DIAGNOSIS OF SEMICONDUCTOR PROCESSES

In very deep-submicron semiconductor manufacturing (VDSM) technologies (i.e., 0.18 μm or below), the cost of setting up a fabrication facility is about 3 to 4 billion dollars, and the equipment cost for each stage is easily a million dollars or more. According to one study (1), if the production of an Intel Pentium chip were delayed by four months in 1997, it would cost Intel about 400 million dollars. In other words, if a one-day delay were caused by a malfunction in any one of hundreds of pieces of processing equipment (process flow should have been fully functioning during the pilot fabrication phase and have had no problem during manufacturing), the cost for the malfunction would be about 3.3 million dollars on average. Therefore, timely diagnostic capability for equipment/process malfunction during manufacturing is required.

Process diagnosis is important not only during mass manufacturing, but also in the ramp-up phase of the manufacturing line (2). However, during the ramp-up phase, little manufacturing information or equipment/process diagnosis history is known; therefore the key during this phase is the modeling and simulation of the equipment/process and development of predictive models for mass manufacturing. The resulting models can therefore be used for diagnosis during mass manufacturing.

Semiconductor processing of a functional chip usually consists of several hundreds of steps and can be divided into five distinct operations: wafer preparation, wafer processing, wafer probe test, packaging, and final test. Wafer processing is usually considered the most important step for yield improvement. Therefore, most monitoring and diagnosis focuses on the wafer processing stage. In this stage, four sources of information can be obtained for diagnostic purposes:

- First, the equipment maintenance history, which contains the preventive maintenance and repair records that can help correctly diagnose malfunctioning equipment.
- Second, the in situ monitoring data, which are read directly from the embedded sensors or gauges that are connected to the equipment, such as pressure, gas flow, and temperature sensors. These data are becoming more and more important because they provide real-time feedback to the equipment/process control and also obviate further processing of the wafer if faults are detected.
- Third, the in-line measurement data obtained after completing a process step, such as the thickness of a film after deposition or the uniformity of a process profile after etching. During a short-loop process involving one or more pieces of equipment or multichamber equipment, in-line measurement data can clearly indicate the health of the process. They can also be used to correlate to the in situ monitoring data.
- Finally, the wafer probe or functional test data, which provide the final check point for the success of the processing. Electrical/functional testing can easily include tens of tests (IDDQ, open–short, etc.), which, when correlated with in-line measurement data, can help identify which process step caused the failure of the test(s).

Other sources of information, such as spot defect density and layout density, are also important in helping diagnose the process.

This article focuses on the working algorithms and existing systems for equipment/process monitoring and diagnosis during ramp-up and manufacturing. The disciplines involved include statistical process control techniques, expert systems, reasoning methods, neural networks, web/Java technologies, networking, distributed systems, quantitive process/equipment/device modeling, and semiconductor manufacturing.

General methods for process diagnosis and yield improvement in semiconductor manufacturing are selectively illustrated in the next section. The section after describes techniques used in monitoring and diagnosis for the unit process step and equipment levels in several systems. It is followed by a summary of the algorithms and systems used in monitoring and diagnosis of the process flow level during ramp-up and manufacturing. Finally, conclusions and future work are outlined.

GENERAL METHODS FOR YIELD IMPROVEMENT IN SEMICONDUCTOR MANUFACTURING

In this section, we overview the general methods for yield diagnosis in semiconductor manufacturing. In semiconductor manufacturing terms, the *die yield* is the fraction of dies on the yielding wafers that is not discarded before reaching assembly and final test (3). There may be further classification of performance bands with respect to the functional dies. The following list includes the ones we think important for today's very deep-submicron semiconductor manufacturing technologies. The list is by no means complete and is in random order.

Correlation between Defect Localization and In-Line Wafer Inspection

Defect localization can be obtained from functional testing, such as Boolean, scan, or IDDQ. This correlation used to be rarely made in practice, but it is getting more and more important.

