SEMICONDUCTOR FACTORY CONTROL AND OPTIMIZATION

The integrated circuit (IC) is the fundamental building block of modern electronics and is one of the most significant developments in technology of the twentieth century. The semiconductor industry, created 40 years ago, has fueled the high-tech industries that have changed the way that the world works, communicates, and plays today. Developments in semiconductor manufacturing have come about as a result of the increasing pace of scientific and technological breakthroughs and the rapidity with which they have been adopted for commercial production. Companies have learned that the market demand for faster, smaller consumer products with increased functionality determines their profitability and future growth. In today's market, several cycles for product releases exist simultaneously in different phases. The competitive advantage of companies is realized by shortening the time to market of each product release and anticipating the demands and opportunities of the marketplace. However, the increasing complexity and the shrinking cycle of product development and introduction to market also increase the risk of failure. Disruptions in this economic chain stemming from late deliveries or consumer recall can mean the difference between huge profits or catastrophic losses.

To prevent disruptions and reduce the time to achieve full-ramp product quality, many different monitoring and control methods are utilized in the modern semiconductor fab (factory). This synergistic combination of methods is known as factory control. The combination must provide coverage for a wide variety of possible sources of variation and abnormalities (control in breadth), as well as mitigate risk as early as possible (control in depth).

CONTROL IN BREADTH

Factory control in breadth is controlling all the factors in the wafer fab that have an impact on or may cause variation in the product characteristics. These sources are the "whats" that should be controlled in order to reduce product variability and to eliminate disruptions of product flow. These potential sources of variation for a typical wafer fab have been identified and classified on the Ishikawa (or fishbone) diagram shown in Fig. 1. Note that the diagram is generic and that it would be tailored to the type of technology of the wafer fab (e.g., bipolar versus MOS or mixed signal versus logic) Table 1 describes each branch of Fig. 1.

Considerably more space would be required to discuss all the methods used to control all the "whats" in Fig. 1. Consequently, just some of the key controllers will be examined in detail. These controllers are associated with the following branches: Methods (SPC, Outliers), Systems (Changes), and Technology (Defects). However, first one must understand the essential elements of any control system in order to comprehend the control systems put in place for any of the branches in Fig. 1.

GENERIC MODEL OF THE ELEMENTS OF A CONTROLLER

Figure 2 is a generic control model that illustrates the controller elements and their relationships to the process. The control cycle begins with a plan that provides instructions or actions for a process based on the input target value, feedforward data, and an expectation of how those actions will have an impact on the process. An example would the machine settings and conditions to achieve a target thickness on a deposition process. Machine sensors or measurements of the process output are compared with the expectation of the process to produce information. This feedback information regarding the state of the process is passed to the correction procedure that analyzes what type of corrective action should be taken to adjust or correct the problem. The feedback information may be either analog (e.g., the deposition rate is 10 A/s greater than expected) or digital (e.g., an indication of normal versus abnormal condition of the process). The digital aspect of control is more generally known as fault detection. The correction procedure is based upon the feedback information. One possible corrective procedure is to change the process by a given amount (e.g., to change the process time by 2 s) to achieve the

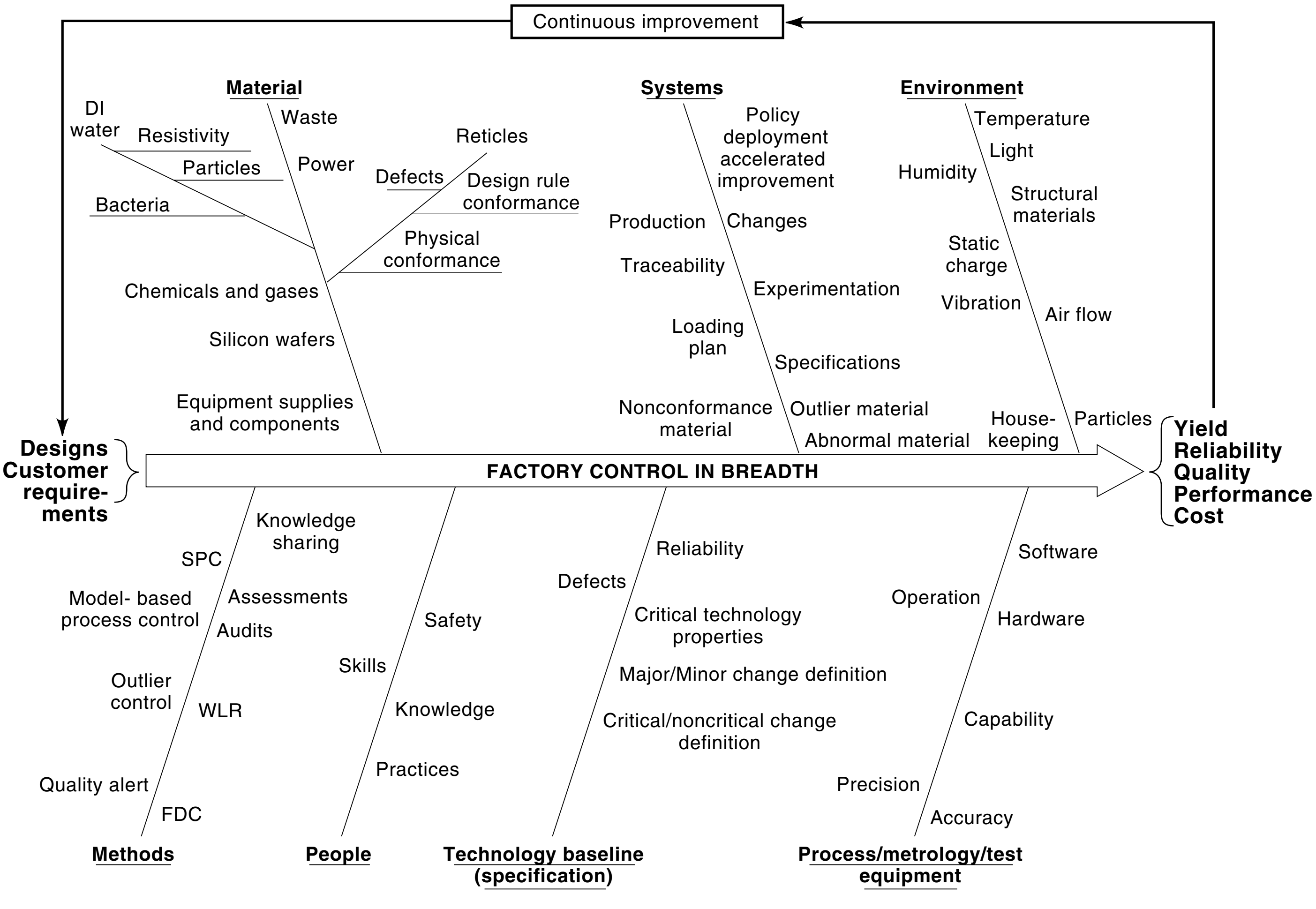

Figure 1. Ishikawa diagram illustrating the concept of factory control in breadth. All sources of variation are identified, mapped, and risk assessed to establish the controls within the wafer manufacturing factory.

Table 1. Definition of the Primary Sources of Variation Within a Wafer Fab from Figure 1

Branches	Definition
Environment	The conditions of a wafer fab to which wafers are exposed. Examples include temperature, humidity, light, airborne particles, air flow, static charge, structural materials, vibration, and housekeeping.
Material	Consumable items that are used in manufacturing semiconductors or in operating the wafer fab. Examples are silicon wafers, DI water, chemicals, gases, waste, power, reticles, and equipment supplies and components.
People	Personnel with responsibility for manufacturing or the operation of the wafer fab. Examples of people variations are skills, knowledge, and practices.
Equipment	All wafer fab machinery and hardware used in manufacturing, measuring, or testing of wafers. This includes test, process, and metrology equipment. Examples include accuracy, precision, capability, hardware, and software.
Methods	Standardized practices used to control or improve processes or factors of variations. Examples include statistical methods, model-based process control, audits, assessments, knowledge sharing, quality alerts, outlier control, and wafer-level reliability.
Systems	Policies, practices, procedures, and business automation used to effectively operate the wafer fab. Examples include production, specifications, experimentation, changes, classification and handling of material, traceability, policy deployment, and continuous improvement.
Technology Baseline	The electrical, reliability, and yield requirements and the fabrication process that *define* the product performance and characteristics. Examples of factors include defects, charging (for MOS), mobile ion contamination, major/minor changes, critical/noncritical changes, and critical technology properties.

desired results. For faults, the corrective procedure is usually first to confirm the abnormality and then, if the fault is confirmed, to perform maintenance on the offending machinery.

Also shown in Fig. 2 are two types of control: feedforward control and feedback control. Feedforward control uses the information from the previous process and enters material to make adjustments to drive the output of the current process to a desired target. An example of feedforward control is using the postpatterned feature size measurements to adjust the etch process to achieve the targeted feature size. The second type of control is feedback control, which uses the output information to adjust the procedure for the next processing. Feedback is also called closed loop control because of the loop created by the feedback information, and the correction action as shown in Fig. 2. Because feedback control is more widely practiced in the industry, future references to control systems will refer to feedback control.

Note that the model is a closed loop series consisting of action based upon initial information, data, new information, and corrective action. The time from when the fault or change occurs until when corrective action is implemented is the response time of the controller. This response time is an important measure of the risk of material in the process loop that may be in jeopardy if the output is very far off from target or if a fault has occurred. Thus, speeding or improving the quality of any of the components of the controller (feedforward information collection, procedure identification, information extraction, or measurement) can reduce the amount of material at risk. Thus, a controller's effectiveness is not only a function of the time to collect data after a fault has occurred, but also its ability to use that data to detect process shifts or events and decide what corrective actions to take. This idea of reducing risk by speeding data collection versus the data's innate information content about a fault is the foundation for the concept of control in depth.

CONTROL IN DEPTH

Testing the electrical function (known as multiprobe) of the integrated circuit provides the highest confidence that all the processes used in its manufacture are in control. Multiprobe

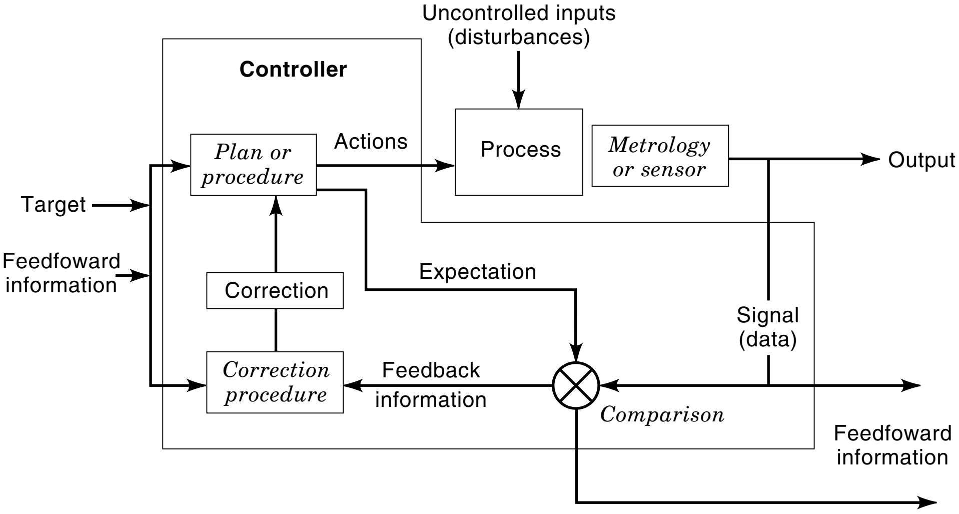

Figure 2. A generic control model showing the major elements of a closed loop control system. The response time is equal to the time from a fault or change occurring and the control system implementing a correction. The response time can be measured in the number of wafers at risk. Improving any of the components [i.e., speed/frequency of data collection, ability to detect a change (quality of data and comparator), or accuracy of the planning or correction procedure] will decrease the response time.

results are normally represented by the yield, defined as the number of "good" die divided by the number of "possible" die. Yield, which represents the overall integration of control for the wafer fab, has a direct impact on the financial cost of manufacturing. Therefore, it is the major index for driving improvement. Unfortunately, it can be very difficult to determine exactly which processes are the cause for reduced yield. As stated previously, an effective control system requires the ability to decide what actions to take. Therefore, to isolate information on the process results and interactions, electrical parametric testing of discrete devices (e.g., diodes, transistors) is essential. The parametric test structures can be embedded in the scribe lines between dies or special structures within the die itself. In addition, some individual dies on a product wafer may be entirely test structures. Besides test structures on product wafers, special short loop test wafers may also be used. These test structures allow the measurement of specific electrical parameters, such as gate oxide integrity, isolation, sheet resistance, and breakdown. Parametric testing produces a few parameters that can be compared with well-defined limits (or expectations) derived by simulation models and experimentation. In addition, these test structures provide some isolation of the fault to certain processes and films. Exactly what and how many test structures are used is a function of the maturity of the product. (See SEMICONDUCTOR MANUFACTURING TEST STRUCTURES for more details.)

Yield and parametric data provide accurate data relating to the control and capability of a wafer fab and may be useful in comparing fabs running the same technology. However, the effectiveness of a control system depends on its data collection time and its ability to detect process shifts or events. Although the use of yield and parametric data provides a control system with very good ability to detect faults or changes, the data collection time is very long because of the amount of time it takes to manufacture a device (3 to 12 weeks depending on complexity and maturity level). In addition, even with parametric data, the ability to decide what corrective actions to take can be difficult. Thus, other levels of controls must be established at the process level and equipment level for early detection of problems and easier linkage to specific machines and processes. The idea of yield and parametric data being used for comparing fabs but the bulk of the factory control system being at the process and equipment level is illustrated by the iceberg concept shown in Fig. 3. The tip represents the visible electrical parametric data, and the majority of the control indices internal to the fab are below the surface. Although the tip should be common for any factory running a particular product, what is below the surface is dependent upon the equipment, people, environment, systems, and materials used by a particular fab. To decide which process and equipment controllers to use, a systematic analysis must be performed to link each parametric variable, such as speed, to material properties, such as the physical dimensions of the polysilicon gate. In turn, how the fab's equipment, processes, and metrology affect each material property is estimated. Based upon a careful analysis of possible risks associated with each piece of equipment and the capability of the metrology, the overall impact to the parametric results can be assessed. Using this assessment, necessary process and equipment control methods can be determined.

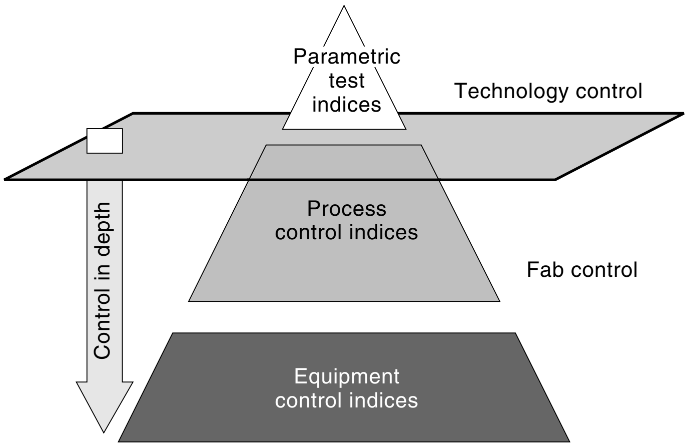

Figure 3. Control "iceberg" showing the levels of control. The pinnacle or "tip of the iceberg" is the electrical parametric indices. Because they have well-defined specification limits, they are good performance indices to compare wafer fabs running the same technology. Process controls and equipment controls indicated below the surface are internal fab controls tailored to each fab's equipment set and process capabilities.

Control systems in a wafer fab can be ranked by their associated risk and quality. Figure 4 illustrates this concept of risk versus quality for several methods of control. The risk on the left axis is measured in the approximate number of wafers in jeopardy from when the fault/change occurs to when the control method implements a corrective action (i.e., the response time of the control system). The horizontal axis in Fig. 4 shows the controller quality by using the concept of error rates (α, β). Table 2 demonstrates the concept of Type I and II errors and the associated error rates (α, β). For example, if in reality a result is good, in α of the cases, the statistical test will indicate a bad result. Conversely, if in reality a result is bad, in β of the cases, the test will indicate a good result. The "truth" for Fig. 4 to determine α and β is whether the device performs correctly in the customer's system. For Fig. 4, Test power ($1 - \beta$) is defined as the probability of detecting a process shift or a failure that results in defective material. Also represented along the bottom of Fig. 4 is the false positive rate (α) which is the probability of the control method saying a shift has occurred when in reality the final product is not impacted. For example, a controller using an in situ particle monitor may create an alarm based upon detecting an increased number of particles, but none of these particles actually deposit on the wafer in such a way as to cause the device to fail. Note that the values for α and β in Fig. 4 are only an approximation for illustrating the relationships of various control methods and the concept of control in depth. The true values of α and β may be quite different.

