to run multivariable control (Hamby *et al.*, 1998), as well as MIT on different sensor and control technologies (Boning *et al.*, 1995; Boning *et al.*, 1996). Due to lack of in-situ sensors much of the control work is developed from the run-to-run (R2R) control strategy (Butler and Stefani, 1994; Sachs *et al.*, 1995). Research groups at University of Maryland contributed in the area of run to run control (Adivikolanu and Zafiriou, 2000; Baras and Patel, 1996; Zafiriou *et al.*, 1995). Sematech, a consortium of leading semiconductor manufacturers, posted several benchmark problems on plasma equipment fault detection and diagnosis (Bakshi, 1997). Adaptive and nonlinear control for R2R operations is proposed by (Del Castillo and Yeh, 1998). Model predictive control is applied to R2R control as well which has additional capability in handling constraints explicitly (Edgar *et al.*, 1999). At UT-Austin we have developed (i) stability conditions and tuning guidelines for multivariable EWMA and double EWMA control with metrology delays (Good and Qin, 2002; Good and Qin, 2003), (ii) multivariate statistical monitoring of RTA and etchers (Yue *et al.*, 2000; Yue *et al.*, 2001), and (iii) multivariate statistical control of CD metrology data from lithography (Cherry *et al.*, 2002). Other new development and applications of control and fault detection are reported at recent SPIE conferences and AEC/APC Symposia organized by Sematech and summarized in Del Castillo and Hurwitz (Del Castillo and Hurwitz, 1997) and Moyne et al. (Moyne *et al.*, 2001). Manufacturing companies like AMD, Intel, Motorola, and TI and vendors like Applied Materials, Brooks-PRI Automation, and Yield Dynamics are leaders in deploying APC technologies at the manufacturing lines.

In this paper we draw the analogy between semiconductor manufacturing fabs and chemical plants and propose a hierarchical optimization and control system for semiconductor fab control. A schematic diagram is shown in Figure 1 for this analogy. The equipment level control involves automatic feedback control of tool parameters and small scale run-to-run control using integrated metrology. The next level run-to-run control involves sharing information from multiple steps to achieve feedforward and predictive control. Since there are multiple tools in each module and each of them are different in terms of manufacturing effectiveness, a module level optimization is needed to make sure that each wafer is processed by the best combination of tools (or threads). The top level of the hierarchy is the fab-wide control which is the highest level optimization to achieve desired electrical properties by recalculating the optimal geometric targets and dosage for the lower level.

The organization of the paper is given as follows. We first propose a hierarchical fab-wide control strategy with the integration of 300mm equipment and metrology tools and highly automated material handling system. Relevant existing run-to-run technology is reviewed and analyzed in the fab-wide control context, process and metrology data monitoring are discussed with an example, and missing components are pointed out as opportunities for future research and development. Concluding remarks are given at the end of the paper.

2. A FRAMEWORK FOR FAB-WIDE CONTROL

Almost all existing development is on R2R control which adjusts recipes of a step based on metrology data at the equipment level. These are known as islands of control as illustrated in the lower part of Figure 1. None of the existing control strategies examine the coordination of multiple manufacturing steps to improve the overall product quality in terms of electrical parameters. The R2R controllers compensate for equipment drifts through metrology feedback, but they cannot compensate for metrology drifts and uncertainties. The direct control of electrical parameters proposed here can compensate for metrology drifts and systematic errors in the geometric measurements. It is be-

lieved that the control and optimization of electrical parameters represent the next generation of semiconductor manufacturing control system as it directly controls the electrical properties to a desired product profile by manipulating the operation requirements for lower level R2R controllers. The electrical parametric control and optimization will maximize the yield of high-grade products or reduce operational cost when a demand profile is specified by market orders.

The fab-wide control framework in Figure 1 provides optimization and coordination from step to step to reduce variability, reworks, and scraps, thus improving the overall equipment effectiveness and reducing manufacturing cost. This framework was first presented by Qin and Sonderman (Qin and Sonderman, 2002) after having deployed many R2R controllers at AMD and analyzed the need for a higher level control. The equipment level control involves automatic feedback control of tool parameters. The next level is run-to-run control using integrated or in-line metrology to achieve a specified target. The third level is a module level control that shares information from multiple steps to perform feedforward and feedback control and tool performance matching. The top level of the hierarchy, which is one of the focuses of the proposal, is electrical parametric control (EPC) to achieve desired electrical properties by recalculating the optimal targets for the lower levels. Equipment drifts, metrology drifts, and material variations are compensated by feedback at the EPC level, leading to improved process and metrology availability and reduced use of calibration and test wafers.

This multiple level control framework resembles the hierarchical control framework that has been successful in the refinery industry (Qin and Badgwell, 1997), but significant differences exist: (i) the lowest level control is mostly batch operations; (ii) the middle level R2R control has virtually no R2R process dynamics except for disturbance dynamics; and (iii) the top level EPC is a multi-step operation control that aims to compensate for errors made in prior steps, regardless of the nature of the errors as long as step-wise metrology measurement is available. This makes it different from model predictive control (MPC) of batch processes with shrinking horizons.

The E-test data are used to update the device model parameters based on mismatch between the E-test data and the model. After parameter estimation is performed, the estimated parameters are sent to a fab-wide optimizer, which distributes targets to lower-level controllers that regulate steps within the manufacturing process. The model updated with the new set of model parameters is used for EPC control.