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*. Copyright © 1999 John Wiley & Sons, Inc.

Correlation between Defect Sensitivity and Integrated Circuit Layout

Although defect sizes are decreasing as clean-room technology and equipment contamination control techniques improve, the sizes of device/interconnect features are decreasing equally rapidly. The effects of defect size and spot defect locations on functioning of the chip have a lot to do with the layout. The denser the features on the layout, the more susceptible the circuit is to defects. This subject has been covered in Refs. 4–6.

Statistical Monitoring, Diagnosis, and Control of Equipment/Process

A major force behind the evolution of statistical process monitoring, diagnosis, and control is the recent availability of automated in situ data collection and real time data processing capabilities. This is one of the major subjects covered in this article.

Minimization of Die-to-Die Variation, Wafer-to-Wafer Variation, and Lot-to-Lot Variation

This information highlights different aspects of how controllable a process is. Modeling and simulation of these variations have been carried out in Refs. 2 and 7.

Simultaneous Correlation between Defect, Testing, and Layout

From data obtained from a functional test (a test that does not stop when a fault is discovered), one can identify the location of a fault. From in-line wafer monitoring data, one has information on where the defects land and what their shapes are. From layout information, one knows the probability that a fault will occur. By correlating these three data, one should have a fairly clear picture of how a defect grows (in-line wafer inspection in different steps reveals the growth history of the defects), how killing defects are formed from layout and defect localization, how effective the functional test is, etc. Much more work is needed on identifying defects directly from functional tests. This in turn can save a lot of money spent in layer-by-layer stripping for wafer failure analysis. Each wafer stripping can cost about $10,000.

Correlation between Contamination and Faults

Monitoring the contamination history can possibly map to faults.

Equipment Drift Detection given Functional-Test Data

Since wafers may traverse different pieces of equipment in the process, one needs to correlate the functional-test data with the path a wafer takes. Using this correlation, drifting equipment can be identified. Of course, another way is to have a better equipment monitoring system.

Contribution of Process Variation to Device Characteristics

Device characteristics should be modeled to reflect the statistical variation in a process. If the measured device characteristics do not match the predicted ones, one may be able to locate the process step that caused the problem using simulation. The simulation package pdDiagnosis from PDF Inc. (8) provides such a capability.

Contribution of Process Variation to Interconnect Performance

The process variation directly affects the variations in multilevel interconnect geometry. This in turn directly affects the performance of the circuit, due to the different delay characteristics caused by the variation in the interconnect. Since interconnect delay dominates the total delay for a global route signal in very deep-submicron technologies, more attention should be paid to the modeling, simulation, monitoring, and control of the processes, such as the CMP (chemical mechanical polishing) process, that have direct influence on the variation in interconnects.

Use of Short-loop Electrical Measurements for Yield Improvement

Short-loop electrical metrology can be used to carefully characterize and decouple wafer-level variability of critical processing steps (9). This technique, if widely applied to all critical processing steps, can help reduce the systematic variation from processes and therefore increase the process error margin.

MONITORING AND DIAGNOSIS AT THE UNIT PROCESS AND EQUIPMENT LEVEL

Introduction

During the manufacturing phase of a process, the process flow is fixed and must have few problems. Low yield or low performance in the circuit may be due to spot defects from contamination, malfunction, or drift, or recalibration in the equipment level may be needed. Traditionally, in-line measurements have been used to diagnose the failures or drift in the unit process/equipment. However, with the advent of distributed measurement technology, more in situ measurements are taken to help monitor the health of the equipment. The systems described below exploit both in situ and in-line measurement data to monitor and diagnose using different diagnostic algorithms.