The right axis of Fig. 4 illustrates the classification of the levels of control: preventive, concurrent, and failure. Preventive control is the use of systems or actions taken to reduce variability or prevent abnormal conditions from occurring. Concurrent control is the use of systems that detect abnormal conditions or problems and that react to correct the problem before there is a high risk of material in jeopardy. Failure control refers to those systems that detect abnormal conditions or problems past the point of making corrections. Within failure control, there may be containment control, which pre-

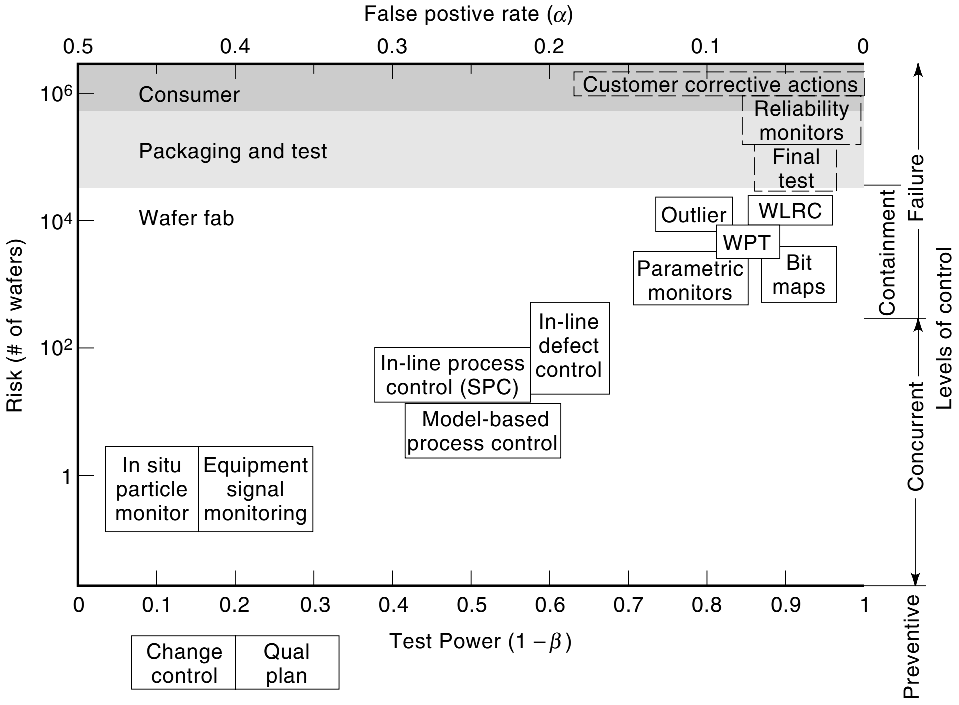

Figure 4. Factory control in depth is the methodology of defining a control system based on the risk of disruption to Test Power $(1 - \beta)$ and False Positive Rate (α). Tradeoffs between risk and confidence must be balanced with the economical costs of controls. Note that the values of β and α are for illustration purposes only; the true values could be different. The bold boxes will be discussed in more detail in later sections. (WPT = Wafer Position Tracking; WLRC = Wafer Level Reliability Control; Outlier = Multiprobe and Parametric Outlier Program.)

vents the abnormal material or problems from effecting the customer. Beyond the containment of the fab, other control methods are directed to discovering problems and providing corrective actions, for example, methods using customer-identified failures have the highest confidence, but unacceptable risk.

In order to minimize the risk of customer disruptions and to maximize yield, careful analysis is required to define an integrated control system that uses controls at each level to minimize risk and maximize information. Thus, different methods operating at different levels may be used for controlling the same source of variation. For example, in situ particle monitors, in-line defect control, and bit map matching are all at least partially focused on detecting and eliminating particles. Not only do the different levels provide risk reduction, data from the higher levels are also used to fine-tune the methods operating at lower levels in order to increase the power and decrease the false positives of the lower levels.

Further details on the selected control methods identified in Fig. 4 will be explained in the following sections. Change management will be discussed first as an example of a preventive method. In-line process statistical process control (SPC) will be discussed next because other control methods [wafer level reliability (WLR) and in-line defect control] use mathematics. Model-based process control will be discussed after SPC, and its relationship to in-line process SPC will be highlighted. Then equipment signal monitoring will be presented leading into a discussion on sensors. The sensor discussion ties into in situ particle monitors that lead to an overview of in-line defect control. In-line defect control also highlights the relationship with higher-level methods, such as bitmapping, which is then reviewed. The order of discussion of in situ particle monitors, in-line defect data, and bitmapping is done to stress how methods at higher and lower levels are aimed at the same source of variation. After bitmapping, the rest of the higher-order methods will be discussed, beginning with wafer position tracking, followed by data mining, outliers, and WLR control. Parametric monitors have already been discussed. The article will conclude with a very detailed discussion of multivariate SPC, predominately focused on its use in equipment signal monitoring. In general, hardware and software will not be discussed in detail because of the speed at which hardware and software are evolving. However, the Reading List provides contact information for major suppliers and Web pages dedicated to semiconductor manufacturing, especially defects and control. Because of the breadth of this article, the reader will probably encounter many new terms and acronyms. Thus, a glossary is provided at the end of the article to assist the reader.

Table 2. Explanation of Type I and Type II Error and Associated Distribution as Function of α, β Error Rates

	Reality	
Test Result	Good	Bad
Good	$(1 - \alpha)$	$\text{II} = \beta$
Bad	$\text{I} = \alpha$	$(1 - \beta)$

CHANGE MANAGEMENT

It is often a misconception that changes within wafer manufacturing are undesirable because deliberate changes must occur for continuous improvement, increased yield, and increased profit margins. Change control is a preventive control method to manage risk systematically and to obtain these results. It is essential that a predetermined methodology exists for making changes to each of the branches of factory control; material, systems, environment, people, equipment, technology baseline, and methods. An effective change control system will contain all of the elements of the generic control model in Fig. 2. Usually, a control procedure is defined for reoccurring changes such as processes, equipment (these procedures are sometimes called engineering or equipment change notices, ECNs), or new employees (orientation or termination procedures), specific to the type of change that is being made. Program management practices, which also follow the generic control model, are used for a one-time event change such as upgrading equipment to larger wafer sizes. The change result data are compared with the change expectation. This information is used to correct the change or actions to obtain the desired result (target). The corrective result may also change the procedure or program plan itself. An additional and important aspect of change control is the record of the change for traceability and dissemination of information. Knowing what, why, and when the change occurred is important if some of the side effects of the change are not discovered until later or if the reason for the change is no longer applicable. Communicating this information prior to initiating the change is important in order to get buy-in and other inputs relating to the effect of the change.

STATISTICAL PROCESS CONTROL

Statistical process control is the most widely used control method in a wafer fab. It has generally been as a quality program focused on eliminating product variations. However, SPC is a highly effective control tool that can be used to increase yield, reduce process variations, and minimize the impact of equipment failures. With respect to the control model presented in Fig. 4, SPC encompasses all the elements of the control model: the output data, the comparison with expectation, and the correction procedure. Each one of these subcomponents will be addressed in detail later.

Process, Data, and Expectation

Understanding the relation of the process factors (or inputs) and interaction to responses (outputs) of the process is of paramount importance prior to applying a statistical control system. Some of the tools that can be used to identify and study this relationship are design of experiments (DOE), failure mode and effects analysis (FMEA) (1), quality function deployment (QFD) (2), computer simulations, fault-tree analysis (3), cause-and-effect analysis, and analysis of the variance (ANOVA). DOE is the key tool for determining the critical factors affecting the output target values, choosing optimum settings for the factors and building empirical models of the process that can be used for adjusting the process back into control. In a series of carefully designed experimental runs, the levels of many factors can be simultaneously varied, and the effects can be observed on the resulting responses. DOE can make its most dramatic contribution in the design phase of process, when it is least expensive to make changes. (See STATISTICAL METHODS FOR SEMICONDUCTOR MANUFACTURING for more information on design of experiments.)

Comparison

The function of the comparitor in SPC control is to determine whether the process state is in control or not in control. The detection of a change in the control state (i.e., a process drift or failure event) is dependent on the noise, or variation, of the data, the sampling frequency, and the sensitivity of the filter (i.e., the type of SPC chart and the alarm settings). If there is large variation in the process or the measurement system, then any signal indicating a drift or special cause event will be masked by the noise and not detected. Likewise, if the incorrect SPC chart is applied and insufficient alarm settings are used, then the out-of-control signal will not be detected. However, if filter is too sensitive (i.e., too many alarm levels), then there will be frequent, false, out-of-control events.

Gauge Studies. Understanding the measurement systems contribution to the total variation is of paramount importance to ensure that the control system does not respond to the noise of the measurement system. Sources of variation could consist of bias, repeatability, reproducibility, and linearity. Bias, or accuracy, is the difference between the observed average of measurements and the reference value. Linearity is the difference in bias values over the range of the measurement system. Repeatability is the variation of the measurement system under identical conditions. Reproducibility is the variation of the measurement system induced by different conditions (e.g., operator, location). Gauge repeatability and reproducibility (GR&R) (4,5) studies will determine if the measurement system is acceptable for control purposes. Whether a measurement system is satisfactory depends largely on the percentage of tolerance that is consumed by the measurement system variation. This is expressed as %GR&R or as a measurement capability (Cp) index. The generally acceptable ranges of measurement Cp or %GR&R are listed in Table 3. The equations for Cp and %GR&R follow:

$$\sigma_{\mathrm{R\&R}} = \sqrt{(S_{\mathrm{r}})^2 + (S_{\mathrm{R}})^2} \tag{1}$$

$$\%\mathrm{GR\&R} = \frac{6 \times \sigma_{\mathrm{R\&R}}}{\mathrm{USL} - \mathrm{LSL}} \times 100\% \tag{2}$$

where

S_{r} is the standard deviation for repeatability
S_{R} is the standard deviation for repeatability
USL, LSL are the upper and lower specification limits, respectively

$$\text{Measurement Cp} = \frac{1}{\%\mathrm{GR\&R}} \times 100 \tag{3}$$

Sampling Plans, Univariate SPC Charts, and Alarm Rules. The bulk of data collected daily in most wafer fab operations may not be time or cost efficient. The proper choice of a representative sample from the population allows predictions about the process and its state. The objective of defining a sampling plan is to provide accurate process information while decreas-

Table 3. The Criteria for Acceptance of Gauge Repeatability and Reproducibility

Measurement Cp	%GR&R	Rating
Cp < 3	%GR&R > 33%	Unacceptable
3 ≤ Cp ≤ 10	10% ≤ %GR&R ≤ 33%	Marginal
Cp ≥ 10	%GR&R ≤ 10%	Acceptable

Table 4. Most Commonly Used Control Chart Types for Continuous Data in a Wafer Fab

Control Chart Type	Subgroup Size	Data Plotted	Typical Use
Xbar and Range	$2 \le n \le 5$	Averages and ranges of subgroups	Process Control
Xbar and Sigma	$n > 2$	Averages and standard deviation of subgroups	Process Control
X-Moving Range	$n = 1$	Individuals data and moving ranges of individuals data	Process Control
X-Sigma	$n = 1$	Individual data	Process Control
Xbar-Moving Range	$n > 1$	Averages of subgroups	Process Control
Xbar-Moving Range and Range	$n > 1$	Averages and moving ranges of averages	Process Control

ing the production cost. Analysis of variance (ANOVA) is used to analyze the different sources of variation in a process and to determine the proper subgroups for control charting. A sampling plan should be selected so that if assignable causes are present, the chance for differences between subgroups will be maximized, while the chance for differences resulting from these assignable causes within a subgroup will be minimized. For example, most wafer fab manufacturing processes are run in a batch rather than a continuous flow. This results in hierarchical, or nested, design structure, where each run, lot, wafer, and measurement is a unique term adding to the total variability. If the run-to-run variability is the greatest, then the sampling plan should be based on run-to-run samples rather than lot-to-lot samples.

The effectiveness of SPC (6–9) depends in a large part on the selection of the control chart. Process data can be classified as four types: a defect, which is an individual failure to a specification; a defective, which is a unit of product that contains one or more defects; variable data, which can be measured on a continuous scale; and attribute data, which can be classified as either conforming or not conforming. The control chart type selection is based on the type of data, sampling method, and the type of variation observed. Tables 4 and 5 list the most commonly used univariate control chart types for variable and discrete data. Univariate denotes a single variable. Most SPC charts used today are univariate. Multivariate SPC charts will be discussed at the end of the article. Note that an underlying assumption for the charts which use groups of data is that the within-subgroup variation is the same as the subgroup-to-subgroup variation in charts based on subgroups (such as the XBar, R). Because the process has considerable systematic nonuniformity across the wafer and the metrology is wafer-based, such an assumption is rarely true in semiconductor processing, where "natural" subgrouping would be at sites on a wafer. The random lot-to-lot variation is not the same as the random variation across a wafer. In addition, the variation across the wafer is mainly the result of systematic nonuniformities of the process rather than random behavior. Thus, charts for individuals usually are more appropriate for the semiconductor processing industry.

The final component of the comparator subsystem is the alarm levels or trigger conditions for indicating an out of control condition. The Western Electric (WECO) (9) rules are the most generally used rules. Referring to Fig. 5, they are: (1) one point outside of the control limits; (2) two out of three successive points on the same side of the centerline in Zone A or beyond; (3) four out of five successive points on the same side of the centerline in Zone B or beyond; (4) eight successive points on one side of the centerline; and (5) seven consecutive points increasing or decreasing. Note that not all situations warrant all the rules applied. Underusage of the appropriate rules will lower the sensitivity to detect changes and faults, whereas overusage will cause the controller to overreact.

Corrective Procedure

The last component of the SPC controller is the corrective procedure. If an out-of-control event has been determined, the process should be stopped and a corrective procedure initiated promptly. The correction procedure should contain diagnostic procedures with associated recommended actions. Typically, these contain a hierarchy of different levels of authorized actions, which specifies which conditions allow different levels of authority to make corrective actions. For example, an operator may be required to verify the metrology and equipment settings, whereas the authorization to stop production may be given only by the supervisor. If the process has been well characterized, there may be one or more settings that can be adjusted to bring the output back to its target value.

Qual Plans

A formal procedure for implementing SPC and qualifying a process is typically termed a qual plan (7,10). Part of the qual plan would include performing a gauge study and determining the sampling plan, both of which were discussed earlier. The importance of executing a formal qual plan has even led to the marketing of software for this specific purpose (11). A qual plan is a control method that is considered a preventive control because it involves techniques to prevent the installation of a process that could easily produce scrap. Thus, good qual plans that are executed well will result in effective

Table 5. Most Commonly Used Control Chart Types for Discrete Data Such as Particle Count Data and Yield

Control Chart Type	Subgroup Size	Data Plotted	Typical Use
C Chart	Constant	Number of defects	Product Inspection
U Chart	Constant or variable	Average number of defects per item	Product Inspection
NP Chart	Constant	Number of defective items	Product Inspection
P Chart	Constant or variable	Percentage of defective items	Product Inspection

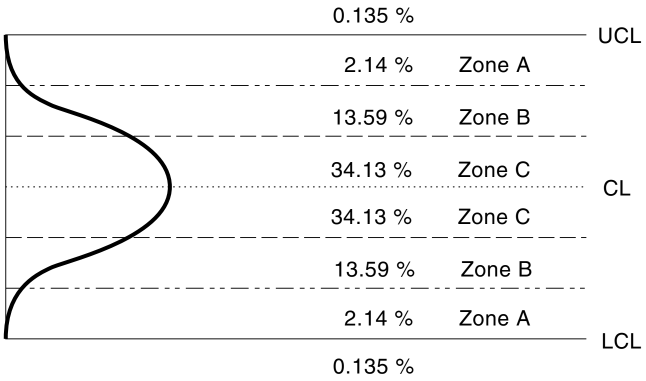

Figure 5. WECO rules are a method for triggering out-of-control events. Each zone represents one standard deviation of normal variation of the process. The percentage indicates how much of the data will be contained in that zone, based on the probabilities for a normal distribution.

SPC and fewer process problems. As model-based process control (MBPC) becomes more widespread, qual plans are being modified to include steps required for successful MBPC implementation.

RUN-TO-RUN MODEL-BASED PROCESS CONTROL

Historically, the process recipe, the set of setpoints for the equipment, does not vary from batch to batch. Correspondingly, in traditional SPC, the process is assumed not to drift or shift in typical behavior. However, many processes do experience drifts or shifts, and such behavior is not considered a “fault” by the process or equipment engineers. Even though the drift or shift is not a fault, it does cause undesired variation in the product. Therefore, a technique is needed to compensate for this undesired variation by varying the recipe on an as-needed basis to maintain a constant output. This technique is known as model-based process control because models are used to describe the expected process behavior (12). As the process shifts or drifts, the models are tuned to predict the new output. The tuned model is used to decide how to change the recipe to counteract the shift or drift. MBPC is also known as run-to-run (RtR) control because the recipe is changed on a run-to-run basis, if need be. This contrasts with the real-time controllers on the equipment that change actuators during processing to maintain the process on setpoint. Real-time controllers may also be model based, but are not discussed further because they are in the jurisdiction of the equipment supplier.

In comparison with traditional in-line SPC, the data sources for RtR MBPC are the same. However, the information filters and procedure for determining corrective action are different. SPC charts may still be used to determine when a shift or drift has occurred and the model should be tuned (13). In addition, SPC concepts are employed to decide whether the recipe should be changed, or the process behavior has changed drastically and manual repair should be performed. MBPC results in fewer wafers at risk than traditional in-line SPC because the fab typically maintains tighter management of measurement and control actions for MBPC because the results are needed to determine how to run future lots, not just decide whether to shut down the tool.

EQUIPMENT SIGNAL MONITORING, REAL-TIME FAULT DETECTION AND CLASSIFICATION

Most modern processing equipment has a semiconductor equipment communication standard (SECS) port that allows collection of up to approximately 50 different variables (signals, traces) once per second on many machines. In addition, some signals may be gathered using hard-wiring (i.e., splicing into a signal line to obtain the data). Monitoring of these signals is most common in etch, furnaces, CVD, PVD, and implant. In other words, monitoring using data from the SECS port is common in all areas except lithography, but that situation is expected to change in the near future. The equipment variables that provide the most information are actuators used in a real-time feedback control loop on the processing equipment (e.g., a throttle valve used to control pressure) and noncontrol process measurements (e.g., dc bias or an uncontrolled chuck temperature). Sensors that have been added on to the tool also provide real-time traces (for a list of possible sensors, see the section on Sensors).