The multi-step EPC control minimizes an objective function that penalizes the difference between the desired electrical properties and the updated model output subject to constraints. At the beginning before the first step is processed, all geometric input parameters are used to minimize the objective function. After the first step is processed for a wafer or a lot, metrology data for the first step is available as u_1, which will be different from the target $\hat{u}_1$ calculated before implementing the first step. To compensate for the manufacturing error in the first step the objective function is re-optimized over the remaining inputs with the first input fixed at u_1. Then the second processing step is implemented with the newly optimized target. This procedure is repeated until all processing steps are implemented for the processed wafer or lot. The objective function at each step is given as follows:

$$J = \left\| y_{EPC} - f\left(\hat{u}, u\right)\right\|^2 + \lambda \left\| \hat{u} - u_{nom} \right\|^2 \qquad (1)$$

subject to

$$u_{min} \leq \hat{u} \leq u_{max}$$
$$y_{min} \leq f\left(\hat{u}, u\right) \leq y_{max}$$

where u contains implemented inputs at the step, $\hat{u}$ contains the remaining inputs that can still be optimized, and u_{nom} is some nominal values for u that could be the values from previous runs. λ is a weight parameter that balances between the output target error and the input deviation from the nominal values. In the case that the output target is exactly feasible within the constraints, the optimization gives a solution that has minimal change from the nominal values or the values from previous runs. The feature is desirable to generate stable targets for the lower level R2R controllers. The constraints are derived from product specifications and requirements.

The vector y_{EPC} can have multiple entries containing multiple electrical parameters to be controlled simultaneously. Some of the electrical parameters have a target value based on the product specifications; others have only upper and lower constraints to make sure that these parameters meet the specifications. The EPC objective can be adapted to represent several modes of operations. For logic products one can choose to optimize the oscillation frequency subject to constraints that all major electrical parameters meet the specifications. The target y_{EPC} in this case is set at the highest desirable value for the product. In another mode of operation the target y_{EPC} is set based on the demand profile. In this case the target y_{EPC} can be easily achievable; the optimization generates a target that has smallest deviation from the nominal values, thus minimizing variability at lower levels due to target adjustment.

The optimization algorithm can be fulfilled using a nonlinear programming solution when the

device model is nonlinear. There is no perceived difficulty in the implementation of such an algorithm. A case study of a flash memory EPC is presented in (Harrison *et al.*, 2003).

3. RUN TO RUN CONTROL ALGORITHMS

3.1 *Run to Run Control*

In recent years, run-to-run (R2R) control technology has received tremendous interest in semiconductor manufacturing. Moyne and Hurwitz (2001) (Moyne *et al.*, 2001) define the run-to-run control as "a form of discrete process and machine control in which the product recipe with respect to a particular process is modified *ex situ*, i.e., between machine 'runs', so as to minimize process drift, shift, and variability". In order to modify the recipe to address the process drift, shift and other variability, the current tool and wafer states need to be estimated. One class of widely used run-to-run controllers is based on the exponentially weighted moving average (EWMA) statistics to estimate process disturbances.

The EWMA has been used for a long time for quality monitoring purposes (Box and Jenkins, 1963). Its use as a basis for run-to-run control is relatively recent (Sachs *et al.*, 1991). For a time series of measurement $\{\, x[n],\ x[n-1],\ \cdots \}$, where n denotes the run number, the EWMA is given in the following recursive formula:

$$\hat{x}[n] = \omega \hat{x}[n-1] + (1-\omega)x[n] \qquad (2)$$

where $\hat{x}$ is the EWMA estimate of x, ω is the EWMA weight, and $x[n]$ is the measurement of the process disturbance or parameter to be estimated. For a linear process model

$$y[n] = bu[n] + x[n] \qquad (3)$$

the disturbance

$$x[n] = y[n] - bu[n]$$

.

After the EWMA filter has estimated the process offset, a control law is used to determine the control input (or recipe) for the following run. In the unconstrained SISO case, the recipe is determined through simple model inversion,

$$u[n+1] = \frac{T - \hat{x}[n+1]}{b} \qquad (4)$$

where T is the process target.

The MIMO control law is somewhat more complicated as b may be non-square so that an inverse is not attainable. The MIMO control law therefore can take several different forms depending on the objective function of the optimization problem. A few of the commonly seen unconstrained MIMO control laws are listed below.

(1) Minimize the sum of the manipulated variables squared subject to the model hitting the process target:

$$\begin{aligned} \min_{u[n+1]} J &= u[n+1]^T u[n+1] \\ s.t. \quad & T = bu[n+1] + \hat{x}[n+1] \end{aligned}$$

$$u[n+1] = b^T(bb^T)^{-1}(T - \hat{x}[n+1]) \qquad (5)$$

(2) Minimize the sum of the change in the manipulated variables squared subject to the the model hitting the process target (Tseng *et al.*, 2002):

$$\begin{aligned} \min_{u[n+1]} J &= \delta u[n+1]^T \delta u[n+1] \\ s.t. \quad & T = bu[n+1] + \hat{x}[n+1] \end{aligned}$$

$$\begin{aligned} u[n+1] = {} & b^T(bb^T)^{-1}(T - \hat{x}[n+1]) \\ & + (b^T(bb^T)^{-1}b - I)u[n] \end{aligned} \qquad (6)$$

(3) Minimize the sum of squares deviation from target (Del Castillo and Rajagopal, 2002):

$$\begin{aligned} \min_{u[n+1]} J &= \hat{y}[n+1]^T \hat{y}[n+1] \\ s.t. \quad & \hat{y}[n+1] = bu[n+1] + \hat{x}[n+1] \end{aligned}$$

$$u[n+1] = (b^T b)^{-1} b^T (T - \hat{x}[n+1]) \qquad (7)$$

(4) Model predictive control formulation:

$$\begin{aligned} \min_{u[n+1]} J &= (T - \hat{y}[n+1])^T Q (T - \hat{y}[n+1]) \\ & \quad + u[n+1]^T R u[n+1] \\ & \quad + \Delta u[n+1]^T S \Delta u[n+1] \\ s.t. \quad & \hat{y}[n+1] = bu[n+1] + \hat{x}[n+1] \end{aligned}$$

$$\begin{aligned} u[n+1] = {} & (b^T Q b + R + S)^{-1}(Su[n] + \\ & b^T Q (T - \hat{x}[n+1])) \end{aligned} \qquad (8)$$

The first two control laws are used when the number of outputs exceeds the number of inputs. In this case there are an infinite number of control inputs that will bring the process to the expected target, T. An objective function is defined to establish a criteria to choose the 'best' controller input. The first objective function is to minimize the sum of squared controller input. The second objective function is to minimize the sum of squared *change* in the controller input. The third control law is used when the number of outputs exceeds the number of inputs. In this case there exists no controller inputs that will bring the process to the expected target. The objective function in this case is to minimize the sum of squared deviation of the expected output from the target. The final control law is a more general controller as the objective is to find the optimal balance between missing the process target, the absolute controller input, and the change in the

controller input from the previous run. The control law in (8) can be made to return equivalent results as the other three control laws by selecting the appropriate values of Q, R, and S.