Integrated Monitoring and Diagnosis using Evidential Theory

The Berkeley computer-aided manufacturing system (10) uses in-line, maintenance, and real time monitoring data that are collected and stored in an integrated relational database. Six functions that contribute to the profitable operation of manufacturing equipment have been identified and implemented: real time monitoring, statistical process control (SPC), equipment maintenance record keeping, fault diagnosis, the efficient development of new recipes, and the development and maintenance of equipment models. Among these, the BCAM (Berkeley computer-aided manufacturing) diagnostic system supports both qualitative and quantitative information for diagnosis based on the Demster–Shafer model for fault inference (11,12). This method provides for consistent and unambiguous evidence combination. This is accomplished by combining evidence originating from equipment maintenance records, from real time equipment data, and from measurements on the finished process step. Using this information, the causes of equipment malfunctions are inferred through the resolution of qualitative and quantitative constraints. The qualitative constraints describe the normal operation of the

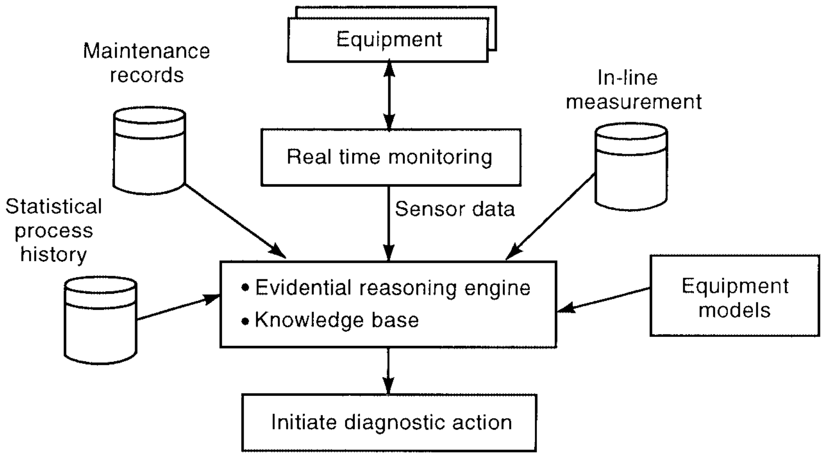

Figure 1. The three stages of diagnosis for low-pressure chemical vapor deposition (LPCVD) reactors.

equipment. The quantitative constraints are numerical models that apply to the manufacturing step in question. These models are specifically created and characterized through experimentation and statistical analysis. The violation of these constraints is linked to the evaluation of continuous *belief functions* for the calculation of the *belief* associated with the various types of failure. The belief functions encapsulate the experience of many equipment maintenance specialists, real-time in situ data monitoring via SPC, and the deviation of in-line measurement from semiphysical equipment models. Once created, the belief functions can be fine-tuned automatically, drawing from historical maintenance and diagnosis records. These records are stored in symbolic form in order to facilitate this task.

The three stages of diagnosis for low-pressure chemical vapor deposition (LPCVD) reactors are shown in Fig. 1. Figure 2 shows the output of an example that uses this method to detect an emerging pressure controller problem in the reactor. On the left side of this graph we start with the beliefs associated with the various faults after examining the maintenance records of the reactor. During the deposition, sensor readings are interpreted and the belief of the various faults is plotted in real time. Finally, after the in-line wafer measurements, the final beliefs are displayed on the right side of the same diagram. For the example in Fig. 2, the system first conducted maintenance diagnosis and found that there was a slight chance for excessive deposition during the next run. The system reached this conclusion by analyzing the tube cleaning history. Since the belief given to this problem was small (0.13 on the scale from 0 to 1), no action was taken, and the process continued.