Monitoring equipment signals to detect a change in the process or tool has come to be known as fault detection and classification (FDC). However, fault detection can occur with any data, and thus real-time fault detection or real-time SPC is a more appropriate name. Also, currently, classification of the fault to a source is rare, but the name FDC is still commonly used. Note that the term *real-time* denotes that the signals are from traces, not that the analysis and interdiction occur in realtime (i.e., analysis and shutdown may occur post-processing rather than during processing of the wafer). Currently, three methods for fault detection are common:

- Guardbanding
- Multiple univariate SPC charts of metrics created from the trace
- Multivariate SPC of metrics created from the trace

In guardbanding, a reference trace is used with a guardband, a zone of %-*X*% around the reference trace. The process trace is compared with the reference trace and the number of out-of-zone samples are counted. If the total number of out of zone samples is greater than a threshold value, then a fault is declared. Because the process time may vary because of automatic endpointing and varying incoming wafer states, dynamic time warping may also be used to stretch or shrink the process trace to match it up with the reference trace (see the subsection on Trace Analysis). Regardless of the fault detection method, once a fault is declared, interdiction may occur. Interdiction normally is to shut down the equipment and perform a diagnostic procedure, similar to that discussed in the section on SPC.

For the other two methods, the trace is decomposed into metrics, such as the average throttle valve position during step 2 of the process and the standard deviation during step 1 of the process. Thus, a single trace for one variable can be decomposed into several metrics. Because the signal may not decompose easily using step number, dynamic time warping may be used to identify the region boundaries. The generated metrics are then used similar to data obtained with in-line measurement tools (i.e., used in SPC charts). However, because of the volume of metrics generated, it requires multiple univariate charts or true multivariate SPC schemes. Because of the level of detail that will be presented, a discussion of multivariate SPC and the challenges of using multiple univariate charts will be done at the end of the article.

SENSORS

Equipment signal monitoring is usually performed with sensors supplied with the equipment. These sensors can be divided into two classes: machine sensors and process sensors. Machine sensors that measure some aspect of a machine actuator setting, such as throttle valve position, capacitor positions, and supplied power. The actuator is usually used in a closed loop controller, such as temperature or pressure control. Process sensors, such as dc bias, pressure, and temperature, measure a result of the equipment and the wafer states. They may, such as for pressure, or may not, such as for dc bias, be controlled by a feedback control loop.

In the past, all sensors were those provided by the equipment supplier. They generally were very limited in number. A new source of sensors is appearing as companies are forming which sell sensors directly to the end user, as well as equipment suppliers. These sensors many times provide more visibility into the process and wafer states. Thus, they provide better measurements for use in feedback control and fault detection. A common process sensor that is beginning to mature is optical emission spectroscopy (OES) (14). Even though single wavelength optical emission spectroscopy has been used for years to endpoint plasma-based processes, only recently has multiwavelength shown promise as being appropriate for the manufacturing environment. Newer processes, such as chemical mechanical polishing (CMP), are also driving development of sensors for measuring both thickness in situ and in-line (i.e., on the tool) but not in the processing chamber (15). Measurement of uniformity is increasing in importance because of the switch to 300 mm wafers. Thus, sensors aimed at uniformity measurements are becoming available, such as the NOVA CMP sensor (15) and the Liebold Full Wafer Interferometer for etch, which uses the light of the plasma to generate an interferometric signal (16). Temperature measurement of the wafer itself is being driven by rapid thermal processing (17,18). One key for success is that the equipment supplier provides necessary kits so that the sensors can be mounted. Such is happening for both the NOVA CMP sensor and, for some etch suppliers, the full wafer interferometer. Some sensors are modifications of existing sensors but with modifications to the hardware or increased algorithmic capabilities (19). The use of advanced mathematics, such as Kalman filters, is also bringing new opportunities to older sensor technology, such as lithography development interferometers (20). Monitoring of the delivery system for contaminants is also now becoming popular because of the availability of the sensors and the increased importance of contamination control (21,22). Development continues in combining novel mathematics with novel sensor technology to allow for key measurements in lithography (23,24). Besides CMP sensors, OES, temperature sensors, and mass spectrometry, including residual gas analyzers (RGAs), the other sensor that is making its way into manufacturing is the RF sensor (25). The RF sensor measures the RF signals either before or after the matching network on plasma systems. The actual measurement of delivered power is demonstrating the potential for tighter control. In addition, fault and endpoint information is being discovered in the harmonic signals. Some of the sensors, while unsuited for use in a manufacturing environment, provide useful information for process development in the R&D environment. An example of such a sensor is the Langmuir probe, which provides valuable information about the electron density (26). Another type of sensor is aimed at monitoring particles in the equipment, known as in situ particle monitors.

IN SITU PARTICLE MONITORS

In situ particle monitors (ISPMs) represent a focus on tool-based defect detection rather than on wafer-based defect detection. In reference to Fig. 4, ISPMs reduce the number of wafers at risk compared with in-line defect control methods. However, the false positive rate is higher with ISPMs, and the power may be lower. ISPMs are sensors placed on processing tool hardware, such as an exhaust line or a recirculation line in a wet process, to truly detect defects as they occur during wafer processing. They consist of a laser that is perpendicular to the flow of air/process gas. As particles pass through the laser beam, they reflect light into a sensor and are counted. ISPMs are small and much less expensive than defect detection tools; consequently, they are being used more and more in modern fabs in an effort to move even closer to monitoring of the sources of defects. A recent article summarizes several successful applications of ISPM (27).

IN-LINE DEFECT MONITORING AND CONTAMINATION CONTROL

Controlling defects during every processing step of semiconductor devices is vital to successfully manufacturing modern integrated circuits. The requirements for tight defect control become increasingly severe with each new generation of semiconductors. Not only must the total number of defects on wafers decrease with each generation, but the defect concentration per mask level must be reduced at an even faster rate because of higher circuit complexity and increased number of mask levels (Table 6). These defect reduction requirements are for DRAMs, commonly used as the technology driver, but must also be achieved in other device families such as ASICs and microprocessors.

In this article, the words *particle, defect,* and *contamination* are used interchangeably. Particulate that falls on a wafer during processing, chemical corrosion, moisture, and pattern anomalies such as missing pattern or extra pattern are but a few examples. Even though they each have their own definition, all are unwanted in semiconductor processing and are treated as one problem here. Particulate contamination in semiconductor processing arises from four general sources: clean rooms, people, equipment, and processes. Although the sources have remained the same over the past decade, the percentage of particles from each has changed quite dramatically. For example, in the mid-1980s, clean room/people and equipment/processes each contributed about an equal amount of particulate. Ten years later, however, the clean rooms have

Table 6. Device Manufacturing Trends: Killing Defect Size Versus Minimum Feature (from 1997 National Technology Roadmap for Semiconductors)

Year of First Product Shipment	1997	1999	2001	2003	2006	2009	2012
Technology generation (nm)	250	180	150	130	100	70	50
Critical defect size (nm)	125	90	75	65	50	35	25
Chip area (mm^2)	300	340	385	430	520	620	750
Mask levels	22	23	23	24	25	27	28
Faults per mask level	88	74	66	56	45	35	28

become much cleaner, as good as Class 1. (Clean room classifications relate to the number of particles per cubic meter of air at a specified particle size and are typically cleaner by orders of magnitude than hospital surgical rooms.) Better clean room garments plus reduced people interaction by use of wafer-handling robotics have reduced the contribution of clean rooms and people to less than 10%. Equipment and processes now have a greater contribution of particles, with processes themselves projected to be the greatest contributor by the year 2001. One method of improving die yield is obviously to reduce particle levels in equipment. The most practiced method is to process in vacuum. The increase in vacuum processing is trending higher, and providing clean processes in vacuum will continue to challenge equipment suppliers for many years to come. Another source of contamination is molecular contaminants such as organics, metals, ions, molecules, and other species that can adsorb to a wafer surface. Metal-ion contamination is also known as mobile ion contamination and is another major issue in wafer processing in that it can diffuse or migrate through silicon and destroy electrical functionality of an integrated circuit. This type of contaminant requires an entirely different set of tools for detection and analysis, and is treated in another chapter (see CLEANING/SURFACE PREPARATION).

Defects have one very important aspect: killing or nonkilling. A killing defect is any kind of defect that destroys the electrical functionality of a device and renders it useless. A nonkilling defect does not affect the electrical functionality of a device and is sometimes viewed as a less serious problem. A nonkilling defect could be in the scribe line between devices or in an open area on the device where there is no active circuitry, or it could be a particle that is removed from the wafer in a clean-up step. Although some fabs are only interested in killing defects, *any* defect is a potential killer, and all attempts should be made to eliminate the defect itself as well as the source. Nuisance defects are "defects" detected by the defect detection tool but that do not actually exist and are artifacts of the defect detection technology. Nuisance counts arise from such process conditions as color variation, metal grain size, or pattern nonuniformity and are not considered true defects.

Tools for Defect Detection, Classification, and Analysis

It is crucial to have the correct tool set to meet the fab-specific requirements for defect detection, both on production and unpatterned wafers. Equally critical are defect review and analysis tools, as well as a methodology that uses all these tools in harmony to deliver the most reliable and complete analysis and data set possible. Furthermore, the production wafer defect detection tools need to have the capability to operate for extended periods of time with few false positives (α), represented by the nuisance counts, and high power ($1 - \beta$), represented by a high defect capture rate.

Classical wafer-based defect detection tools fall into two broad genres: optical image comparison/analysis and laser-based light scattering. Optical image tools use a comparison algorithm and image subtraction across identical structures, either in the same die (memory cells) or across a row of dies (random logic circuitry) to identify portions of the image that do not match the identical structures surrounding it. The tools typically use visible light of either a narrow or broad band of wavelengths. The optics path closely resembles that of a high power microscope, except that the image is fed into a 1-D detector, such as a line of charge-coupled device (CCD) detector. Images are taken by scanning the wafer, line by line, across the fixed optics path and feeding in image data as grayscales to a powerful image processor. Image clean-up/filtering, image subtraction, and application of the set defect thresholds are all done on the image processor, the power of which is a limiting factor for the speed of the tool. Sensitivity depends on the magnification optics. Higher magnification gives greater resolution of smaller defects, but it also increases the scan time for a wafer. Such tools are probably the best in terms of absolute defect capture rate, but they often have higher nuisance rates and are slower than laser-based tools. This type of tool can typically scan an 8 in. diameter wafer in 5 to 20 min, depending on the sensitivity required for that device type.

Laser-based tools work on the principle of light scattering off defects in a way that distinguishes them from the normal pattern of the wafer. Tools for production monitoring invariably have the laser scanning across the wafer at a small oblique angle (2° to 3°). Defects rising above the standard pattern level of the circuit will scatter light at angles other than the angle of the main reflected beam. Dark field detectors at key locations will pick up this light and apply the programmed threshold levels to identify the defects. This type of tool can typically scan an 8 in. diameter wafer in 2 to 5 min. However, in general, this type of tool is less sensitive than the optical imaging tool.

Laser scattering-based tools have actually been in use for many years in the unpatterned wafer market (and later developed into the patterned wafer inspection market). In general, unpatterned wafer inspection tools are much more sensitive simply because there is no need to filter out pattern effects (anything that is not flat silicon or films is a defect); defects under 0.1 μm can be detected on bare silicon. Unpatterned wafer inspection tools come in two varieties: laser with normal incidence (for bare silicon and smooth films) and laser with oblique incidence (for rough films and metals). The normal incidence tools give higher sensitivity to smaller defects, including stacking faults or small pits in the silicon. The oblique incidence tools have a grazing angle of 2° to 3° in order to minimize effects of grain size and film roughness (such as in tungsten chemical vapor deposition and rugged poly).

Simply detecting defects is only the first of many steps in contamination-free manufacturing (CFM) practices. Review of the defects to identify their visual properties is the next logical step. Review information will quickly identify the true defects from any nuisance defects that may have been detected from an inspection recipe that was too sensitive. Classical review tools have been optical microscopes with a computer-controlled stage. Defect coordinates from either the defect detection tool or some central defect database are downloaded to the review station and translated to the coordinate system used by the review tool, and the appropriate wafer is loaded. The user will then align the wafer to the die corners, pick a sample (or all) of the defects to review, and proceed to classify the defects manually according to some preset codes developed by the fab. New advances in optical review stations include confocal optics for suppression of out-of-focus features and integration of laser imaging.

With increasingly small device geometries, we must be concerned about increasingly small defects. Optical review, even with new advancements, is limited by the wavelengths of optical light (4000 Å to 7000 Å or 0.4 μm to 0.7 μm). Even now, a large portion of defect review, especially for new or unknown defects, is done on a scanning electron microscope (SEM), where resolution is 100 Å or better. See Fig. 6 for a comparison of optical versus SEM review tool capability. The defect review SEM is an especially powerful tool because of its ability to do much more than just provide a high-resolution image of the defect. Integrated X-ray analysis, usually by energy dispersive spectroscopy (EDS), has been a mainstay of SEM tools for many years. With such integrated capability, composition of particulate contamination can be quickly and easily identified, which is a key piece of information to tracking down the root cause or source tool. Typically such defect review SEMs also include tilt capability. Newer models also include a focused ion beam (FIB) for in-line cross sectioning of defects.

Manual or automatic classification of the defects (based on training from previous defect data) will give the next level of information needed to identify excursions of a particularly crucial defect (such as blocked etch or peeling films) or to give clues about the root cause of a new defect type. An experienced and trained technician can perform optical review very quickly. However, because human judgment is inconsistent from person to person and day to day, there has been a strong movement in the past few years to move to automatic defect classification (ADC). Automatic defect classification algorithms are now available on defect detection, optical review, and SEM tools. ADC is mostly software that uses the visual attributes of a defect to determine a classification. Some of these attributes are color, shape, elongation, contrast, and size. ADC uses an image obtained from defects during either inspection or review, applies the algorithm, and determines what the defect is, based on a training set of similar defects. ADC takes from 2 to 15 s to arrive at a classification, depending on the algorithm. ADC is performed on production wafers at various inspection steps, and allows the process engineer to arrive at the root cause of a processing problem at the time it occurs.

All these tools and methods are suited especially well for defect detection, review, and analysis on production wafers sampled in-line. However, all these methods can also be used for unpatterned pilot wafers to obtain defect information for an individual tool. Unpatterned defect detection tools have been in use much longer as a result of the relatively simple challenges of detecting defects on a smooth surface, compared with one covered with complex circuitry. Recent optical review and SEM analysis tools all have the capability of working with unpatterned wafers as well as production wafers. However, for unpatterned wafers, the fine alignment of the wafer to the coordinate system must be done with the defects themselves, instead of die corners or alignment marks. For this to happen, there must be at least a few defects large enough to be found at low magnification before fine alignment is done. The trend in current manufacturing is to eliminate unpatterned wafers for routine monitoring. Unpatterned wafers add extra cost, take extra time, and tie up tools needed for production wafers. Semiconductor makers are finding

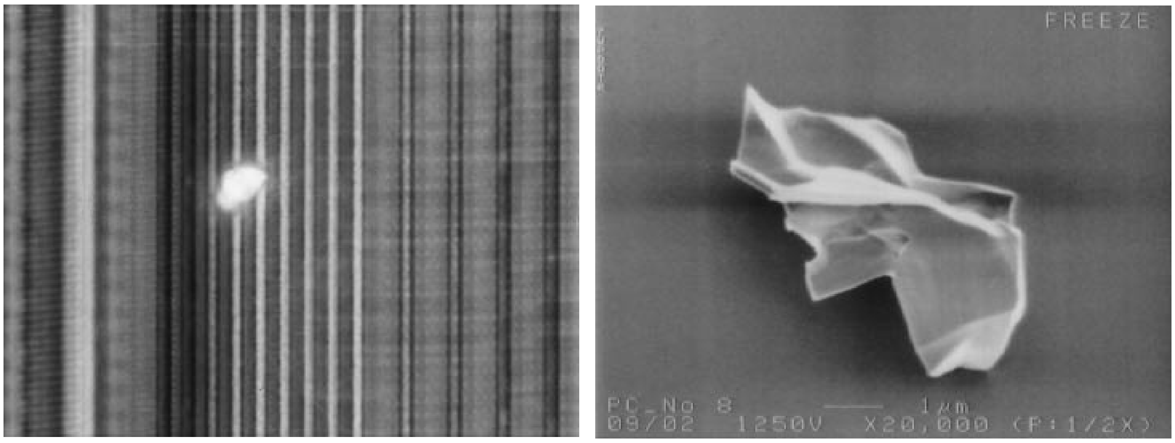

Figure 6. Contrast in optical versus SEM review images and ability. Defect is a via etch defect on top of intermetal dielectric oxide. EDS on SEM tool showed Si/C. (a) Optical; (b) SEM.

ways to perform process monitoring on production wafers as they go through each step of processing, to detect any problems in-line and in real time. However, unpatterned wafers will continue to be used for wafer-handling tests, acceptance and qualification of new tools, and qualification of processes in tools after service or routine maintenance (i.e., unpattern wafers will continue to be used for preventative control).

Methodology

Multiprobe yield (also known as sort yield, nominal yield, or die yield) is arguably the best metric to distinguish between a semiconductor fab that is struggling to perform and maintain financial progress and one that is smoothly operating and thriving in the competitive marketplace. Established semiconductor technologies that have been in volume manufacturing for more than one year typically have probe yields that are defect-limited. Newer technologies are usually still developing and fine-tuning the process and equipment to work out marginalities in the process/design and are limited to lower yields by systematic issues, only some of which may be caught by in-line visual inspection. What is obvious to all in the industry is that no semiconductor manufacturer can hope to be successful, especially with newer technologies and smaller geometries, without adequately clean facilities and equipment.