Ingolfson and Sachs (Ingolfsson and Sachs, 1993) show that the EWMA controller is a discrete integral controller, which explains why it is able to compensate for process shifts and offsets. Butler and Stefani (Butler and Stefani, 1994) noticed that for processes with severe drifts, the EWMA controller is insufficient even when large weights are used. This problem becomes more severe when there is a measurement delay, which is almost inevitable in semiconductor manufacturing.

In order to control drifting processes, a predictor-corrector controller (PCC) (Butler and Stefani, 1994) and a double EWMA (dEWMA) controller (Chen and Guo, 2001) have been developed. The PCC algorithm uses two parameters, ω_1 and ω_2 to weight noise and drift respectively. The double-EWMA is very similar to PCC, as can be seen from the following equations:

$$\begin{aligned} a[n] &= \omega_1 x[n] + (1-\omega_1)(a[n-1]+p[n-1]) \\ p[n] &= \omega_2(x[n]-a[n-1]) + (1-\omega_2)p[n-1] \\ \hat{x}[n] &= a[n]+p[n] \end{aligned}$$

The only difference between double-EWMA and PCC is the estimate of intercept term a_i. Chen and Guo (Chen and Guo, 2001) show that both PCC and double-EWMA controller are in effect Integral-double-Integral (I-II) controllers, which are able to control drifting processes. However, since offset is often coupled with the noise of the process, the second filter may add variability to the control action in the presence of significant noise (Bode, 2001). In addition, tuning the second filter is not as intuitive as a single EWMA filter. Therefore, PCC or double-EWMA controller is not as widely used as EWMA controllers.

Like all feedback controllers, run-to-run control is subject to closed loop instability. The first study on the conditions for stability of the single input-single output (SISO) EWMA controller was published shortly after Ingolfsson and Sachs's first work on run-to-run control (Ingolfsson and Sachs, 1993). They noted that a process will become unstable when the input-output relationship between the tool recipes and quality measurements are not accurately estimated. Their work showed the allowable range of model mismatch that a process can have and still maintain asymptotic stability. This work was later extended by Tseng *et al* to show the stable region of a particular formulation of the multiple input-multiple output (MIMO) EWMA controller (Tseng *et al.*, 2002) . In addition, the effect of metrology delay on the stability of the SISO and multiple input-single output (MISO) EWMA controller was studied by Adivikolanu and Zafiriou (Adivikolanu and Zafiriou, 2000). They utilized an internal model control approach to derive the stability region of an EWMA controller with a delay of one run. A numerical method was then introduced for determining the stability region of the EWMA controller for processes with longer metrology delays. Good and Qin (Good and Qin, 2002; Good and Qin, 2003) extend the work of Tseng *et al* and Ingolffson and Sachs by deriving the stability conditions of the MIMO EWMA controller with metrology delay.

3.2 *Simplified dEWMA with RLS*

To simplify the double EWMA control Wang et al. (Wang *et al.*, 2004) consider the double EWMA control in the recursive least squares (RLS) framework and reduce the tuning parameters to a forgetting factor. Consider the process model as

$$y[n] = g(u[n]) + x[n] \tag{9}$$

where g is the nonlinear input-output model and the disturbance x has the following polynomial form,

$$\begin{aligned} x[n+i] &= \theta_0 + \theta_1 i + \theta_2 \frac{i^2}{2!} + \cdots + \theta_k \frac{i^k}{k!} + \epsilon[n+i] \\ &= \sum_{j=0}^{k} \theta_j \frac{i^j}{j!} + \epsilon[n+i] \end{aligned} \tag{10}$$

where $\theta = \left[\ \theta_0\ \theta_1\ \cdots\ \theta_k\ \right]^T$ are the parameters of the model to be determined, i is the time index, and ϵ is the a sequence of uncorrelated errors with variance σ^2. k denotes the model order, which can be determined by cross-validation. The model adequacy can be checked by calculating the sample autocorrelations of the residuals (Abraham and Ledolter, 1983).

The model (10) assumes that the model parameters are constant over all time periods. However, in many instances, the assumption of a time invariant model is restrictive and a locally constant model would be more reasonable. By applying the forgetting factor in the least squares criterion, more weight is given to more recent observations and past observations are discounted. Letting $\varphi[n+i] = \left[\ 1\ i\ \cdots\ \frac{i^k}{k!}\ \right]^T$, the model parameter estimates are determined by minimizing the following loss function,

$$V(\theta, n) = \frac{1}{2}\sum_{i=1}^{n} \lambda^{n-i}(x[i] - \theta^T\varphi[i])^2 \tag{11}$$

where λ ($0 < \lambda \le 1$) is the forgetting factor that gives more weight to recent prediction error $(x[i]-$

$\theta^T\varphi[i]$). The recursive algorithms to estimate the model parameters is (Åström and Wittenmark, 1995; Ljung, 1999):

$$\hat{\theta}[n] = \hat{\theta}[n-1] + K[n](x[n] - \varphi^T[n]\hat{\theta}[n-1])$$
$$K[n] = P[n-1]\varphi[n](\lambda + \varphi^T[n]P[n-1]\varphi[n])^{-1}$$
$$P[n] = (I_m - K[n]\varphi^T[n])P[n-1]/\lambda$$

where $K[n]$ is an $(k+1)\times 1$ vector of gains and $P[n]$ is an $(k+1)\times(k+1)$ matrix proportional to the covariance matrix of the estimated parameter. In the above recursive algorithm, the initial estimates $\hat{\theta}[0]$ and $P[0]$ can be obtained from a priori knowledge. The better the initial estimates, the smaller the effect of the transient behavior (Del Castillo and Hurwitz, 1997).