At the start of deposition, the system examined the time needed to reach a stable deposition temperature. This was found to be longer than usual and contributed to the belief associated with the following faults: thermocouple out of calibration and temperature-controller problem. During deposition, however, the pressure readings were consistently higher than expected. So the belief in the pressure-controller problem quickly reached a high value (0.76), overshadowing all other faults. Finally, after the wafer measurements, some be-

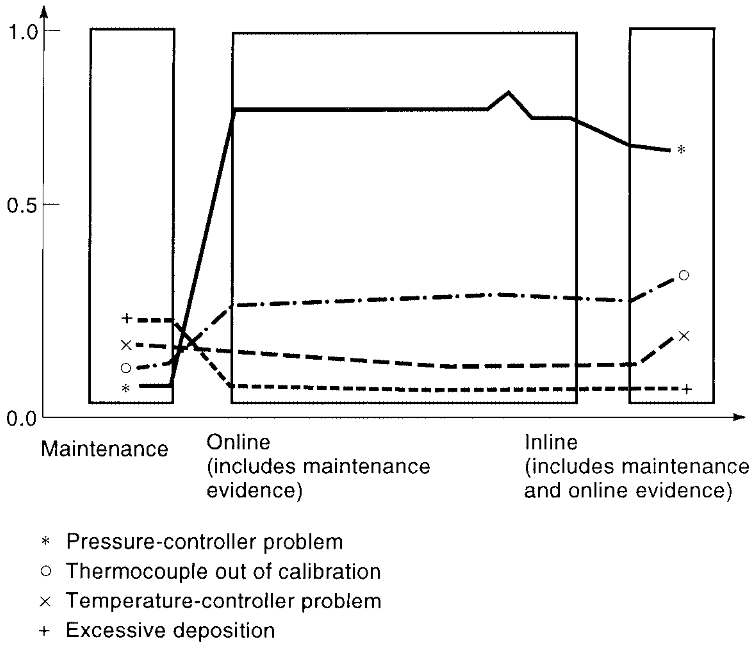

Figure 2. Existing pressure-controller problem. Belief in top faults is shown for the maintenance, real time, and in-line diagnostic stages from a process run on an LPCVD reactor.

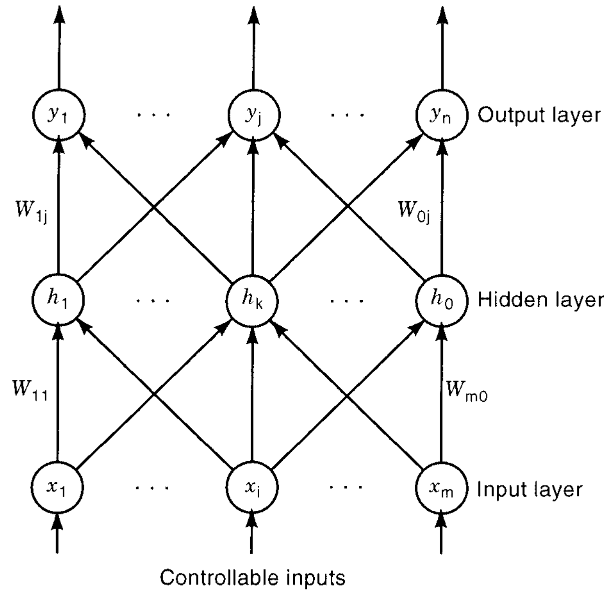

Figure 3. FFEBP neural network, showing input, hidden, and output layers.

lief was assigned to thermocouple out of calibration, while the pressure-controller problem stayed at the top of the ranked fault list. These inferences were later verified by the maintenance technician.

Similarly, this method has successfully been applied to plasma etching diagnosis (13).

Application of Neural Networks in Equipment Diagnosis

Neural networks have been studied for quite some time, but not until recently have they been successfully applied to the modeling of semiconductor fabrication processes, such as plasma etching and LPCVD (14–17). In these applications, neural networks have been shown to exhibit improved accuracy, compared with statistical approaches to modeling the highly nonlinear behavior of the processes in question. The neural network architecture is determined by the number of layers and the number of neurons per layer. In general, the numbers of input- and output-layer neurons is uniquely determined by the number of process inputs and responses in the modeling application (14). The trick to building a good neural network application depends on the selection of the optimal number of hidden-layer neurons based on the criteria of learning capability, prediction (or generalization) capability, and convergence speed. A popular neural network training algorithm applied in semiconductor process diagnosis is the feedforward, error backpropagation (FFEBP) algorithm (Fig. 3), for which the important design parameters are learning rate, initial weight range, momentum, and training tolerance.