Approaches to addressing low yields in the early 1980s relied almost solely on physical failure analysis of failed die at the end of the line. With the relentless advance in technology toward smaller geometries, larger die, and more processing steps, as well as an increasing demand to recover the greater than $1 billion cost of fabs quickly, such techniques are far too slow, expensive, and limited in scope. Extensive in-line monitoring of defects, either particulate contamination or process-induced defects, such as corrosion, is now a standard approach for yield enhancement in all newer fabs. See Fig. 7 for outline of all the various inspections performed. The key benefit of in-line defect monitoring is reduced cycle time of fixing problems with process and equipment compared with using probe data alone (see Fig. 4). A severe issue near the front end of the line (like isolation or gate) might not be caught at test for more than 30 days for an advanced process flow (4+ levels of metal). An in-line inspection plan in the right place might take three days for the problem to be identified and another day or so for the offending loop or equipment to be identified so that hopefully a fix can be implemented quickly afterward. For emerging or developing technologies or controlled experiments, in-line detection gives almost instant feedback on any visual integration problems. For baseline defect reduction, top defects on the yield loss Pareto can be identified by review and classification. Partitioning of the process loop and SEM/EDS characterization of the defects can quickly identify the root cause.

Because it is impractical and unnecessary to inspect every wafer of every lot at every inspection step, some sampling plan must be implemented in order to minimize the cost of inspection. However, this must be done in such a way as to minimize the likelihood of a crucial defect issue going undetected and unresolved for several days, in which time several hundred more wafers would be contaminated and suffer the yield loss associated with the problem. Usually only 2 to 3 wafers per lot are inspected, and the same wafers at every inspection, if possible, to allow calculating the number of defects added between inspection points (i.e., “adder” defects). The results are generalized to represent the condition of the entire lot if results are fairly consistent from wafer to wafer. Lot sampling varies from every lot to every second, third, or even fifth lot. As a rule, within-lot variation is less than lot-to-lot variation, so more value is obtained by inspecting more lots and fewer wafers per lot (in a capacity-limited scenario).

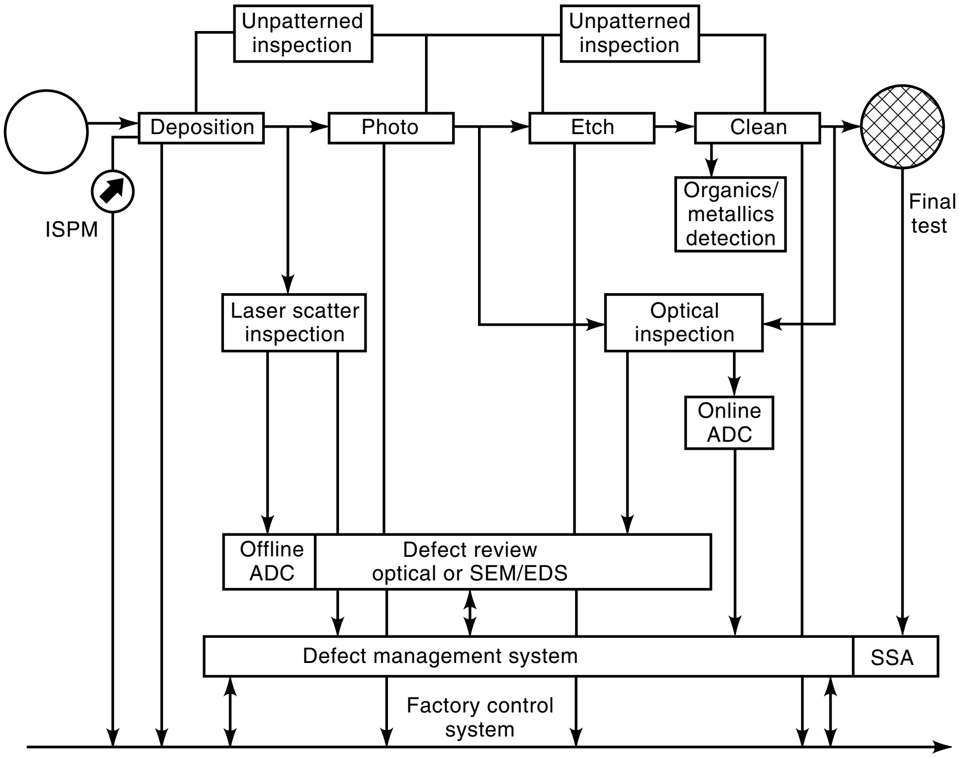

Figure 7. Inspection and data management flowchart for defect and contamination control. (ADC = Automatic Defect Classification; SSA = Spatial Signature Analysis; SEM = Scanning Electron Microscope; EDS = energy dispersive spectroscopy.)

Thought must also be given to where in the process flow to place the inspection steps. Laser-scatter-based patterned wafer inspection tools perform best after film deposition and are adept at detecting particles that rise above the surface of the film. Optical-based inspection tools are often the best choice for postpattern or postetch inspection steps because they can pick up planar defects such as blocked etch or residue between the structures. Intimate knowledge of the fab equipment, device and process flow is essential to choose the best plan to inspect and control defects for a particular situation.

Rigorous SPC control of defects on production wafers is essential. A stable baseline must be established, and any deviation upward from the baseline must be investigated. This adds cycle time to the material being investigated but is crucial in order to drive to the root cause of the defects. Figure 8 shows the desired response action to an out of control (OOC) condition (i.e., when a defect SPC chart alarms). If detailed analysis of the current out-of-control lot does not conclusively give the location of the defect source, the next material coming into the suspect process loop must be partitioned by inspecting at many nonstandard inspection steps in order to isolate the offending process/equipment. Such partitioning of process loops is essential for any baseline reduction effort. At any given inspection step, the defects could be originating from many different defect sources. In order to make steady improvements in the baseline defect levels (and so improve the yield), much effort and analysis must be expended to understand the pareto of defect types and their sources. After a critical or high-level defect can be attributed to a particular process and/or process equipment, teams of experts including process/equipment engineers, tool vendors, and yield enhancement engineers can be chartered to address the issues and implement fixes.

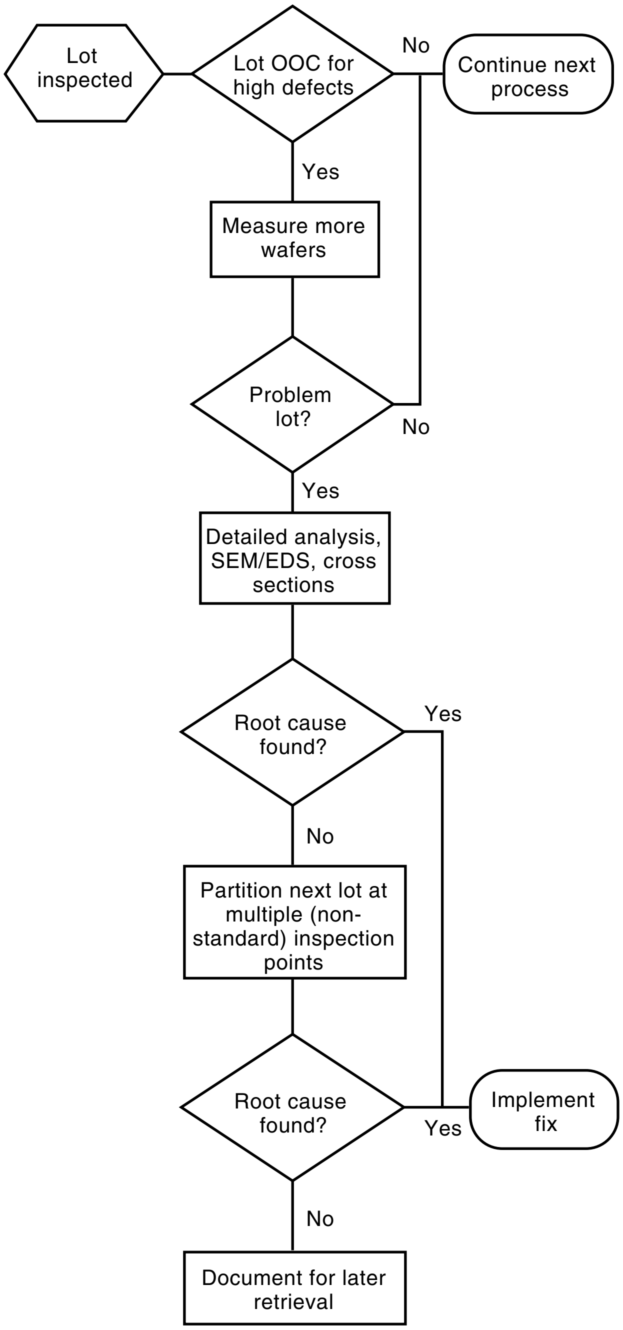

Figure 8. Defect or contamination SPC chart OOC response procedure.

Data Management and Analysis

The primary goal of in-line defect detection and review, especially in a manufacturing fab, is to collect reliable information about defects on the wafers, compile this information quickly and concisely, and use it to manage the (defect-limited) yield in the fab effectively. Information needed includes defect density, spatial layout, process level first detected, size, and classification type. The end goal is to identify which defects and tools/processes need appropriate attention to prevent an excursion from causing significant yield loss (SPC control), or to concentrate limited resources on the top defects in a pareto in order to maximize the impact of such efforts (baseline defect reduction). Increasingly it is crucial to have an integrated system to hold all this historical data for easy access, provide automatic data summary and report generation, track historical performance of inspection steps, and apply SPC methodologies to control the line. Newer analysis systems are emerging with the capability to be proactive and search for correlation and patterns without human intervention. The main idea of data management systems (DMSs) is to turn all the collected data from wafer processing into useful information for the process engineer. Figure 7 highlights the various sources of data that can be used.

All defect coordinate information should be fed to a central database. Defects can be clustered if they are spatially grouped. If clustering is not done, the groups of defects from mechanisms like scratches or corrosion can greatly outnumber the random defects across the wafer. If this is the case, the SPC chart for that inspection will show a very large spike in defect count, indicating an unstable line with an inordinate number of defect excursions. In reality, these large defect spikes may be only one or two clusters of large numbers of defects affecting only a few die and all originating from the same mechanism. The confidence of maintaining a stable manufacturing line is greatly increased with clusters of defects removed from random defect SPC control charts. A software algorithm performing spatial signature analysis (SSA) on defect counts can prove quite beneficial. SSA can be trained to recognize process signatures such as scratches, radial arms, repetitive defects from the mask, or nuisance process variations from the total defect count. This procedure then reports only the random defects that occurred at that step. If manual review and/or ADC are being used, the number of defects needing review or ADC is also greatly reduced using SSA.

Planning and a well-thought-out methodology can maximize data collection and correlation. In order to drive toward the source of defects in-line quickly, the same wafers should be inspected at all steps. After the data are fed to the central database, level-to-level defect overlay can be performed. Because the defect coordinates are also saved in the database, some position overlay tolerance (on the order of 50 to 300 μm, depending on the position accuracy of the inspection tools) can be applied, and maps from all previous inspections can be overlaid with the current inspection. Defects from previous levels that fall within the tolerance bounds of currently detected defects can be attributed to the previous levels, and so indicate the true added defects detected on the wafer since the last inspection point. Additionally, knowing which defects carry over to subsequent levels will also give some indication as to whether the defects are likely to cause electrical fails or not. In addition, the images themselves (both optical and SEM) are often saved electronically and linked to the individual defect positions on the wafer map. From the data management user's interface, a simple click on the marked defect can then bring up the image. A picture can truly be worth a thousand words because defect shape, color, morphology, and interaction with the surrounding circuit can give many clues as to the defect's origin.

Bit Map Matching. One very powerful use of this central defect management system is correlation of in-line defects with end-of-the-line electrical fail information. This is most useful in matching the coordinate position of defects with the bit/row/column fails in a memory structure (DRAM, embedded SRAM, flash memory, etc.). Memory fail testing, by its nature, gives the exact spatial address of the failing capacitors, often in patterns of failed rows or columns or clustered bits. By knowing the spatial positions of both physical and electrical defects, matching can be done using an overlay tolerance (again, depending on the particular system), and assuming that a physical defect that occurs very close to an electrical defect is probably the root cause of that electrical defect. This technique works best with high-yield and low-defect production lines with relatively few fails and therefore a relatively low chance of random matching of physical/electrical defects. After this has been done for many wafers, a Pareto can be developed to identify which defects (by inspection level, size, classification) are causing the highest number of electrical fails or have the highest kill ratio (probability of causing an electrical fail). By using bit mapping to correlate defects, killer defect properties are identified for future use in in-line defect control and optimization of the recipes on the defect detection tools. In other words, bit map correlation is used to increase power $(1 - \beta)$ for a control method that has a shorter response time than end-of-line testing.

WAFER POSITION TRACKING

As mentioned in the discussion of Control in Depth, identifying the source of yield loss from yield numbers alone is quite difficult. Besides bit mapping, another method for trying to determine the source of yield loss is based upon correlation of abnormal yield wafers with their processing position in each piece of equipment. This method of control is known as wafer position tracking (28). It is also known as "Wafer Sleuth," although Wafer Sleuth is a brand name copyrighted by SleuthWorks (29–31). Use of wafer tracking first gained visibility because of work done at SEMATECH, but it is now common in many fabs around the world.

In order to do the correlation, the following components and operational practices are required:

- Readable wafer identification scribes, either character or bar code
- Readers to read the wafer identifications for a lot
- Sorters to randomize the wafers in a lot periodically
- A database to store the positional order of each wafer at each reading and notes the routing and which equipment was used
- A database that stores yield data identified for each wafer
- An analysis package that uses the data from the database(s) and identifies abnormal wafers and determines their positional and equipment commonality; preferably the analysis occurs automatically

Scribe readers are required to ensure quality of data and speed of tracking. Randomization is required to achieve few correlations where each correlation identifies a possible rogue machine. Wafers do not change position frequently, other than to reverse order, as they proceed through their routing. Thus, a wafer is likely to be nth or $24 - n$th for its entire processing life. (Note that some tools obtain wafers in groups, such as 8, so that there is some randomization, but it is not great enough.) The randomization is critical to break this consistency and create a situation where a wafer can be in any position in the boat. In addition, because all the lots are randomized, the chance that all wafers with a particular yield loss have the same position in more than one piece of equipment is small. Thus, correlation between yield loss behavior and the processing position in a given piece of equipment is used to identify rogue equipment. The processing position also provides assistance in determining the source of the fault in the equipment. For example, if the first wafer in a furnace is suffering yield loss, then the technicians know to focus on that end of the furnace. If it is the third wafer in an implanter, then the rotation pattern of the equipment is suspect (some implanters rotate the wafers in groups of three).

DATA MINING AND DATA WAREHOUSING

The importance of using all sources of data to maximize ability to locate sources of yield loss and customer disruptions has been highlighted by several of the previous methods. However, traditionally, data from various sources are in different databases. For example, final yield at assembly/test may be in one database, design information may be in another database, in-line process data may be in another database, and defect data may be in yet another database. Thus, the first need is to get the data into one database. This combining is typically called data warehousing. Looking for the correlations in this massive amount of data is called data mining. Data warehousing and mining have been common in other industries, and these techniques are now being applied to the semiconductor industry. See the Reading List for references.

PARAMETRIC AND YIELD OUTLIER CONTROL

The use of parametric and yield data in SPC charts is common for product engineers to track the performance of their devices. However, a new control method that uses parametric and yield data is being driven by the customer. Outlier control, also known as maverick control, is a method for identifying wafers or lots whose performance is outside of the fab's normal distribution. Today many IC customers want consistent delivery of devices whose performance matches those that were used for their initial system qualification rather than only being compliant to specifications. There is also a correlation between outlier material with low yield and product with poor reliability in the customer's application. Many customers require that outlier material not be shipped to them or that expensive burn-in be used on outlier material. Therefore, the control of outlier material at the wafer fab level must be done to initiate corrective action rapidly and to reduce the cost of further testing and burn-in of deviant material. It is important to note that although outlier material may be within test specifications, it is deviant to the normal population of material.

The identification of outlier material may be determined by either outlier parametric values or by yield—the first being variable type data and the later being attribute data. For a true Gaussian distribution, either normal statistics or Tukey statistics could be used to define outlier controls. However, Tukey statistics develop more realistic limits because of insensitivity to the presence of outliers in the data set used to derive the control limits.

In the Tukey method, limits are determined by ordering the data from smallest to largest. The data are then divided into four equal parts or quartiles. The first quartile (Q_1) occurs at the 25% percentile, the point below which 25% of the data fall. The third quartile (Q_3) is the point below which 75% of the data fall. The interquartile range (IQR) is defined as IQR = $Q_3 - Q_1$. Two sets of limits are then defined.

$$\text{Inner Limits: } Q_1 - 1.5 \times \text{IQR} \quad \text{and} \quad Q_3 + 1.5 \times \text{IQR} \qquad (4)$$

$$\text{Outer Limits: } Q_1 - 3.0 \times \text{IQR} \quad \text{and} \quad Q_3 + 1.5 \times \text{IQR} \qquad (5)$$

The inner limits are defined such that any data beyond these limits may be considered as possible or near outliers from the central distribution. The outer limits are defined such that any data beyond these limits may be considered to be serious or far outliers.

As stated earlier, the determination of the Tukey limits is insensitive to the presence of outliers in the data set used to calculate the limits. This is a result of the fact that the limits are calculated using quartiles. Because outliers usually appear beyond the first and third quartiles, their presence does not significantly change the values of Q_1 and Q_3. Thus the Tukey limits remain the same. On the other hand, the standard deviation of a sample is very sensitive to the presence of outliers. Their presence causes the estimate of σ to become inflated. As a result, limits based on standard deviations become too large and will not detect the outliers.