For $k = 0$, model (10) becomes the constant mean model,

$$x[n+i] = \theta_0 + \epsilon[n+i] \tag{12}$$

In this case the RLS gives a form that is very similar to EWMA except that $K[n]$ depends on the number data points n. Wang et al. (2004) show that the RLS converges exactly to EWMA with $\omega = 1 - \lambda$ when n approaches infinity. For finite n, $K[n]$ varies with n which is similar to the unsteady state Kalman filter.

For $k = 1$, model (10) becomes a linear trend model,

$$x[n+i] = \theta_0 + \theta_1 i + \epsilon[n+i] \tag{13}$$

In this case the RLS gives a form that is very similar to double EWMA except that $K[n]$ depends on the number data points n. Wang et al. (2004) show that the RLS converges exactly to double EWMA when n approaches infinity. The equivalence is achieved with

$$\omega_1 = 1 - \lambda^2 \tag{14}$$
$$\omega_2 = (1-\lambda)^2 \tag{15}$$

This result relates the double EWMA tuning parameters to the forgetting factor in RLS. The number of tuning parameters reduces from two to one which has the physical meaning of a forgetting factor.

3.3 *Kalman Filter Implementations*

Realizing (9) and 10 in a state space form,

$$x[n+1] = Ax[n] + w[n] \tag{16}$$
$$y[n] = Cx[n] + g(u[n]) + v[n] \tag{17}$$

where w and v are process and measurement noise, the estimated disturbance is

$$\hat{x}[n+1] = A\hat{x}[n] + K[n](y[n] - g(u[n]) - C\hat{x}[n])$$
$$= (A - K[n]C)\hat{x}[n] + K[n](y[n] - g(u[n]))$$

For a constant disturbance model, $A = 1, C = 1$. In this case the Kalman filter is equivalent to EWMA with $K[n] = \omega$ as n approaches infinity. For a linear trend model $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 1 \end{bmatrix}$. In this case the Kalman filter is equivalent to the double EWMA with $K[n] = \begin{bmatrix} \omega_1 & \omega_2 \end{bmatrix}^T$ as n approaches infinity. In the Kalman filter interpretation both tuning constants are necessary and are determined by the process and measurement noise.

3.4 *Time Control as R2R Control*

One of the most effective manipulated variables in R2R control is the processing time within a processing step such as etch time, exposure time, and planarization time. The controlled variables in the case are typically the extent to which the process develops under the processing time, such as depth of etch and critical dimensions. Given the processing rate as $r(t,n)$ for Run n, the process model can be described as

$$y[n] = \int_0^{T_f} r(t,n)dt \tag{18}$$

The processing rate $r(t,n)$ typically varies with run and time in a run. In the rare case that the rate can be measured in real time, the control problem is trivial. If the rate is not measurable in real time but some kind of end pointing mechanism or indirect measurement is available, the monitoring of the indirect measurements are effective, such as optical emission spectra in dielectric etch (Yue *et al.*, 2001). In the most typical case no real time measurement is available. In this case end of the step metrology measurement can be used for R2R control, with or without metrology delay (Wang *et al.*, 2004). Only the average rate can be estimated from previous processing steps and (18) becomes

$$y[n] = \bar{r}[n]T_f[n]$$

This multiplicative model does not fit into the typical linear state space model presented earlier, but it can be converted to the linear state space model with process and measurement noise by simply taking the logarithm. Therefore, all the control algorithms presented earlier in this are applicable to time control.

4. FAULT DETECTION AND MONITORING

Processing tool data such as temperatures, pressures, and gas flow rates will be used to monitor recipes applied to single wafers or batches

of wafers. Some typical processing operations include plasma etching, thin film deposition, rapid thermal annealing, ion implantation, and chemical mechanical planarization. At most processing steps, sensors collect data for each wafer or batch of wafers that are processed on the tool. This data can be in the form of real-time traces for a recipe, summary statistics available at the end of each run, or data from more advanced sensor platforms such as optical emission spectroscopy.

While the batch nature of semiconductor manufacturing provides plenty of opportunities for applying multi-way process monitoring (Nomikos and MacGregor, 1995), many forms of semiconductor metrology data are naturally organized in three dimensions. One such case is CD metrology, where the three dimensions are wafer, site, and parameter. Batch data are also commonly available from processing tools, which exhibits the dimensions of batch, time, and parameter (Figure 2).

Multi-way PCA has been successfully applied for batch process monitoring across many different industries. In the field of semiconductor manufacturing, Yue *et al.* (Yue *et al.*, 2000) demonstrated the concept of unfolding data by applying multi-way PCA to optical emission spectra for plasma etchers. For metrology and processing tool monitoring, the data can be unfolded by site or time (every row represents one site on a wafer or time instant in a batch) or by wafer (every row represents one wafer). In this work, wafer level fault detection and identification is desired, so the latter has been chosen as the more appropriate unfolding method (Figure 4). However, as will be discussed later, the advantages of analyzing data by site or time can be realized simply by implementing the multiblock approach.

4.1 *Metrology Monitoring*

While processing operations build the structures, metrology operations characterize them. Some examples of metrology measurements include development inspection critical dimension (DICD), final inspection critical dimension (FICD), and film thickness. Metrology measurements are normally taken at several locations on the semiconductor wafers, oftentimes for multiple features at the same site (i.e., top and bottom DICD). Fault detection and identification applied to site-level metrology data is intended to validate whether the structures built on the semiconductor wafers hit their targets and do so uniformly across the wafer surfaces.

As an example, we will use PCA to perform fault detection and identification on DICD data from Advanced Micro Devices' Fab25 in Austin, Texas. The DICD is the width of the pattern in the photoresist after the photoresist has been developed and before the next trim step. As shown in Figure 3, isotropy in development results in a small difference between the top of the photoresist and the bottom. The data set consists of 700 wafers, where both the top and bottom of the resist are measured at nine sites on each wafer.