For example, in Ref. 16, the goal is to design an optimal neural network for a specific semiconductor manufacturing problem: modeling the etch rate of polysilicon in a CCl_4-based plasma under the variation of chamber pressure, RF power, electrode spacing, and gas composition. The effects of network structure and FFEBP learning parameters, as mentioned above, were optimized by means of an efficient statistical design-of-experiment technique (i.e., D-optimal design).

Some commercial vendors, such as Verity Instruments, Inc., have developed neural-network-based tools for plasma etch endpoint detection, which at present is often done by operators. The endpoint is the point at which one would like to shut off the plasma when the etch of a layer is finished. It is known that if the shutoff time is not controlled well, the wafer may be underetched or overetched. Neither of these conditions is acceptable. Verity's tool has a graphical user interface (GUI) that selects normal samples for training on neural networks and then uses the network to detect the endpoint. Verity claims a 99.5% success rate.

Another successful commercial application of neural network technology is in detecting meaningful wafer bin patterns from electrical test parameter systems and defect databases, and then correlating these patterns with process equipment. NEDA of DYM Inc. features NeuralNet™, a custom-designed, class-sensitive neural network engine that learns a fabrication's specific bin patterns and then correlates similar patterns on production wafers with various in-fabrication processes, thereby suggesting a corrective course of action (18).

Other techniques, such as directed-graph classifier (19) or ID3 (20) (a class of classification algorithms), can be used to detect the malfunctioning wafer-test patterns. These sets of techniques classify the failure patterns for incoming wafers on a wafer map and look for similarity of patterns for diagnosing problems automatically. Using the processing history, a correlation can be established between the failure patterns and possible process/equipment faults.

Application of Fuzzy Logic in Equipment Diagnosis

Fuzzy logic is quite popular in diagnostics and control applications (21,22). In essence, fuzzy logic transforms a quantitative space into a qualitative one, which facilitates the fuzzy reasoning method. For example, a temperature between $-12°$ and $-1°C$ can be classified as "very low," with a membership function u assigned to the range. The rest of the assignments are shown in Fig. 4. With other relevant parameters similarly expressed in terms of fuzzy membership, a set of fuzzy rules (22) can then be applied to derive the outcome. This approach avoids defining infinite combinations of expert system rules in a quantitative space, thereby producing humanlike reasoning such as is used in parking a car (21).

In Ref. 23, a self-learning fuzzy logic system was developed for in situ and in-process diagnosis of a mass flow controller (MFC) that controlled the flow of gas into a process chamber. Mainly, the unacceptable drift in its calibration was diag-

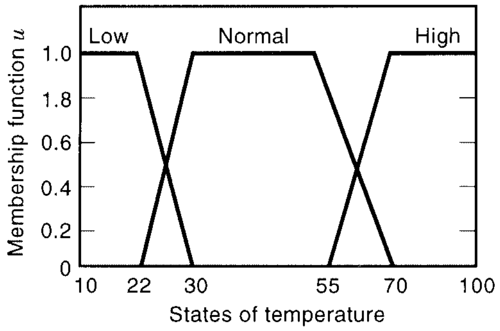

Figure 4. Description of temperature in a fuzzy membership mapping.

nosed, and accordingly, an automated calibration procedure was invoked before processing any wafer. Malfunctioning arising out of catastrophic failure was also addressed. Fuzzy logic was used in the diagnostic system to detect the problem while a self-learning system automatically built the knowledge base consisting of fuzzy rules used for diagnosis.

Detecting Faults using Real Time Statistical Process Control on Plasma Etching

Plasma etch is considered a very important step in integrated circuit (IC) processing because the density of the IC depends on how fine a pattern one can etch. Therefore, there has been a lot of interest in better control of plasma etch processing and diagnosis.