Unfortunately, the Tukey method for determining outlier limits for yield data does not work well directly because of the distribution of yield data. Because yield data are bounded between 0% and 100%, using Tukey statistics directly could possibly result with limits defined outside of the 0% to 100% boundaries. As a result, no serious outliers would be identified. The Tukey method works on normally distributed (unbounded) data. The yield data can be transformed so that Tukey limits can then be applied to the transformed data. The logit transformation is recommended in these situations when using proportion data, such as yield data. The logit transformation is defined to be

$$\text{logit}(p) = \log[p/(1-p)] \qquad (6)$$

For yield data, which is between 0% and 100%, logit (yield) is given as

$$\text{logit(yield)} = \log[\text{yield}/(100 - \text{yield})] \qquad (7)$$

Logit(yield) will then range from negative infinity to positive infinity. Tukey limits can be determined based on the quartiles of the logit distribution. The logit limits can then be transformed back into yield units using the inverse transformation:

$$p = 1/[1 + \exp(-\text{logit})] \qquad (8)$$

or for yield data:

$$\text{yield limit} = 100/[1 + \exp(-\text{logit})] \qquad (9)$$

WAFER LEVEL RELIABILITY CONTROL

Wafer level reliability (WLR) is an important method to monitor the reliability performance of devices, materials, and their interactions prior to packaging (32). Typically, product qualification occurs at the package level prior to full-scale production to verify the product's robustness for operational life, resistance to corrosion, and tolerance to mechanical stress. These tests are conducted under dynamic operation at elevated temperatures and voltages, in high humidity and temperature, and under conditions of temperature cycling. Although such stress tests are effective in projecting failure rates for similarly processed units, the entire qualification process represents only a snapshot in time (i.e., the process could deviate in the future from that used for the qualification lots). Even though an intentional "major" process change is strictly forbidden without requalification, unintentional process changes may occur, or a series of "minor" process changes may effectively add, unknowingly, to a cumulative "major" process change. At the wafer level, WLR detects unintentional changes and process drifts that change the intrinsic reliability of the device from its initial qualification. Because WLR is done in the fab, failures detected can be quickly addressed as opposed to discovering a failure at packaging.

WLR testing is a series of accelerated tests, done at the wafer level, which can be performed rapidly to assess the intrinsic reliability of the IC technology/process. Because the acceleration is normally achieved through the use of elevated voltage, current, and/or temperature, the essential elements of a WLR probe station should include a high voltage source, a high current source, and a rampable, hot temperature stage. The hot probe chuck should permit a temperature elevation of 300°C for mobile-ion testing; the high voltage unit should permit up to 100 V for interlevel dielectric leakage measurements; the high current module should permit up to 200 mA for electromigration testing of leads, contacts, and vias. Several key reliability parameters of an IC technology that can be accelerated in order to obtain a real-time monitoring of the reliability robustness at the wafer level are listed in Table 7. Table 7 also provides corresponding issues with respect to these parameters.

Because WLR testing is done under highly accelerated conditions (stress times must be kept short so that sufficient statistics can be gathered), extrapolation of such greatly accelerated data, to precise failure rate prediction for the field, requires many time-decades of extrapolation. For this reason, it is better to use WLR for "reliability fingerprinting" of the qualification lots rather than absolute failure rate prediction. Reliability fingerprinting simply means that the individual components of reliability (metallization, contacts, vias, gate oxide, transistors, etc.) are stressed for the qualification lots, and the shifts (in metal resistance, contact resistance, via resistance, gate oxide breakdown strength, transistor V_t, etc.) are carefully documented. This documented shift becomes the reliability fingerprint that is used as a "benchmark" to detect deviations of the process in the future and to support the continuous product improvement efforts. By using this reliability fingerprinting methodology, the reliability of the process can be continuously controlled.

If the WLR data become "out of control," then efforts must be taken to contain the affected material, to determine the root cause, and to implement corrective actions. To determine the affected lots, all the lots at risk must be sampled for WLR testing. The lots at risk are those lots processed since the last normal WLR test. In a parallel effort, a root cause analysis of all the factors that could contribute to the failure signature of the WLR test is performed. For example, there are many factors that can affect gate oxide reliability. These include intrinsic factors such as starting wafer quality, preoxidation cleanups used for silicon surface preparation, furnace growth conditions, poly deposition, and annealing. There are also many extrinsic factors such as particles, implantation damage, and wafer charging during processing that can have an impact on the reliability of the gate-oxide after it is grown and fabricated into devices. To illustrate the processing variables that can have an impact on the gate oxide reliability, an Ishikawa (or fishbone) diagram is useful and is shown in Fig. 9. Each of the bones on this diagram can, of course, be further expanded and detailed. It soon becomes obvious that tracing a gate oxide issue back to its "root cause" is a complex and time-consuming task. Even though WLR control provides high-quality information on the interaction of the processing variables, controlling the variation of variables and conditions at the lowest level possible which may impact reliability per-

Table 7. Definition of Key Reliability Parameters Tested Using WLR with Associated Issues

Junctions	For a CMOS technology, both $n+/p$ and $p+/n$ junctions must show low leakage, good kinetics, low defect density, and good stability under voltage, current, and temperature stressing.
Gate oxide	The gate oxide for the MOSFET must have low leakage at use electric fields, high breakdown electric fields, high charge-to-breakdown values, and good V_T stability under gate stressing at high voltage for both low and high temperatures.
Mobile-ions	Device isolation depends on the thick field or shallow trench oxide film being relatively free of mobile-ions so as to prevent surface inversion during high-temperature and high-voltage stressing.
Channel hot-carriers	N-channel short channel devices must be stressed under the conditions of maximum substrate current and the transistor parameters (e.g., V_T, g_m, I_{DS}) monitored for shifts. P-channel short channel devices must be stressed under the conditions of maximum gate current and the changes in off-state leakage monitored.
Metal integrity	All metal levels should be tested for electromigration robustness under the conditions of high current density and high temperature. Prior to electromigration testing, some of the electromigration samples should be baked at ~175°C/1000 h so as to induce any stress migration effects. Both NIST-type and via-fed electromigration test structures are recommended.
Contact and via integrity	Contacts to diffusions and metal-to-metal contacts (vias) should be tested at high-current and high-temperature conditions. The electromigration performance should be determined in both current-flow directions.
Interlevel dielectric integrity	Both intralevel and interlevel dielectric leakage should be measured at high voltage and high temperature. Etching residues or side-hillock formations can present leakage or breakdown issues for the intralevel dielectric. Interlevel dielectric issues can develop because of top-hillocks on the metallization or poor planarization.
Passivation integrity	Passivation over the final metal level should be planarized so as to minimize the thermomechanical interaction with the plastic package and must be pin-hole free to prevent corrosion. The passivation pin-hole density determination can be accelerated by an exposure to a simple metal etch.
Corrosive residues	No corrosive residues should be left on the wafer after metal etching and photoresist removal. Also, no corrosive residues should be left on the bonding pads after back-grind cleanup. A simple water-box storage test (24 h/100% RH) can be used to accelerate the detection of corrosive residues.
ESD/latchup	The ESD robustness should be assessed by measuring the high current–voltage characteristics of the n-channel output transistor [e.g., the trigger and snapback voltages and I_{t_2} (the second breakdown current)]. The latchup robustness can be assessed by measuring the trigger current and holding voltage of a four terminal *pnpn* device.
SER robustness	The soft-error-rate (SER) robustness of the process and/or design can be assessed by measuring single-event upsets using an accelerated alpha-particle source such as thorium or americium.

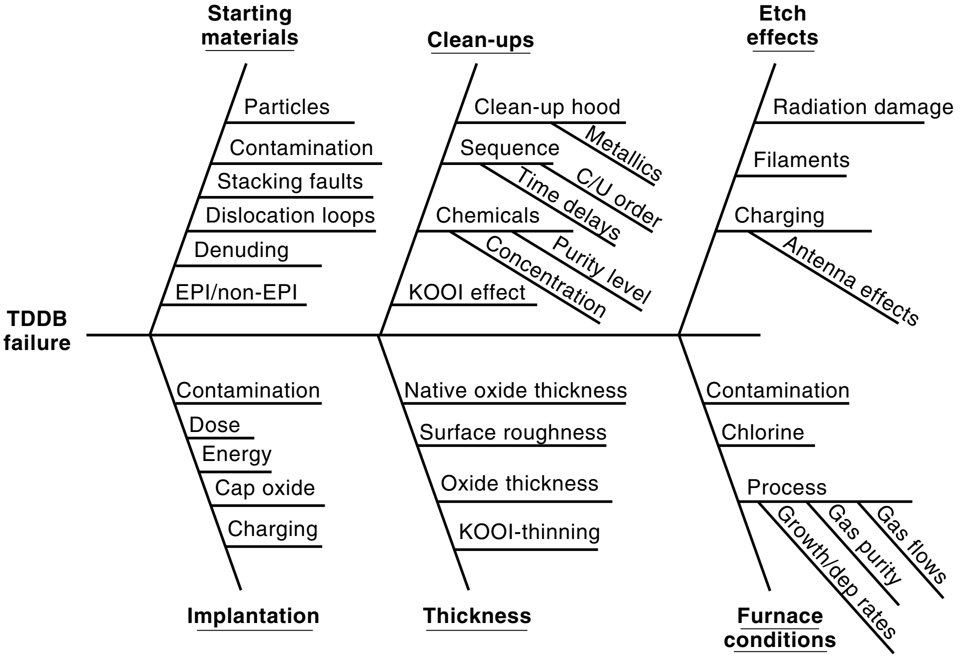

Figure 9. Fishbone diagram showing the areas of the process that can impact gate oxide reliability.

formance is extremely important to guarantee disruption-free delivery of product. For example, for gate oxide integrity, some of these lower-level control methods are careful preventive maintenance of furnaces, sensitive equipment monitoring with real-time traces, and the use of high-purity chemicals for surface preparation and oxide growth. This enforces the concept of control in depth.

MULTIVARIATE SPC, ESPECIALLY FOR EQUIPMENT SIGNAL MONITORING

Typically only a single measurement is taken in-line, such as thickness. If more than a single thickness measurement is taken on a wafer or across a lot, then multivariate statistical process control (MSPC) would be applicable. Multiple measurements across a wafer and within a lot are gathered at final probe. These data are obviously highly correlated (33). Thus, multivariate SPC would be of significant value. However, currently, it is rare to see MSPC applied to final probe, although application is expected to increase as customers demand increased quality, and the business environment requires less unnecessary scrap and reduced burn-in. There have been discussions about replacing the univariate outlier Tukey method with MSPC, but that has not occurred yet. The most common application of MSPC is to equipment and sensor signals. Using the semiconductor equipment communication standard port, it is easy to collect 50 different variables (signals, traces) once per second on many machines. Because most of the focus of MSPC is for equipment signal and sensor applications, the discussion will focus mainly on the mathematics necessary for such applications.

Methods that examine only a single variable are called univariate. Use of multiple univariate SPC charts for the case of multiple variables has been cited as being too cumbersome for a human to handle. However, with the advent of computers, such an issue is irrelevant because computer technology can be used to set up many charts and perform all calculations. However, using many univariate charts does suffer from two major problems:

1. Unacceptably high overall error rate for false positives for uncorrelated variables
2. Unexpected false positive and false negative rates when the variables are correlated

In this section, we will describe the different methods for performing SPC on multiple variables that solve these two problems. We will first address the case for uncorrelated variables and then examine the case of correlated variables. Many of the issues that arise in practice will be discussed and their common solutions given. Note that many of these issues also arise in univariate SPC charts, and some of the solutions have been extended from the univariate case. The focus in this section is only on the comparison/expectation part of the control model of Fig. 2. The rest of the components that were discussed in the section on univariate SPC, such as corrective procedure, are still required. However, no special changes are required for multivariate SPC except the additional step of isolating which few variables, of the many variables charted, are involved in the fault. This isolation is necessary because in univariate SPC, the faulty variable is intrinsically identified.

Controlling Overall False Positive Error Rate

Although equipment signals are usually correlated, it is possible to have a reduced set of equipment signals that are uncorrelated. Such an uncorrelated set of variables may occur because a fab is trying to reduce the amount of data it collects, and so they eliminate any redundant variables (i.e., variables that are correlated with other variables). The correlated variables are assumed to provide no additional information about the process, but, as will be discussed later, monitoring the correlation provides very sensitive and robust fault detection. However, business situations may require a few variables, and the focus will become the use of those variables that in total contain the most information. Even if the variables are uncorrelated, traditional univariate SPC chart set-up procedures can lead to increased false positives.

There are two different approaches to handle the issue of error rates. One approach is based on changing the control limits of the univariate charts, and the other approach is to use multivariate methods. We will discuss the former approach first. The best known of the adjustment methods is Bonferroni inequalities. The method is easy to employ. Let α be the desired Type I error (i.e., rate of false positives). For example, the traditional Shewhart univariate chart set up with limits set at 3σ has an α of 0.27%. With Bonferroni limits, for p tests on p variables, the limits are set at Type I values of α/p. Thus, the overall Type I rate is kept at $p^*\alpha/p = \alpha$. To demonstrate the problem of increased Type I error, suppose a typical fab with 40,000/month wafer starts and with 30 day fab cycle time has a process on which Shewhart charts with 3σ limits are used. If each run is 24 wafers and only one variable per run is monitored, a false positive will occur approximately once per week for that process. In other words, the SPC chart will indicate a fault has occurred when no fault has occurred. Even though time is wasted investigating the alarm, such false positive rates are acceptable in order to ensure that a real fault will be detected. However, if 10 variables are monitored each with Shewhart charts with 3σ limits, then approximately 1.5 false alarms occur *per day!* This rate is unacceptable. Another method is Roy and Bose intervals, which some prefer because Bonferroni may give a slightly shorter average run length (34). A third method is to use a technique for correlated variables that naturally handle the overall Type I error. Such methods will be discussed next.

Hotelling's T^2: The Traditional MPSC Chart

The preceding discussion on false positive rates assumed that the variables are independent. When the variables are correlated, the false positive and negative rates for using univariate charts can be quite different than expected (35–38). For example, a change in correlation may go undetected. This concept is shown in Fig. 10. Two variables (y_1 and y_2) are plotted against each other. Upper control limits (UCLs) and lower control limits (LCLs) for each variable are shown as if univariate charts were set up. The dots represent typical variation. As can be seen, the points all lie within a well-defined ellipsoid. In other words, y_1 and y_2 are correlated. X represents an unexpected point in that it violates the correlation structure of y_1 and y_2. However, the univariate charts would not detect X as abnormal. Thus, a method is needed which will detect a change in the system that considers the correlation structure of the system. The most common and well-known test statistic for MSPC is the Hotelling's T^2 (35–38). This statistic describes an ellipsoid in p-dimensional space that has a probability $1 - \alpha$ of containing all the data sample of p variables. This ellipsoid is shown in Fig. 10. Solid ellipsoids satisfy the following equation with probability $1 - \alpha$ if both the population covariance matrix and mean vector are known (35,36):

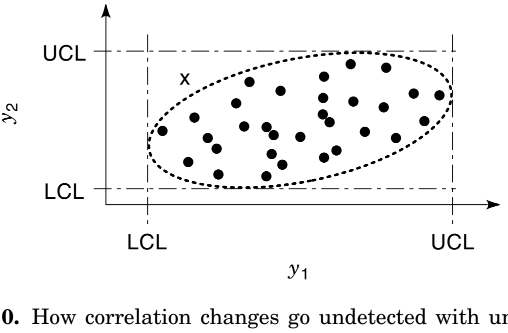

Figure 10. How correlation changes go undetected with univariate charts. (X is a fault.)

$$n(\overline{\boldsymbol{X}} - \mu_0)^T \Sigma^{-1}(\overline{\boldsymbol{X}} - \mu_0) \leq \chi_2^{\mathrm{p}}(\alpha) \quad (10)$$

where

$\overline{\boldsymbol{X}}$ is vector ($p \times 1$) of sampled means of each of the p variables
μ_0 is vector ($p \times 1$) of population means of each of the p variables
Σ is population variance–covariance matrix
p is number of variables
n is sample size used to calculate $\overline{\boldsymbol{X}}$
χ is chi-squared statistic

This equation can be used for MSPC in that it is based upon the probability that the sample mean will lie within a certain range. In other words, assume a hypothesis (H_0) of IIDMN(μ_0, Σ) where IIDMN is identically independently distributed multivariate normal with a multivariate mean of μ_0 and covariance Σ:

$$\begin{aligned}&\text{Null hypothesis } H_0\colon \mu = \mu_0\\ &\text{Alternative hypothesis } H_1\colon \mu \neq \mu_0\end{aligned} \quad (11)$$

To test this hypothesis, a test statistic and limit is needed:

$$\begin{aligned}&\text{If test statistic} \leq \text{test limit, then} H_0 \text{ is accepted}\\ &\quad \text{i.e., the means are not statistically different)}\\ &\text{If test statistic} > \text{test limit, then} H_0 \text{ is rejected}\\ &\quad \text{(i.e., the means are statistically different)}\end{aligned} \quad (12)$$

Using Eqs. (10) and (12), but substituting estimates for μ_0 and Σ, Alt has shown (34,35,37):

$$\begin{aligned}\text{Test statistic} &= T_\alpha^2 = n(\overline{\boldsymbol{X}} - \overline{\overline{\boldsymbol{X}}})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{X}} - \overline{\overline{\boldsymbol{X}}})\\ \text{Test limit} &= \frac{p(m+1)(n-1)}{(mn-m-p+1)} F_{\alpha,p,mn-m-p+1}\end{aligned} \quad (13)$$

where

$\boldsymbol{S}$ is estimated variance–covariance matrix by pooling m samples of size $n = \sum_{i=1}^m \boldsymbol{S}_i$

$$\boldsymbol{S}_i = \frac{1}{n-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)^{\mathrm{T}}(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_i)$$

$F_{\alpha,p,mn-m-p+1}$ is Fisher's F statistic with degrees of freedom p, $mn - m - p + 1$
$T_{\overline{\alpha}}^2$ is Hotelling's T^2 (39)
$\overline{\overline{\boldsymbol{X}}}$ is estimated mean of each of the p variables with sample size n

m is number of sample sets of size n used for estimation
n is sample size

Thus, if $T_\alpha^2 >$ test statistic of Eq. (13), then the hypothesis of Eq. (11) is rejected, and the system is assumed to be no longer the same. Equation (12) is one sided (i.e., there is only an upper control limit) because T^2 can only be positive. However, others have used nonzero lower control limits (40).