An initial PCA model is built using the first 100 wafers and the VRE method (Qin and Dunia, 2000) selects 10 components as the optimal number for reconstruction. While this may seem high, keep in mind that the goal of the VRE method is the reconstruction of variables. After the PCA model is formulated, it is applied to the remaining 600 wafers, with the fault detection indices, SPE, T^2, and the combined index ϕ (Yue and Qin, 2001) provided in Figure 5. Note that the statistics have been scaled against their respective confidence limits and logarithms have been taken. The result is that the value of 1 represents the 99% threshold for each of SPE_r, T_r^2, and φ_r.

The indices charted in Figure 5 demonstrate that there is some behavior within those 600 wafers that is not consistent with the initial set of 100 wafers. In order to identify the cause of the excursion, multiblock contributions to the combined index are calculated. For the case of site-level metrology, the most logical blocking of the 18 variables are by parameter (top and bottom) and by site (sites 1 through 9). These multiblock indices are tracked by wafer in Figures 6 and 7.

Inspection of Figure 6 allows one to quickly reach the conclusion that the general trend is caused by drifts in both the bottom and top DICDs as time goes on. On the other hand, the larger outliers are more pronounced in either the top or bottom, but not both. Because the two CDs are highly correlated, the presence of such extreme outliers is likely to be caused by inaccurate values provided by the CD metrology tool, rather than problems with the physical structures on the wafers. While sensor faults are of interest, the more important issue is the process drift, which is first experienced in the bottom CD around wafer 200 and then propagates to the top 100 wafers later, where the signal becomes even more pronounced.

While Figure 6 grouped all 9 sites together for each of the two parameters, the Figure 7 contributions take into account both parameters at each individual site. These nine plots make it easy to identify problems based on wafer location. For the data set provided, the excursion appears strongest on sites 2, 3, and 4, while it is hardly noticeable at sites 6, 8, and 9. With knowledge of each site's location on the wafer it would be possible to use these plots to troubleshoot possible tilt or focus issues with the masking tool.

While tracking the block contributions is good practice for identifying excursions that influence a large number of wafers, one must also consider the case where a problem is identified on a single wafer, and the cause needs to be identified. To demonstrate this functionality, contribution plots have been generated for wafers 395 and 450 (indicated with arrows in Figure 5.

The contribution plot for wafer 395 is provided in Figure 8. It is easy to see that measurements 4 ($Bottom{-}Site$ 4), 12 ($Top{-}Site$ 3), and 13 ($Top{-}Site$ 4) are suspect. Problems are also indicated in both *Bottom* and *Top* parameters, along with *Site* 4 as the only extreme site contribution. A logical interpretation of these plots are that there was a significant issue with *Site* 4 and further investigation into that location on the wafer may be warranted to explore any issues that may affect product yield or performance. Although the top dimension on *Site* 3 was also singled out, the overall site contribution was normal when both the top and bottom were considered collectively.

The contribution plot for wafer 450 is provided in Figure 9. As previously shown in the φ_r plot (Figure 5), this was a less extreme fault than 395, but it did lie within the set of wafers experiencing some form of process drift. Although none of the contributions are as extreme as those for wafer 395, the plots suggest that four of the nine sites were faulty for both the bottom and top CDs. This fault signature appears to be quite typical of many wafers during the same processing time span, as corroborated in the variable and site contribution plots tracked for the entire sequence. For the case of wafer 450, the contribution plot showed that all faults indicated in the measurement contribution also propagated themselves to their corresponding multiblock contributions for both parameters and sites.

5. CHALLENGES AND OPPORTUNITIES

5.1 *Modeling of Electrical Parameters*

To implement the fab-wide control it is important to develop a physics-based device model that maps from geometric dimensions to electrical parameters such as oscillation frequency, erase time of flash memories, and sheet resistance. This model is different from process models used in R2R controllers that describe the relation between process operation conditions to geometric parameters such as critical dimensions, depth, or thickness. Since optimization is involved in EPC, which could use the model in fairly wide operation regions, nonlinear physics based models are chosen to accomplish this task. The models suitable for EPC must be implementable in real-time, making it different from simulation and design models. To illustrate the modeling task we use a flash memory cell to demonstrate the steps needed for these tasks.

5.2 *Model Update with Long Delay*

The developed models can work well for one operating condition. As the process metrology and material change over time it is important to adapt the model from real data using parameter estimation. In order to estimate a set of parameters, a nonlinear least-squares method is employed for the nonlinear physical model. The least squares objective is a dual of the EPC objective that minimizes the difference between the E-test data and the model output of finished wafers or lots subject to possible constraints. An effective parameter for model update is the intercept term or constant disturbance model. A challenging task in model updating is the long measurement delay in E-test data. The updating mechanism should respond only to long term persistent changes, not short-lived temporary errors. The integrated learning control and real time feedback control framework by (Chin et al., 2003) is a possible solution. The updated model is then used by the EPC controller to generate input targets for the next incoming lot.

5.3 *Integration of FDC and R2R*

As illustrated in Figure 1 each step in the fab wide control framework has a R2R controller and an FDC module for the step. The FDC are designed to monitor deviation from normal situations based on historical data analysis. One of the approaches is the multi-way PCA approach to equipment monitoring. The co-existence of the FDC and R2R control presents a challenge for their integration in two ways. First, FDC methods usually assumes repeatable batch profiles with similar or equal batch lengths. On the other hand, the R2R module is designed to adjust the recipe, e.g., process time to minimize variability due to normal process drifts. The FDC module, if not properly designed, could consider normal R2R adjustments as deviation from normal situations and signal a false alarm. Another challenge is the impact of R2R feedback on the FDC module. Because of tool control feedback the root cause of the fault could be transferred from one variable to another due to the existence of feedback. The use of feedback invariant subspace for fault diagnosis by (McNabb and Qin, 2004) is a possible solution.

5.4 *Integrated Metrology for Control*

The transition from 200mm to 300mm technology and beyond makes it possible to replace inline metrology with integrated metrology into the tools. The integrated metrology reduces the time delay in the measurement and provides the possibility of wafer to wafer (W2W) control which can reduce the variability further. Another possibility is within wafer (WiW) control that allows one to control from die to die (Sonderman and Bode, 2004).

6. CONCLUDING REMARKS

The semiconductor industry is becoming one of the most capital-intensive industries with a high ratio of capital investment to revenue. On the other hand, the optimization and control of manufacturing operations have received significant attentions only recently and shown to be a necessary competitive advantage. A well designed fab-wide control framework provides improved competitiveness of the semiconductor manufacturers as they transition to 300mm technology and 450mm technology in the foreseeable future. The automated material handling systems and automated R2R control capabilities provide the necessary foundation for implementing fab-wide control and fault detection in all levels of the hierarchy. It is envisioned by the leading manufacturers that most of the routine operations will move from clean rooms to a centralized control room in the future. This transition provides exciting challenges and opportunities to process control researchers and engineers to develop a new standard for this vigorously growing industry.

ACKNOWLEDGMENTS

Financial support from National Science Foundation under CTS-9985074 and a Faculty Research Assignment grant from University of Texas is gratefully acknowledged.

7. REFERENCES

Abraham, B. and J. Ledolter (1983). *Statistical Methods for Forecasting*. John Wiley and Sons.

Adivikolanu, S. and E. Zafiriou (2000). Extensions and performance/robustness tradeoffs of the EWMA run-to-run controller by using the internal model control structure.. *IEEE Transactions on Electronics Packaging Manufacturing* **23**, 56–68.

Åström, Karl J. and Björn Wittenmark (1995). *Adaptive Control*. Prentice-Hall. Addison-Wesley Publishing Company, Inc.

Bakshi, V. (1997). Benchmarking of commercial software for fault detection and classification (fdc) of plasma etchers for semiconductor manufacturing equipment. In: *Proceedings of the American Control Conference*. Albuquerque, New Mexico.

Baras, J.S. and N.S. Patel (1996). Designing response surface model-based run-by-run controllers: A worst case approach. *IEEE Transactions on Components, Packaging, and Manufacturing Technology - part C* **19**, 98–104.

Bode, C.A. (2001). Run-to-Run Control of Overlay and Linewidth in Semiconductor Manufacturing. PhD thesis. The University of Texas at Austin.

Boning, D., W. Moyne and T. Smith (1995). Run by run control of chemical-mechanical polishing. In: *1995 IEEE/CPMT International Electronics Manufacturing Technology Symposium*. pp. 81–87.

Boning, D.S., W.P. Moyne, T.H. Smith, J. Moyne, R. Telfeyan, A. Hurwitz, S. Shellman and J. Taylor (1996). Run by run control of chemical-mechanical polishing. *IEEE Transactions on Semiconductor Manufacturing* **9**, 307–314.

Box, G.E.P. and G.M. Jenkins (1963). Further contributions to adaptive quality control: Simultaneous estimation of dynamics: Nonzero costs. *Bulletin of the International Statistical Institute* **34**, 943–974.

Butler, S.W. and J.A. Stefani (1994). Supervisory run-to-run control of a polysilicon gate etch using in situ ellipsometry. *IEEE Transactions on Semiconductor Manufacturing* **7**, 193–201.

Chen, A. and R.S. Guo (2001). Age-based double EWMA controller and its application to CMP processes. *IEEE Transactions on Semiconductor Manufacturing* **14**, 11–19.

Cherry, G., R. Good, and S.J. Qin (2002). Semiconductor process monitoring and fault detection with recursive multiway pca based on a combined index. In: *AEC/APC Symposium XIV*. Salt Lake City, Utah.

Del Castillo, E. and A. Hurwitz (1997). Run-to-run process control: Literature review and extensions. *Journal of Quality Technology* **29**, 184–196.

Del Castillo, E. and J.Y. Yeh (1998). An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes. *IEEE Transactions on Semiconductor Manufacturing* **11**, 285–295.

Del Castillo, E. and R. Rajagopal (2002). A multivariate double EWMA process adjustment

scheme for drifting processes. *IIE Transactions* **34**, 1055–1068.

Edgar, T.F., W.J. Campbell and C. Bode (1999). Model based control in microelectronics manufacturing. In: *Proceedings of the Conference on Decision and Control*. Vol. 38. pp. 4185–4191.

Good, R. and S.J. Qin (2002). Stability analysis of double EWMA run-to-run control with metrology delay. In: *Proceedings of the American Control Conference*. Anchorage, AK. pp. 2156–2161.

Good, R. and S.J. Qin (2003). On the stability of MIMO EWMA run-to-run controllers with metrology delay. *IEEE Transactions on Semiconductor Manufacturing*. submitted.

Hamby, E.S., P.T. Kabamba and P.P. Khargonekar (1998). A probabilistic approach to run-to-run control. *IEEE Transactions on Semiconductor Manufacturing* **11(4)**, 654–669.

Harrison, Christopher A., Richard Good, Daniel Kadosh and S. Joe Qin (2003). Multi-step supervisory control of flash memory device production via a simple first-principles model. In: *AEC/APC Symposium XV*. Colorado Spring, Denver.

Ingolfsson, A. and E. Sachs (1993). Stability and sensitivity of an EWMA controller. *Journal of Quality Technology* **25 (4)**, 271–287.

Lee, S.F. and C.J. Spanos (1995). Prediction of wafer state after plasma processing using real-time tool data. *IEEE Transactions on Semiconductor Manufacturing*.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice-Hall, Inc.. Englewood Cliffs, New Jersey.

May, G.S., J. Huang and C.J. Spanos (1991). Statistical experimental design in plasma etch modeling. *IEEE Transactions on Semiconductor Manufacturing* **4**, 83–98.

McNabb, C. A. and S. Joe Qin (2004). Fault diagnosis in the control invariant subspace of closed-loop systems. *Ind. Eng. Chem. Res.* to be submitted.