Hotelling's T^2 is the multivariate analog of the univariate t^2 statistic. Consequently, it has also been called the multivariate Shewhart chart, although Shewhart personally had no association with its development or use. It has several useful properties:

- It has a quadratic form.
- T^2 is unaffected by changes of units or shifts of origins of the response variates, but it is also invariant under all affine transformations ($Wx + b$) of the observations and hypothesis (affine equivariant) (34,41). Thus, the test is unaffected by scaling of individual measurements in x (34).
- It is the optimal affine invariant test statistic for a shift in the mean vector of the single observation vector $X(n = 1)$ or for a shift in the mean in all n observations of group size n (34).

Test for Individuals. The preceding equations are for samples of size n. In semiconductor manufacturing, it is rare to take from a batch more than one sample that meets the necessary requirement that the within-sample and sample-to-sample expected variation is the same. Thus, a statistic is needed for single sample sizes, also known as an individuals test. If a large sample size is taken to estimate parameters, then the following equation holds (37):

$$\begin{aligned}\text{Test statistic} &= T_\alpha^2 = (\boldsymbol{X} - \overline{\boldsymbol{X}})^T \boldsymbol{S}^{-1} (\boldsymbol{X} - \overline{\boldsymbol{X}}) \\ \text{Test limit} &= \frac{p(n+1)(n-1)}{n(n-p)} F_{\alpha,p,n-p}\end{aligned} \tag{14}$$

where

n is sample size used to calculate $\boldsymbol{S}$ and $\overline{\boldsymbol{X}}$
$\boldsymbol{S}$ is estimated covariance matrix from sample size n
$F_{\alpha,p,n-p}$ is Fisher's F statistic with degrees of freedom p, $n - p$

Tests for Dispersion

In the univariate case, the chart used in combination with the individuals is a moving range chart. Unfortunately, the multivariate analog of moving range chart is intractable (35,37). Thus, no equivalent exists for a moving range chart to be used in the multivariate individuals case. However, Smith has proposed an analog to the range chart when discussing the calculation of T^2 for groups (38), as have Prins and Mader (42). Other types of charts to monitor dispersion (variance) are reviewed by Alt and Bedewi (43). Healy showed the CUSUM of T^2 (COT^2) is an appropriate test statistic for inflation of the covariance matrix (44) (i.e., to test for a scalar multiplication of the covariance matrix).

Issues with T^2 in Practice

Even though T^2 is the most commonly seen and the oldest multivariate technique, it suffers from several problems.

- Even though Eqs. (13) and (14) give theoretical limits that produce a Type I error of α, these limits are found in practice to yield a much greater Type I error. Crosier gives figures for out-of control average run lengths (ARLs) based upon size of shift, number of variables, and in-control ARL (45). In practice, simulations, boot-strapping, and actual data are used to set the control limits. Tracy et al. discusses the issue of limits and provides alternative equations (40).
- The values used to calculate $\boldsymbol{S}$ and $\overline{\overline{\boldsymbol{X}}}$ must be "good" data (i.e., data from when the system is in control). A large data set (>100 lots) is required to calculate variances, preferably a data set with greater than a 1000 lots would be used. Thus, manually identifying bad data points is impossible. Use of automatic outlier rejection (e.g., testing the data, removing data outside the control limits, and recalculating the tests) is easy with today's computers. However, the resulting test limit may be overly sensitive because extreme, but expected, data points were removed from the data set by this method.
- Although Eq. (11) assumes that the variance is constant, dispersion (variance) and mean shifts are confounded in T^2 (40). Several people have used T^2 failure to signify a change in the variance (44,46).
- Sample size needed to detect shifts in the process means does not always decrease as the magnitude of the shifts increase (35). For a relatively large positive correlation, the needed sample size increases with increasing positive shifts.
- For the bivariate case, when the two variables are positively correlated, the probability of detecting a shift, known as power, is not a monotonically decreasing function of the standard deviation, as it is in the univariate case. Thus, a smaller noise level does not necessarily translate to a higher probability of detecting shifts.
- Single-test optimality does not imply optimality in repeated use, which is the case for univariate charts (34).
- If the variables are highly correlated, then $\boldsymbol{S}$ is singular (i.e., it is not invertable and therefore $\boldsymbol{S}^{-1}$ does not exist). In such a case, data reduction methods, such as principal component analysis, must be employed. Such methods will be discussed in a later section.
- Even though single sample sizes are common, if more than one sample can be taken, the issue of sample size should be carefully investigated. Aparisi does such a study and shows the answer depends on the particular situation (47).

Note that T^2 may still be used even if the data are not correlated. It provides an easy way to overcome the overall Type I error problem instead of using Bonferroni limits. However, sometimes the result is decreased sensitivity to a fault that appears in only one variable. It is very difficult to achieve simultaneously sensitivity for all variables and yet not have an unacceptable Type I rate.

Applications. As mentioned in the introduction to this section, the most common application of MSPC is to equipment and sensor signals. Using the SECS port, it is easy to collect 50 different variables (signals, traces) once per second on many machines. Because of equipment aging and chamber build-up, these signals change over time (i.e., they are autocorrelated run to run). Within a run, one would expect the signals to be autocorrelated because of within-process dynamics and the result of real-time controllers. Within a lot has a particular autocorrelation because of the first wafer effect (48,49) which is associated with chamber warm-up and degassing. Further explanation of the autocorrelation and variation time scales can be found in Ref. 50.

When discussing Eqs. (10) and (11), the assumptions of normality and IID were noted. IID also assumes independence (i.e., that each data point is not autocorrelated with the next one). Autocorrelation is shown to have an impact on the Type I, Type II errors (51–62). Another assumption is constant variance over the entire space, also known as homoscedasticity. Nonconstant variance is known as heteroscedasticity. The correlation structure is also assumed to be constant. While a changing correlation structure is uncommon, nonnormality, autocorrelation, and heteroscedasticity are encountered frequently. Thus, a method for "removing" the nonnormality, autocorrelation, and heteroscedasticity are needed for Eqs. (10)–(14) to be valid.

Models and Transformations for Application of MSPC to Equipment Signals

One way to "remove" the nonnormality, autocorrelation, and heteroscedasticity is to create new variables. These IIDN variables are the residuals of a model that predicts the autocorrelation for a transformed variable. The transformation accounts for the nonnormality and heteroscedasity, whereas the model accounts for the autocorrelation. This concept is shown in Fig. 11. For model residuals, the $\overline{\boldsymbol{X}}$ in Eq. (14) has the value of 0 because the model is expected on average to predict the output. Another way to view the use of a model is that it predicts the value of $\overline{\boldsymbol{X}}$ in Eq. (14). In other words, models are used to adapt the null hypothesis H_0 of Eq. (11) by adapting μ_0 to match the expected changes. Thus, faults are changes that occur faster or larger than expected. In summary, autocorrelation models are implemented in one of two ways:

1. μ_0 in Eq. (11) is approximated by $\overline{\boldsymbol{X}}$ in the equation $D = 0$; $\boldsymbol{X}$ in Eq. (14) = Residual = Measured (transformed) Value − Predicted Value.
2. μ_0 in Eq. (11) is approximated by $\overline{\boldsymbol{X}}$ in the equation D = Model Prediction; $\boldsymbol{X}$ in Eq. (14) = Measured (transformed) Value.

The equivalency between the two methods can be seen by substituting either implementation into Eq. (14), which yields

$$(\boldsymbol{X} - \overline{\boldsymbol{X}}) = \text{Measured (transformed) Value} - \text{Predicted Value}) \tag{15}$$

The Correlation matrix (**S**) is the same in both cases as well, using Eq. (15) for its calculation [see Eq. (13)].

By accounting for the autocorrelation, heteroscedasticity, and nonnormality, increased sensitivity (power, reduced Type II error) becomes possible while simultaneously reducing the Type I error (α). Note that even though new variables are used for analysis to detect faults, the system itself is not changed. To improve the signal-to-noise ratio even more, feedforward variables may also be used in the model, such as to account for the impact of wafer state upon sensor signals, such as the optical emission intensity decreasing with increasing percent open area during etch. Different devices have different percentages of open area for the same step in the flow (routing), and the same device may have different percentages of open area for different steps in its flow. Thus, the percentages of open area can be used as a feedforward variable to predict the change in intensity resulting from the

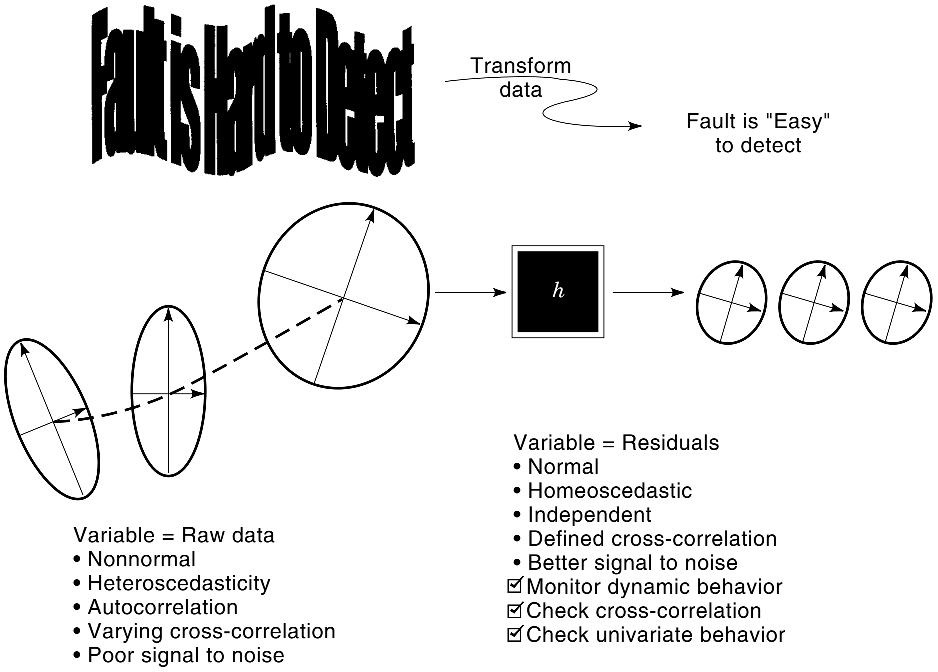

Figure 11. Purpose of process state model.

changes in the open area and then data from different devices can be analyzed together.

Logarithm and square root are the transformations most commonly used to create new variables that are normally distributed and homoscedastic (63). Time series models are the most common model form for predicting autocorrelation (64–70). The most common time series model representation is known as an ARIMA (p, d, q) model where p is the order of the autoregressive part, d is the integration order, and q is the order of the moving average part. The most well-known ARIMA order is the (0, 1, 1) order. The IMA model (it has no autoregressive part) is equivalent to a first-order digital filter. It is also equivalent to the EWMA metric used in an exponentially weighted moving average (EWMA) chart, a common univariate SPC chart. An ARIMA (p, d, q) of time series y_t is represented as

$$w_t = -\sum_{k=1}^{p}\phi_k w_{t-k} + \sum_{i=0}^{q}\theta_i a_{t-1} \tag{16}$$

where

$\theta_0 = 1$
a_t = forecasting error = $w_t - \hat{w}_t = N(0, \sigma)$
$\hat{w}_t$ = prediction of w at time t
w_t = Differenced data = $\nabla^d y_t$
y_t = Variable being modeled with time series (may be a transformation of raw data)
∇^d = dth order of differencing operator
$\nabla^1 y_t = y_t - y_{t-1}$
$\nabla^2 y_t = \nabla^1 y_t - \nabla^1 y_{t-1} = y_t - y_{t-1} - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$

Rearranging Eq. (16) by expanding the right term for $i = 0$ yields an equation to solve for $\hat{w}_t$:

$$\hat{w}_t = -\sum_{k=1}^{p}\phi_k w_{t-k} + \sum_{i=1}^{q}\theta_i a_{t-i} \tag{17}$$

Thus, Eq. (17) can be used to predict the value for w for the next sampling period. The preceding equations will work best if the y values are homoscedastic and normally distributed; consequently, transformations of the variables may be used for variable y instead of the raw data itself. The prediction residuals of a time series model of the transformed variables should produce IID Normal homoscedastistic variables (a_t). Consequently, a_t becomes the variable to be monitored by a MSPC chart as shown in Fig. 11.

The main challenge with respect to use of time series is that data across a SECS port is not at a constant sampling rate. Variations of plus or minus 20% of the sampling rate are not uncommon. However, time-series models assume constant sampling rates. Thus, techniques may need to be used to create a model that works on nonconstant sampling (71,72).

Real-Time SPC

Trying to use a single autocorrelation model for each 1-second sample across all wafers and lots has been shown not to work (64–69). Thus, Spanos et al. have decomposed the problem into three models representing the three dominant time scales over which the variation occurs: lot to lot, within lot, and within run. The lot average is used as the lot data, the wafer average is the within lot data, and the within run is a single sample or a group of samples. Transformations are used as needed on any of the signals. A separate T^2 is used for each of the three time scales. The three T^2 values are plotted in a single plot. The use of time-series models for real-time signals was termed real-time SPC (64–69). Note that even though some authors call the longer time scale lot to lot, others call it within a maintenance cycle (50). This is because the aging really occurs across the entire maintenance cycle. There is also a lot-to-lot effect generally caused by the incoming material (i.e., due to the lot itself).

Trace Analysis Using Dynamic Time Warping or Step Number to Generate Metrics

The cause of the most significant variation that occurs during the processing of a single wafer is generally caused by switching chemistries, ramping of power, or switching between films. Thus, these changes denote significant regions. Many, but not all, of these regions correspond to steps within a recipe. Thus, metrics could be generated for each signal during a particular step to handle the within-wafer autocorrelation. Another way is to generate metrics for these significant regions found by decomposing the signal using dynamic time warping (73,74). Such metrics can include average, standard deviation, coefficients from a curve fit through the data, the maximum, or the minimum. These metrics can be used together in a single T^2 with the mean predicted from a run-to-run autocorrelated model. Thus, the autocorrelation within a wafer is handle by treating it as cross-correlation. However, autocorrelation between wafers must still be treated. Instead of a formal creation of time-series models, a simple first-order filter (i.e., an EWMA) can be used with the filter factor picked using heuristics. In other words, the mean is adapted using an EWMA to account for wafer-to-wafer autocorrelation. The biggest issue found in a 7-month study of MSPC (75,76) using the preceding techniques was that the biggest change in the system occurred whenever maintenance was performed. However, this variation is not a fault. Thus, a method was needed to adapt the system to changes caused by maintenance. The EWMA adaptation of the mean was found to be almost adequate after maintenance. In other words, the correlation structure only changed slightly. However, the slight change required an exponentially weighted moving covariance (EWMC) to account for maintenance-to-maintenance changes and within maintenance aging. In addition, a large number of variables were being analyzed. This study also examined data reduction methods and found them to have fewer problems.

Data Reduction Methods, Such as Principal Components Analysis

To illustrate the usage of principal component analysis (PCA) for MSPC, a two-dimensional example will be given. However, in practice, it is the reduction of several hundred dimensions to a couple of dimensions where PCA finds its strengths. Figure 12 is similar to Fig. 10, but now the height of the ellipse has been shrunk. The data now fall in approximately one dimension defined by a vector p_1. Good data would be expected to lie along dimension p_1 within the UCL and LCL drawn on vector p_1; faulty data are expected to lie along

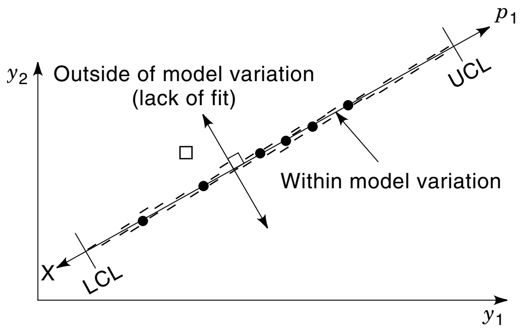

Figure 12. Dimensionality reduction—use of PCA for MSPC. p_1 is the principal component describing direction of expected (typical) variation; — is good data; X is a fault within the model; □ is a fault outside the model.

the dimension perpendicular to p_1 (outside the model) or to lie along p_1 (within the model) but be outside the control limits. Thus, the original data in dimensions (y_1 and y_2) can be translated into data in one dimension (p_1). p_1 is the eigenvector of **Y** and translating the original data into the p_1 space produces scores (t_1) as "data." Thus, the scores are expected to lie within the control limits similar to a univariate chart. If more than one eigenvector is required to describe the reduced dimension, each eigenvector will be perpendicular to all the others. Thus, univariate charts are still valid because PCA not only reduces the dimensionality of the data but also translates it into uncorrelated variables 5. However, note that the scores many times will be highly autocorrelated. Thus, a method to deal with the autocorrelation will be necessary. In practice, a T^2 chart is used in order to address the issue of inflated overall Type I error. To check for variance not captured by the PCA model, a $\boldsymbol{Q}$ statistic is used. $\boldsymbol{Q}$ is the sum of the squares of the residuals for each original variable y_i. Note that if the PCA model uses ALL eigenvectors (i.e., the full model is used and no data reduction occurs) then the original T^2 MSPC chart results, and there is no $\boldsymbol{Q}$ chart. One benefit of PCA is that it reduces the directional dependency of fault sensitivity compared to the full model (77). For more details on the mathematics and additional tests, such as on variance, see Ref. 78.