Moyne, J., E. del Castillo and A. M. Hurwitz (2001). *Run-to-Run Control in Semiconductor Manufacturing*. CRC Press.

Nomikos, P. and J.F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics* **37**(1), 41–59.

Qin, S. J. and R. Dunia (2000). Determining the number of principal components for best reconstruction. *J. Proc. Cont.* **10**, 245–250.

Qin, S. J. and T. Sonderman (2002). From chemical process control to semiconductor manufacturing control. In: *Keynote at the AEC/APC Symposium XIV*. Salt Lake City, Utah.

Qin, S. Joe and T.A. Badgwell (1997). An overview of industrial model predictive control technology. In: *Chemical Process Control - V* (J. Kantor, C. Garcia and B. Carnahan, Eds.). Fifth International Conference on Chemical Process Control. CACHE and AIChE. pp. 232–256.

Sachs, E., A. Hu and A. Ingolfsson (1991). Modeling and control of a epitaxial silicon deposition process with step disturbance. In: *Proceedings of 1991 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*. pp. 104–107.

Sachs, E., A. Hu and A. Ingolfsson (1995). Run by run process control: Combining SPC and feedback control. *IEEE Transactions on Semiconductor Manufacturing* **8**, 26–43.

Semiconductor Industry Association (2003). The international technology roadmap for semiconductors. San Jose, CA.

Sonderman, T. and C. Bode (2004). Automated precision manufacturing at AMD. In: *Spring Meeting of Texas-Wisconsin Modeling and Control Consortium*. http://www.che.utexas.edu/twmcc/.

Tseng, S.T., R.J. Chou and S.P. Lee (2002). A study on a multivariate EWMA controller. *IIE Transactions* **34**, 541–549.

Wang, Jin, Q. Peter He, S. Joe Qin, Christopher A. Bode and Matthew A. Purdy (2004). Recursive least squares estimation for run-to-run control with metrology delay and its application to an STI etch process. *IEEE Trans. on Semiconductor Manufacturing*. revised for publication.

Yue, H. and S. Joe Qin (2001). Reconstruction based fault identification using a combined index. *Ind. Eng. Chem. Res.* **40**, 4403–4414.

Yue, H., S.J. Qin, J. Wiseman and A. Toprac (2001). Plasma etching endpoint detection using multiple wavelengths for small open-area wafers. *J. of Vacuum Science and Technology A* **19**, 66–75.

Yue, H., S.J. Qin, R. Markle, C. Nauert and M. Gatto (2000). Fault detection of plasma etchers using optical emission spectra. *IEEE Trans. on Semiconductor Manufacturing* **13**, 374–385.

Zafiriou, Evanghelos, Hung-Wen Chiou and R.A. Adomaitis (1995). Nonlinear model-based run-to-run control for rapid thermal processing with unmeasured variable estimation. In: *187th Electrochemical Society Meeting*.

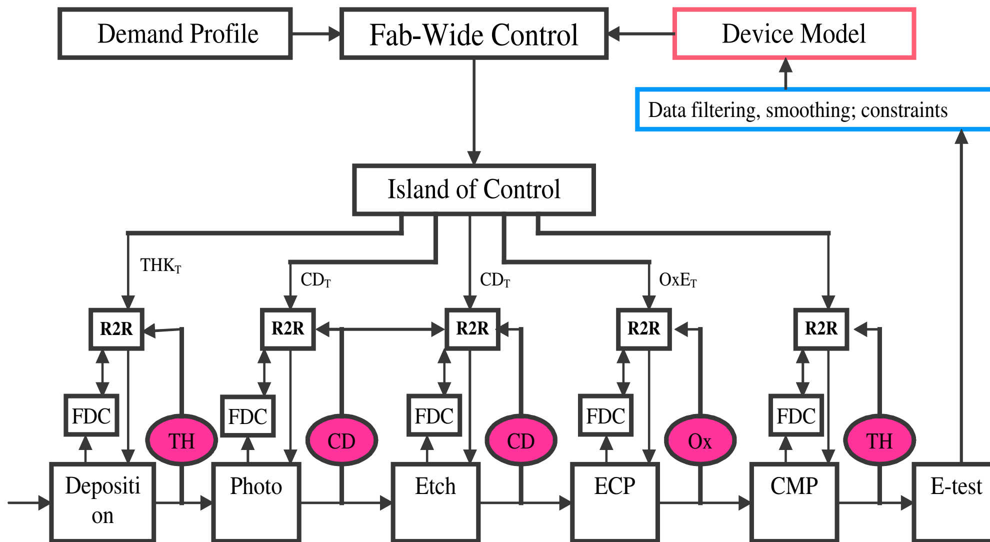

Fig. 1. Electrical parameter control as the top level of fab-wide control

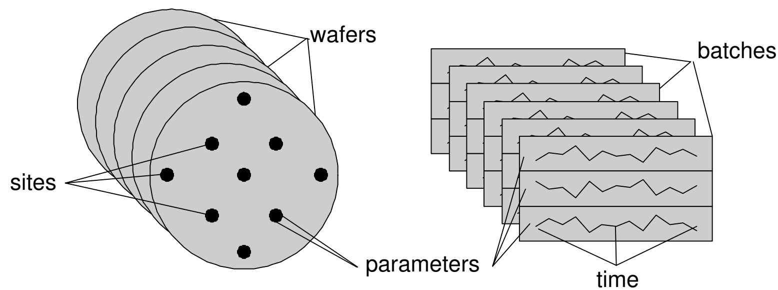

Fig. 2. Organization of site-level and batch data.

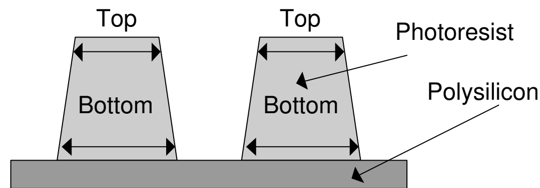

Fig. 3. Development inspection critical dimension (DICD).

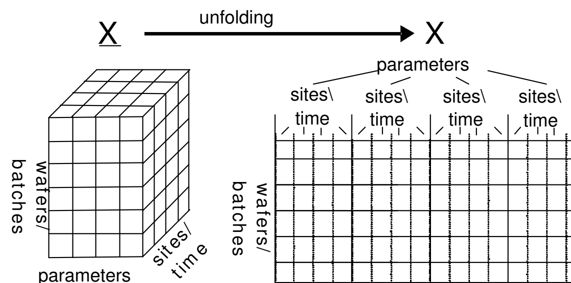

Fig. 4. Unfolding of site-level and batch data.

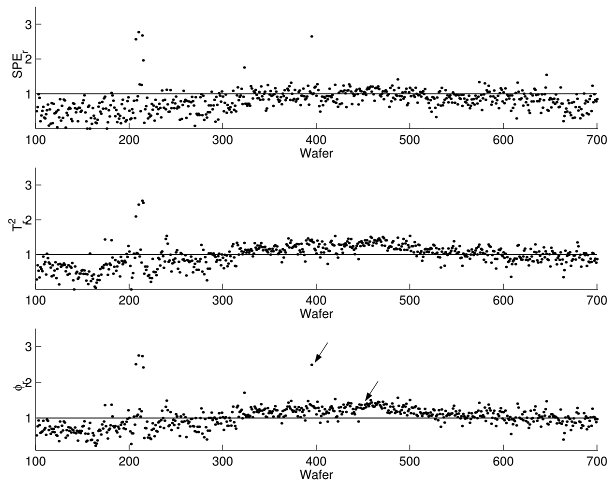

Fig. 5. DICD fault detection using SPE_r, T_r^2, and φ_r.

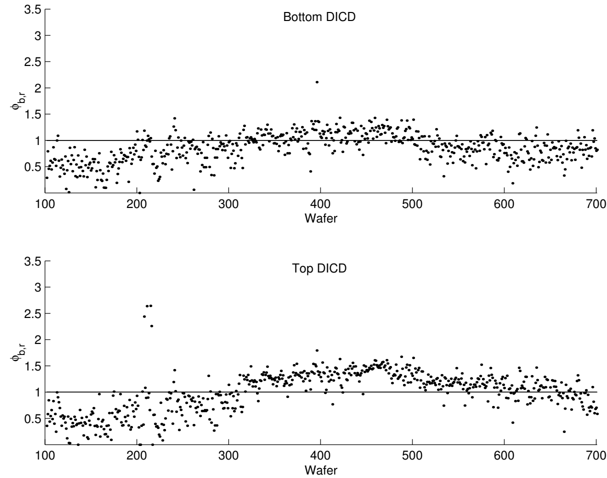

Fig. 6. DICD fault identification using parameter contributions.

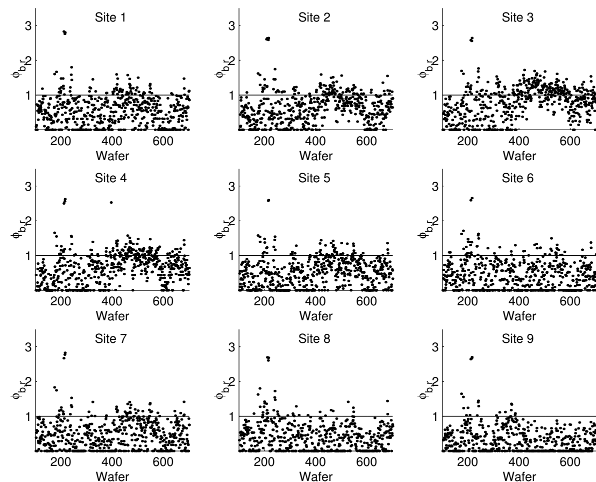

Fig. 7. DICD fault identification using site contributions.

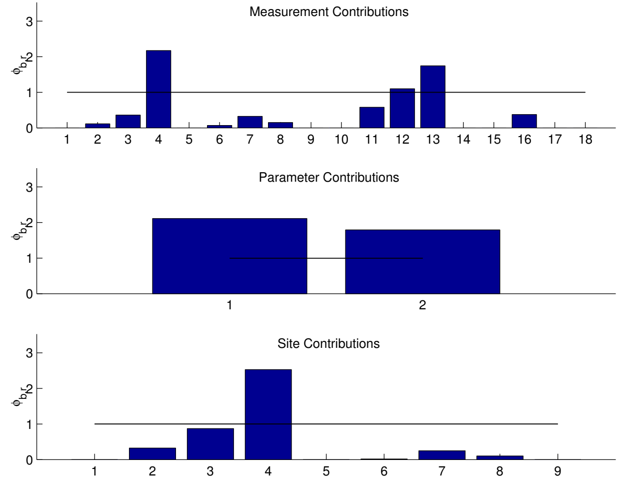

Fig. 8. Wafer 395 - DICD fault identification using measurement, parameter, and site contributions.

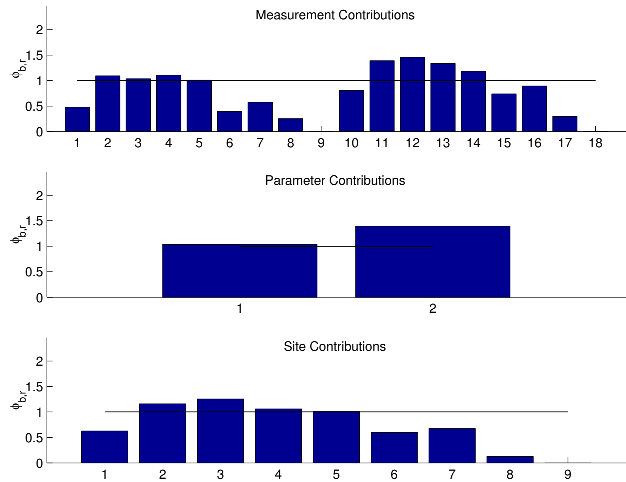

Fig. 9. Wafer 450 - DICD fault identification using measurement, parameter, and site contributions.