A 7-month study was performed to compare the full model versus a reduced model (PCA) (75,76). For both PCA and the full model, autocorrelation was present so that the mean needed to be adapted using a EWMA [i.e., a (0, 1, 1) time series]. Two scenarios will occur if a model for autocorrelation is *not* used:

- A data set that covers considerable aging and maintenances is used to develop the model (full or reduced) resulting in almost no sensitivity to real faults because the faults are much smaller than the normal aging or maintenance-induced changes
- A data set that covers very little aging and no maintenances is used to develop the model (full or reduced), which results in almost constant false positives caused by normal aging or maintenance-induced changes

Neither scenario is acceptable. By adapting the mean, normal variations are tracked, and faults then are detectable, as shown in Fig. 11.

In the study it was found that the covariance needed to be adapted for the full-model case, as described in the section on dynamic time warping. However, even though improvements were seen in the PCA case if the covariance was adapted, the improvements were so small as to not justify the effort. The overall result was that PCA provided more robustness (decreased false positives) and more sensitivity (decreased false negatives) than the full model case. This study, partially funded by SEMATECH, has led to increased emphasis on the use of PCA in the industry.

Data Reconstruction Methods with Regression Adjusted Variables

Another methodology for multivariate monitoring is to reconstruct (i.e., predict, the value for each variable from a model based upon all other variables). These predicted variables have also been called regression adjusted variables (34). A quick review will be given of the various modeling techniques used for the predictions.

Projection to latent structures (PLS), also known as partial least squares, is a technique for estimating the model

$$y = \boldsymbol{BX} \tag{18}$$

when the data are correlated. PLS attempts to maximize covariance by decomposing the **X** and **Y** matrices into vectors that are highly correlated. Thus, it is related to principal component analysis. PLS can be used for monitoring by creating a model for every variable as a function of all other variables:

$$\hat{x}_i = f_i(x_{\text{all}\,k \neq i}) \tag{19}$$

producing p PLS models, one for each of the p variables. The set of $\hat{x}_i$ are also called regression-adjusted variables (34,78). The residuals, $x_i - \hat{x}_i$, can be monitored in much the same way that the residuals in PCA are monitored. Again, the residuals may be autocorrelated and require a technique for addressing this autocorrelation. This technique provides increased sensitivity but is more cumbersome as a result of the number of models needing to be generated.

Hawkins applied the preceding technique but used linear regression to fit the model in Eq. 19 (34). The resulting variables are still correlated (79). However, if the fault direction is known a priori, this method provides increased sensitivity to faults.

Triant (80) uses a similar concept with a modeling technique similar to k-nearest neighbors. Triant calls their technique universal process modeling (UPM). An overall "health" metric is provided based upon the residuals. In addition, a "bulls-eye" plot is used to designate which variables have a problem.

Multivariate CUSUM and Other Methods

Other methods based upon a multivariate extension of the cumulative sum chart (CUSUM) have been developed to try to address some of the issues with T^2 (34,38,43,45,63,77). One difficulty with CUSUM is that it requires the specification of the direction of the shift. In the univariate case, that amounts to testing plus or minus. However, in the multivariate case, as the number of variables increases, the number of possible directions grows dramatically. Crosier developed a method that considers the size of the fault but attempts to be independent of the direction of the fault (45). Besides direction, the Type I and II errors of different techniques may be sensitive to the number of variables and the correlation structure. Thus, the issues of the impact of the number of variables, correlation structure, variance level, size of fault, and direction of fault on the Type I and II errors have not been solved. In addition, the issues of normality, homoscedasticity, and autocorrelation must also be addressed for CUSUM techniques, although theoretically the sequential probability ratio test, upon which some CUSUMs are based, is valid for any distribution. Another technique that can be used for finding outliers in multivariate data is described by Rocke and Woodruff (79). Their paper also examines the difficulty of increasing dimensionality (i.e., number of variables). A multivariate exponentially weighted moving average control chart is examined by Lowry et al. (81). All these alternatives to T^2 can be used with time-series model residuals and data reduction techniques, too.

Neural Net Methods

The methods described previously are based upon statistical methods. Another method being used for fault monitoring and control based on multiple equipment traces is neural nets. The neural nets can be used to predict a wafer result as a function of the traces. If the prediction is outside of acceptable regions, then a fault is declared. The inputs to the neural net can include previous values to provide a method of handling autocorrelation (82–86). The neural net can also be used to predict in-control and out-of-control conditions (i.e., a go/no-go type of sensor).

Isolating the Fault

One additional issue is true for any of the multivariate detection methods. Fault detection is only one step of a three-step process:

- Detection (identification)—to detect the occurrence of a fault
- Isolation—to isolate which variables have changed (e.g., throttle valve variable is different)
- Classification (diagnosis)—to assign a cause of the fault (e.g., a malfunction has occurred in the pump)

Another component is prognosis, which is to predict that a fault will occur in the future. Prognosis may or may not include isolating and classifying the fault.

Table 8. Overview of Constructing MSPC

1.	Obtain data set representing in-control conditions with all sources of expected variation
2.	Determine transformations to make all variables normally distributed and homoscedastic
	Logarithmic
	Square root
3.	Select method for within-wafer autocorrelation
	Time series model
	Metrics from Dynamic Time Warping
4.	Select method for run-to-run and higher autocorrelations
	Time series model for mean
	EWMC for Co-variance matrix
	Neural nets
5.	Select method for monitoring dynamic behavior (autocorrelations)
	SPC charts on amount of change total and single time period
6.	Determine if data reduction will be used
	PCA
7.	Determine if data reconstruction will be used
	PLS
	Regression-adjusted variables
	K-nearest neighbors
	Neural nets
8.	Select MSPC chart if monitored variables are correlated (within-model variation)
	T^2
	MCUSUM
	MEWMA
9.	If data reduction is used, select method for outside model variations
	Q
	Univariate charts for each residual
10.	Select MSPC chart for monitoring dispersion
	COT^2
	No widely accepted chart for monitoring dispersion
11.	Select method for controlling overall Type I error
	Roy and Bose intervals
	Bonferroni
	MSPC Chart (T^2, MCUSUM, MEWMA)
12.	Use data to calculate control limits for various charts
13.	Determine method for isolating fault
	Contribution plots
	Data reconstruction (PLS, regression-adjusted variables, universal process modeling)

Thus, T^2 detects only out-of-control situations, but it does not identify which variable is out of control. There are two methods used for isolation. The first method is to also use univariate charts to try to assign the problem to a variable (35–37). However, the problem may appear in the univariate or in the multivariate chart; it is not required to appear in both. Limits from the Bonferroni inequalities or Roy and Bose intervals should be used to set up the univariate charts.

The second method is the single-sample variant of the linear discriminant function coefficient vector $\boldsymbol{a}$:

$$\boldsymbol{a} = \boldsymbol{S}^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}}) \tag{20}$$

This variable arises in the derivation of the quadratic form for T^2. It provides an indication of the contribution of each response to T^2. If the standard deviations of the variables are nearly equal, it has been proposed to use this variable to determine which response variable is responsible for the failure (41). It was suggested to make it dimensionless by scaling each a_i by its standard deviation (s_i).

This concept is similar to the contributions plot of principal components-based MSPC. In PCA, the contribution of each variable to the T^2 are presented in a single plot (i.e., each variable's "contribution" is plotted). Variables with large contributions are most likely to be involved in the fault. In addition, the residual (Q_i) for each variable is also used to identify variables that may be involved. One item to note is that "smearing" has been cited as a possible problem with contribution plots (i.e., a fault in one variable is smeared to other variables for which it is highly correlated).

Methods based on data reconstruction naturally provide isolation. The variable most likely at fault is the one whose residual is biggest. Again note that correlation can result in smearing (i.e., a faulty variable may produce a poor prediction in models for which it is heavily weighted giving a residual larger than the residual for its own prediction).

Methods for fault classification are beyond the scope of this article. However, many times, bright engineers and technicians can guess the cause of the fault given the variables responsible for triggering an alarm. Determining the cause of the fault is part of the corrective procedure of Fig. 2. Thus, classification is also a necessary step in univariate SPC.

Summary of Steps to Create MSPC

All the various aspects and options are listed in Table 8. The table shows the steps that must be addressed to produce robust and sensitive MSPC. Note that many of the same steps are needed for univariate SPC, although the issue of which variable to isolate in the fault is not an issue for univariate SPC. For neural nets, the resulting required steps might be quite different depending upon the output of the neural net. If the output is a go/no-go value, many of the steps are skipped. However, if the model's output is a prediction of each variable, then most steps apply.

ACKNOWLEDGMENTS

The concepts of control in depth and breadth to achieve factory control were developed by the TI Semiconductor Division 0.72 μm Factory Control Team.

GLOSSARY OF FACTORY CONTROL TERMS

α. Type I error rate

ADC. Automatic defect classification

Analysis of Variance (ANOVA). One of the statistical methods used to evaluate the data from an experimental design to determine sources of variability

Assignable Cause. A source of variation that is nonrandom; a change in the source will produce a significant change of some magnitude in the response

Attribute. A characteristic that may take on only one value (e.g., 0 or 1)

Autocorrelation. Correlation between samples of the same variables; implies a dynamic process

β. Type II error rate

Capability Index. The index of the process spread versus specification width; the potential process capability (C_p is the index used to measure the process capability with respect to centering between the specification limits.)

Cause. That which produces an effect or brings about a change

CFM. Contamination-free manufacturing; generic term used for the practices to control contamination and reduce defects

Change Control. The process of managing changes through appropriate documentation, validation and notification

Common Cause. The combined effect of multiple sources of variation that are inherent in a process. These causes define the natural fluctuation of the process.

Containment Control. The prevention of moving abnormal or out-of-specification material to the next process or shipping to the customer

Control Chart. A graphical method for evaluating whether a process is in a state of statistical control (The decisions are made by comparing values of some statistical measure calculated from the data with control limits.)

Control Limits. Lines on a control chart that serve as a basis for judging whether a set of values is in a state of statistical control (These limits are calculated from process data.)

Control System. A set of closed loop activities that provide instructions to processes and detects and responds to nonexpected conditions

Corrective Action. Integral part of a control system that responds to information generated by a monitoring system

Customer Disruption. Any event caused by a supplier that interrupts the normal economic cycle of business (e.g., late or missed deliveries, customer production line fall outs, consumer recalls due to reliability failures)

Design of Experiments (DOE). The process of planning and analyzing experimental data to derive statistically valid conclusions (The objective of the experiment is to discover the cause-and-effect relationship between control factors and responses.)

EDS. Energy dispersive spectroscopy, a type of X-ray analysis used on SEMs to perform composition analysis of particles/defects

Failure Analysis. The process to determine the failure mode and mechanism of a product or process

Failure Control. The process to detect, contain, analyze root cause, and implement corrective actions to problems past the concurrent control methods to prevent reoccurrence of the same problem

Fault Tree Analysis. The technique of "top-down" methodical analysis depicting the interrelationship between an undesired system state and its subsystem states (It begins with an assumed undesirable event at the top or system level and identifies the events at subsequent lower levels in the system that can cause the undesirable top event.)

FDC. Fault detection and classification, used currently to refer to monitoring and SPC using real-time equipment traces (signals)

FMEA. Failure mode and effects analysis, a structured procedure for identifying and minimizing effects of as many potential failure modes as possible

Heteroscedasticity. Nonuniform variance (i.e., the variance of a variable is a function of the value of that variable or conditions)

In Control. The condition describing a process that is only being influenced by common causes.

ISPM. In situ process monitor; traditionally had meant in situ particle monitor

Machine Capability. The measure of the ability of a machine to meet specification limits with a controlled set of conditions

Measurement Bias. The difference between the observed average of measurements and the standard sample for which the "true standard" value is known

Measurement Linearity. The systematic differences in the bias values of a measuring system throughout the expected operating range of the gage

Measurement Repeatability. The variation of a measurement system obtained by repeating measurements on the same sample back-to-back using the same measurement conditions

Measurement Reproducibility. The variation among the averages of measurements made at different measurement conditions (e.g., different operators, different environments, and possibly different laboratories)

Measurement Stability. The total variation in the measurements obtained with a measurement system on the same master or parts when measuring a single characteristic over an extended time period

Measurement System. The process for gauging a parameter (The inputs for this process are the gauge, the operator, specification procedures, and management methods.)

Methods. Procedures, processes, techniques, or evaluations used to monitor and control a particular aspect of a business operation

Metrology. Measurement science and the application of measurement science

Model-Based Process Control (MBPC). A specific form of feedback/feedforward control using process models (See *Run to Run Control.*)

Multivariate. Statistics with more than one variable

Normal Distribution. A bell-shaped curve that extends indefinitely in both directions (It also may be referred to as Gaussian.)

NTRS. National Technology Roadmap for Semiconductors, the roadmap, created by the Semiconductor Industry Association (SIA), that predicts what device technology will be in production and what will be needed to allow manufacturing of devices is a cost-effective manner

Out of Control. The condition describing a process from which all the special causes of variation have not been eliminated (This condition is evident on a control chart when a point falls outside a control limit or a nonrandom pattern is produced.)

OOC. Out of control (i.e., to fail a SPC chart test), usually used as percent of all SPC charts in factory which have "alarmed" in a certain period of time (i.e., 3% OOC for all of last week

Outlier Control. Control methods that detect material that is outside a predetermined distribution for one or more critical parameters and that applies appropriate actions to correct the assignable cause responsible for the event

Outlier Material. Material that is within specification but outside a predetermined distribution for one or more critical parameters

Pareto Chart. The graphical depiction of data in bar chart format that identifies the major contributors in an analysis

Percent GRR. The percent of the specification tolerance consumed by the measurement system repeatability and reproducibility variations

Prevention Control. Actions or designed in system used to prevent potential problems

Process. A set of interrelated work activities that are characterized by specific inputs and value-added tasks that produce a set of specific outputs

Process Capability. The measure of process variation resulting from common causes; has a spread of plus or minus three standard deviations

Qualification. The methodologies to demonstrate the inherent quality and reliability of the process or product that meets qualification objectives and customer requirements

Quality Function Deployment (QFD). A method for translating user requirements into the appropriate technical requirements for each stage of marketing, product planning, product design, manufacturing engineering, production, and sales and service

Reliability Monitor. A set of stresses and tests performed on partial or fully assembled product to identify potential reliability problems

Root Cause. The condition that is the origin or source of a fault/failure

Run to Run (RtR) Control. Control by changing the recipe as needed to keep the process output on target (See also *Model-Based Process Control.*)

SEM. Scanning Electron Microscope, used both for critical dimension measurement as well as a high-resolution microscope with chemical analysis capability

Shewhart Chart. Most common SPC chart

Special Cause. The variation that is not inherent in a process (It is a source of intermittent variation that is unpredictable or unstable.)

SRAM. Static random access memory, type of memory chip and also used to perform bit mapping

SSA. Spatial signature analysis, used with defect and electrical (parametric, yield) data to assist in identifying root cause

Standard Deviation. The unit of measure that is used to describe the width or spread of a distribution or pattern

Statistical Process Control (SPC). A control method that applies statistical techniques to understand and analyze variation in a process and that applies appropriate actions to achieve and maintain a state of statistical control

Test Power. The probability a change of a particular size will be detected by the particular fault-detection method

Tolerance. The specification range within which a product is considered acceptable

Type I Error. The error of a test declaring a sample "bad" when in fact it is "good"

Type II Error. The error of a test declaring a sample "good" when in fact it is "bad"

Univariate. Statistics with a single variable

Wafer Level Reliability Control (WLRC). A control method to detect, analyze, and correct reliability problems early by stress testing with voltages, currents, and temperature

BIBLIOGRAPHY

1. Chrysler Corporation, Ford Motor Company, General Motors Corporation, *Potential Failure Mode and Effects Analysis (FMEA),* 1995.
2. American Supplier Institute, *Quality Function Deployment for Products,* 1995.
3. B. S. Dhilon, *Quality Control, Reliability and Engineering Design,* New York: Dekker, 1985.
4. Chrysler Corporation, Ford Motor Company, General Motors Corporation, *Measurement Systems Analysis,* 2nd ed., 1995.
5. Evaluating Automated Wafer Measurement Instruments, SEMATECH Technology Transfer Document #94112638A-XFR.
6. D. C. Montgomery, *Introduction to Statistical Quality Control,* 2nd ed., New York: Wiley, 1991.
7. Texas Instruments Statistical Process Control Guidelines.
8. M. J. Harry and J. R. Lawson, *Six Sigma Producibility Analysis and Process Characterization,* Reading, MA: Addison-Wesley, 1992.
9. Western Electric Company, *Statistical Quality Control Handbook,* 2nd ed., Mack, 1956.
10. A. V. Czitrom and K. Horrell, SEMATECH Qual Plan: A qualification plan for process and equipment characterization, *Future Fab Int.,* **1** (1): 45, 1996.
11. Starfire from Domain Solution Corp (formerly, BBN Domain Corp), Cambridge, MA [Online]. Available www: http://www.domaincorp.com
12. S. W. Butler, Process control in semiconductor manufacturing, *J. Vac. Sci. Technol. B, Microelectron. Process. Phenom.,* **13**: 1917–1923, 1995.
13. T. Smith et al., Run by run advanced process control of metal sputter deposition, *Electrochem. Soc. Proc.,* **97** (9): 11–18, 1997.
14. T. E. Bensen et al., Sensor systems for real-time feedback control of reactive ion etching, *J. Vac. Sci. Technol. B, Microelectron. Process. Phenom.,* **14**: 483–488, 1996.
15. R. DeJule, CMP challenges below a quarter micron, *Semicond. Int.,* **20** (13): 54–60, 1997.
16. I. Tepermeister et al., In situ monitoring of product wafers, *Solid State Technol.,* **39** (3): 63–68, 1996.
17. P. Timans, Temperature measurement in rapid thermal processing, *Solid State Technol.,* **40** (4): 63–74, 1997.
18. Y. Lee, B. Khuri-Yakub, and K. Saraswat, Temperature measurement in rapid thermal processing using the acoustic temperature sensor, *IEEE Trans. Semicond. Manuf.,* **9**: 115–121, 1996.
19. P. Biolsi et al., An advanced endpoint detection solution for <1% open areas, *Solid State Technol.,* **39** (12): 59–67, 1996.
20. T. Carroll and W. Ramirez, On-line state and parameter identification of positive photoresist development, *AIChE J.,* **36**: 1046–1053, 1990.
21. G. Lu, G. Rubloff, and J. Durham, Contamination control for gas delivery from a liquid source in semiconductor manufacturing, *IEEE Trans. Semicond. Manuf.,* **10**: 425–432, 1997.
22. Ferran Scientific [Online]. Available www: http://www.ferran.com/main.html
23. S. Leang and C. Spanos, A novel in-line automated metrology for photolithography, *IEEE Trans. Semicond. Manuf.,* **9**: 101–107, 1996.
24. S. Bushman and S. Farrer, Scatterometry measurements for process monitoring of polysilicon gate etch, *Proc. SPIE: Process, Equipment, Materials Control Integrated Circuit Manufacturing III,* **3213**: 79–90, 1997.
25. R. Patrick, N. Williams, and C. Lee, Application of RF sensors for real time control of inductively coupled plasma etching equipment, *Proc. SPIE: Process, Equipment, Materials Control Integrated Circuit Manufacturing III,* **3213**: 67–72, 1997.
26. N. Hershkowitz and H. L. Maynard, Plasma characterization and process control diagnostics, *J. Vac. Sci. Technol. A., Vac. Surf. Films,* **11**: 1172–1178, 1993.
27. L. Peters, In situ particle monitoring slowly matures, *Semicond. Int.,* **48**, 1998.
28. Z. M. Ling et al., Analysis of within-run process variations using automated wafer-position tracking in workstream, *Extended Abstracts 187th Meeting Electrochemical Soc.,* **95** (1): 524–525, 1995.
29. Wafer Sleuth Implementation Guide, SEMATECH Technol. Transfer Document 91060587A-ENG, 1991.
30. G. Scher, Wafer tracking comes of age, *Semicond. Int.,* **14** (6): 126–131, 1991.
31. Silicon supplier in line statistical process control and feedback for VLSI Manufacturing, *IEEE Trans. Semicond. Manuf.,* **13**: 1990.
32. McPherson, Rost, Dickerson, Wafer Level Reliability Testing, Internal Document, Texas Instruments.
33. C. K. Chow, Projection of circuit performance distributions by multivariate statistics, *IEEE Trans. Semicond. Manuf.,* **2**: 60–65, 1989.
34. D. M. Hawkins, Multivariate quality control based on regression-adjusted variables, *Technometrics,* **33**: 61, 1991.
35. F. B. Alt, Multivariate quality control, in S. Kotz and N. L. Johnson (eds.), *Encyclopedia of Statistical Sciences,* Vol. 6, New York: Wiley, 1985, pp. 110–122.
36. N. F. Hubele, A multivariate and stochastic framework for statistical process control, in J. B. Keats and N. F. Hubele (eds.), *Statistical Process Control in Automated Manufacturing,* New York: Dekker, 1989.
37. T. P. Ryan, *Statistical methods for quality improvement,* New York: Wiley, 1989.
38. N. D. Smith, Multivariate cumulative sum control charts, Ph.D. Dissertation, Univ. of Maryland, College Park, MD, 1987.
39. H. Hotelling, Multivariate quality control, in E. Eisenhart, M. Hastay, and W. A. Wallis (eds.), *Techniques of Statistical Analysis,* New York: McGraw-Hill, 1947, pp. 111–184.
40. N. Tracy, J. Young, and R. Mason, Multivariate control charts for individual observations, *J. Qual. Technol.,* **24** (2): 88–95, 1992.

41. D. F. Morrison, *Multivariate Statistical Methods,* New York: McGraw-Hill, 1990.
42. J. Prins and D. Mader, Multivariate control charts for grouped and individual observations, *Quality Eng.,* **10** (1): 49–57, 1997–98.
43. F. B. Alt and G. E. Bedewi, SPC of dispersion for multivariate data, *ASQC Quality Congress Trans.,* Anaheim, 1986, p. 248.
44. J. D. Healy, A note on multivariate CUSUM procedures, *Technometrics,* **29**: 409–412, 1987.
45. R. B. Crosier, Multivariate generalizations of cumulative sum quality-control schemes, *Technometrics,* **30**: 291, 1988.
46. S. Leang and C. J. Spanos, Statistically based feedback control of photoresist application, *Proc. ASM,* Boston, 1991, pp. 185–190.
47. F. Aparisi, Sampling plans for the multivariate T2 control chart, *Qual. Eng.,* **10** (1): 141–147, 1997–98.
48. J. Stefani, L. Loewestein, and M. Sullivan, On-line diagnostic monitoring of photoresist ashing, *IEEE Trans. Semicond. Manuf.,* **8**: 2–9, 1995.
49. L. Loewenstein, J. Stefani, and S. W. Butler, A first-wafer effect in remote plasma processing: The stripping of photoresist, silicon nitride and polysilicon, *J. Vac. Soc. Tech. B, Microelectron. Process. Phenom.,* **12**: 2810, 1994.
50. S. W. Butler, Issues and solutions for applying process control to semiconductor manufacturing, in P. F. Williams (ed.), *Plasma Processing of Semiconductors. Series E: Applied Sciences,* Vol. 336, Norwell, MA: Kluwer, 1997.
51. B. M. Wise, N. L. Ricker, and D. J. Veltkamp, Upset and sensor failure detection in multivariate processes, *AICHE Meeting,* 1989.
52. D. Wardell, H. Moskowitz, and R. Plante, Run-length distributions of special-cause control charts for correlated processes, *Technometrics,* **36** (1): 3–17, 1994.
53. J. Lucas, Discussion, *Technometrics,* **36** (1): 17–19, 1994.
54. B. Adams, W. Woodall, and C. Superville, Discussion, *Technometrics,* **36** (1): 19–22, 1994.
55. W. Fellner, Discussion, *Technometrics,* **36** (1): 22–23, 1994.
56. D. Wardell, H. Moskowitz, and R. Plante, Rejoinder, *Techometrics,* **36** (1): 23–27, 1994.
57. A. Sweet, Using coupled EWMA control charts for monitoring processes with linear trends, *IIE Trans.,* **20**: 404–408, 1988.
58. D. Montgomery and C. Mastrangelo, Some statistical process control methods for autocorrelated data, *J. Qual. Technol.,* **23** (3): 179–193, 1991.
59. F. Faltin and W. Woodall, Discussion, *J. Qual. Technol.,* **23** (3): 194–197, 1991.
60. J. MacGregor, Discussion, *J. Qual. Technol.,* **23** (3): 198–199, 1991.
61. T. Ryan, Discussion, *J. Qual. Technol.,* **23** (3): 200–202, 1991.
62. D. Montgomery and C. Mastrangelo, Response, *J. Qual. Technol.,* **23** (3): 203–204, 1991.
63. D. M. Hawkins, A CUSUM for a scale parameter, *J. Qual. Technol.,* **13** (4): 228, 1981.
64. H. Guo, C. Spanos, and A. Miller, Real time statistical process control for plasma etching, *IEEE/SEMI Int. Semicond. Manuf. Sci. Symp.,* 1991, pp. 113–118.
65. S. Lee and C. Spanos, Equipment analysis and wafer parameter prediction using real-time tool data, *1994 Int. Symp. Semiconductor Manufacturing,* 1995, pp. 133–136.
66. H.-F. Guo, Real time statistical process control for plasma etching, Masters Thesis, Univ. California, Berkeley, CA, 1991.
67. S. Lee and C. Spanos, Prediction of wafer state after plasma processing using real-time tool data, *IEEE Trans. Semicond. Manuf.,* **8**: 252–261, 1995.
68. C. Spanos et al., Real-time statistical process control using tool data, *IEEE Trans. Semicond. Manuf.,* **5**: 308–318, 1992.
69. S. Lee et al., RTSPC: A software utility for real-time SPC and tool data analysis, *IEEE Trans. Semicond. Manuf.,* **8**: 17–25, 1995.
70. G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis, Forecasting and Control,* 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, 1988.
71. D. J. Wright, Forecasting data published at irregular time intervals using an extension of Holt's method, *Manage. Sci.,* **32** (4): 499–510, 1986 OR 1980.
72. T. H. Smith and D. Boning, Non-periodic lot processing, random measurement delays, and intermittent lot processing with an extended predictor corrector controller, *44th Nat. Symp. Amer. Vacuum Soc.,* Oct. 1997; *J. Vac. Sci. Technol.,* 1998, submitted for publication.
73. S. B. Dolins, A. Srivastava, and B. E. Flinchbaugh, Monitoring and diagnosis of plasma etch processes, *IEEE Trans. Semicond. Manuf.,* **1**: 23–27, 1988.
74. S. B. Dolins et al., *Apparatus and method for production process diagnosis using dynamic time warping.* U.S. Patent No. 4,861,419, 1989.
75. D. White et al., Methodology for robust and sensitive fault detection, *Electrochemical Soc. Proc.,* **97** (9): 55–63, 1997.
76. N. B. Gallagher et al., Development and benchmarking of multivariate statistical process control tools for a semiconductor etch process: Improving robustness through model updating, *IFAC ADCHEM'97,* Banff, Canada, 1997.
77. W. H. Woodall and M. M. Ncube, Multivariate CUSUM quality-control procedures, *Technometrics,* **27** (3): 285, 1985.
78. B. Wise and N. Gallagher, The process chemometrics approach to process monitoring and fault detection, *J. Proc Cont.,* **6**: 329–348, 1996.
79. D. Rocke and D. Woodruff, Identification of outliers in multivariate data, *J. Amer. Statistical Assoc.—Theory and Methods,* **91** (432): 1047–1061, 1996.
80. Triant Technologies, [Online]. Available www.triant.com.
81. C. A. Lowry et al., A multivariate exponentially weighted moving average control chart, *Technometrics,* **30**: 291–303, 1988.
82. E. A. Rietman and E. R. Lory, Use of neural networks in modeling semiconductor manufacturing processes: An example for plasma etch modeling, *IEEE Trans. Semicond. Manuf.,* **6**: 343–347, 1993.
83. R. Shadmehr et al., Principal component analysis of optical emission spectroscopy and mass spectrometry: Application to reactive ion etch process parameter estimation using neural networks, *J. Electrochem. Soc.,* **139**: 907–914, 1992.
84. B. Kim and G. May, Real-time diagnosis of semiconductor manufacturing equipment using a hybrid neural network expert system, *IEEE Trans. Compon. Packag. Manuf. Technol. C,* **20**: 39–47, 1997.
85. M. Baker, C. Himmel, and G. May, Time series modeling of reactive ion etching using neural networks, *IEEE Trans. Semicond. Manuf.,* **8**: 62–71, 1995.
86. E. Rietman and S. Patel, A production demonstration of wafer-to-wafer plasma gate etch control by adaptive real-time computation of the over-etch time from in situ process signals, *IEEE Trans. Semicond. Manuf.,* **8**: 304–308, 1995.

Reading List

General

SEMATECH: www.sematech.org

I300I (dedicated to 300mm issues): www.i300i.org

National Technology Roadmap: http://www.sematech.org/public/roadmap/index.htm

Semiconductor Subway: http://www-mtl.mit.edu/semisubway.html

Semiconductor Equipment and Materials International (SEMI): http://www.semi.org

Semiconductor Research Corporation (SRC): http://www.semi.org/

Semiconductor International: *http: / / www.semiconductor-intl.com*

Solid State Technology: http://www.solid-state.com/

Semiconductor Online: http://www.semiconductoronline.com/

Semiconductor SuperSite.Net: http://supersite.net/semin2/docs/home.htm

FabTech: www.fabtech.org

TechWeb: http://www.techweb.com/

Semiconductor Process Equipment and Materials Network: http://www.smartlink.net/~bmcd/semi/cat.html

Semiconductor.Net—The Semiconductor Manufacturing Industry Resource for Products, Services and Information: http://www.semiconductor.net/

SemiSource, Semiconductor Resource Guide, published annually by Semiconductor International

Solid State Technology Resource Guide, published annually by Solid State Technology

American Vacuum Society (AVS) Buyers Guide: http://www.aip.org/avsguide

R. J. Muirhead, *Aspects of Multivariate Statistical Theory,* New York: Wiley, 1982.

Conferences and Supporting Organizations

Electrochemical Society, Inc., http://www.electrochem.org/

American Vacuum Society (AVS) Manufacturing Science and Technology Group (MSTG): http://www.cems.umn.edu/~weaver/mstg/mstg-subway.html

International Symposium on Semiconductor Manufacturing (ISSM): http://www.issm.com

Advanced Semiconductor Manufacturing Conference (ASMC): http://www.semi.org/Education/asmc/main.html

SPIE Microelectronic Manufacturing: http://www.spie.org/info/mm/

SC Control and Control Software

University of Michigan Controls Group: http://www.engin.umich.edu/research/controls/

Berkeley Computer Aided Manufacturing (BCAM): http://radon.eecs.berkeley.edu/

Maryland University, The Institute for Systems Research: http://www.isr.umd.edu/

SEMATECH & MIT Run by Run Benchmarking: http://www-mtl.mit.edu/rbrBench/

TRIANT Technologies, Inc.: http://www.triant.com/

Semy: www.semy.com

Domain Solution Corp (formerly, BBN Domain Corp.), Cambridge, MA: http://www.domaincorp.com

Umetrics, Winchester, MA: http://www.umetri.se (also good Chemometrics links)

Brookside Software: http://www.brooksidesoftware.com/

Brooks Automation, Richmond, BC, Canada: http://www.brooks.com/bac.htm

Real Time Performance, Sunnyvale, CA: http://www.rp.com

ControlWORKS, Dallas, TX: http://www.ti.com/control

Fastech: http://www.fastech.com

Voyan Technology, Santa Clara, CA

V. Bakshi, Fault Detection and Classification (FDC) Software Benchmarking Results, SEMATECH Technol. Rep. 97123433A-TR, 1998.

V. Bakshi, Fault Detection and Classification Software for Plasma Etchers: Summary of Commercial Product Information, SEMATECH Technol. Rep. 97083337A-XFR, 1997.

Manufacturing Execution Systems (MES) / Computer Integrated Manufacturing (CIM) / Equipment Integration Automation: Software Used to Run and Track Fab, Perform SPC, etc.

Fastech, http://www.fastech.com

Real Time Performance, Sunnyvale, CA: http://www.rp.com

Consillium: http://www.consilium.com/about/about.htm

Promis: http://www.promis.com

T. Byrd and A. Maggi, Challenges to plug and play CIM, *Future Fab International,* pp. 77–81.

M. Greig and A. Weber, AMD & ObjectSpace, Inc., *Future Fab International,* pp. 73–74.

Inspection Tools

KLA-Tencor, San Jose, CA: www.kla-tencor.com

Orbot (owned by Applied Materials), Santa Clara, CA: http://www.appliedmaterials.com/products/pdc.html

Inspex: (508) 667-5500.

OSI, Fremont, California: (510) 490-6400.

Optical Review / SEM / EDS Analysis

Ultrapointe (distributed by KLA-Tencor), San Jose, CA: http://www.ultrapointe.com/Ultrapointe/home.htm

Leica, Deerfield, IL: http://www.leica.com

JEOL, Peabody, MA: http://www.jeol.com

FEI, Hillsboro, OR: semisales@feico.com

Hitachi Scientific Instruments, Mountain View, CA: (650) 969-1100.

Opal (owned by Applied Materials), Santa Clara, CA: http://www.appliedmaterials.com/products/pdc.html

Noran Instruments Inc., Middleton, WI: http://www.noran.com

Oxford Instruments Inc., Concord, MA: http://www.oxinst.com

Data Analysis, Data Warehousing, Data Mining, Bit Mapping, Wafer Tracking, Etc.

Knight's Technlogy, Sunnyvale, CA: http://www.knights.com

DYM, Bedford, MA: http://www.dym.com

LPA Software, South Burlington, VT: (802) 862-2068

Quadrillion, Company Information: http://www.quadrillion.com/quadinfo.htm

DeviceWare Corporation: http://www.dware.com/

Maestro, Data Management [JJT Inc.]: http://www.jjt.com/data.-man.html

Sleuthworks: http://www.sleuthworks.com/doc/

SAS: http://www.sas.com

KLA-Tencor, San Jose, CA: www.kla-tencor.com

Reliability, Parametric Testing

Keithley Instruments Semiconductor Products: http://www.keithley.com/TIG/SBU/

STEPHANIE WATTS BUTLER
RUDY YORK
MARYLYN HOY BENNETT
TOM WINTER
Texas Instruments