The recent trend in plasma etch diagnosis is to install more in situ sensors for monitoring. These may include optical emission spectrum (OES) tools for chemical process emission; quadrupole mass spectrum analyzers for gas analysis; residual gas analyzers for residual chemical component analysis; and monitoring for RF, pressure, temperature, flow, etc. So, for a typical plasma etcher, there can be from 15 to 25 waveforms (including spectra) one can monitor. The problem is how to determine faults by looking at these waveforms simultaneously in real time.

Traditional SPC is based on the application of either Shewhart or cumulative sum (CUSUM) charts. Shewhart charts can monitor large shifts efficiently, while CUSUM charts are more applicable when small continuous drifts are present (24,25). However, these techniques look at only one parameter at a time and also assume that the parameters are not cross-correlated and are independently and identically normally distributed (IIND). Under the IIND assumption, the arithmetic average can be shown to be distributed according to another known distribution given as

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{1}$$

where μ is the mean and σ is the standard deviation of one parameter. However, during real time monitoring of in situ parameters in plasma etcher, these parameters are typically cross-correlated and non-IIND. One effective approach to detecting faults by monitoring these parameters, called real time SPC, was developed at the University of California at Berkeley. They achieved great success in applying this method to plasma etcher (26,27). Their approach is shown in Fig. 5 and is described below.

During the rapid and continuous monitoring of in situ parameters, a problem often arises that each new value tends to be statistically related to previously measured values. The existence of autocorrelation in the controlled parameters violates one of the most basic assumptions underlying the design of standard SPC schemes, namely, that all samples are IIND random variables. In order to cope with this problem, the monitored parameter might be modeled by means of an appropriate time series model. Time series models, such as the well-known autoregressive integrated moving average (ARIMA), can be used to forecast each measurement and deduce the forecast error (28). This error can then be assumed to be an independently distributed random variable, and it can be used with traditional SPC schemes. The other problem during real time monitoring of multiple parameters with equipment such as plasma etcher is that these parameters (or residuals of these parameters after ARIMA modeling) are cross-correlated. If we look at a number of independent control charts of these parameters, the overall risk due to cross-correlation cannot be correctly evaluated.

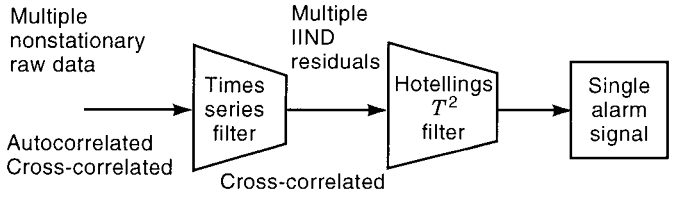

Figure 5. Summary of the real time SPC scheme.

A good multivariate scheme that alerts the operator to changes in the mean vector or the covariance matrix of a group of controlled parameters is the Hotelling's T^2 statistic. This statistic is sensitive to the collective deviations of a number of cross-correlated IIND parameters from their respective targets. In practice, the T^2 statistic presents a far clearer picture of the process status and is much less likely to introduce false alarms. Data streams include pressure (P), ratio (R), power (W), gap (G), total flow (T), and/or OES. This approach has successfully been applied in a couple of real world plasma etchers. The reason to use Hotelling's T^2 technique is that using SPC for each waveform may cause too many false alarms, which makes waveform correlation difficult. Hotelling's T^2 approach can be tuned to different sensitivities for each faulty waveform pattern to reduce false alarms. However, the problem is that Hotelling's T^2 can only signal the existence of a fault, not point to a specific cause. This is still an active research topic.

Factorywide Monitoring and Diagnosis

To make effective the diagnostic techniques discussed in the previous sections, there is a need for a distributed factory-wide equipment/process monitoring system that provides data collection, management, and analysis. The following functionalities have been identified for such a system (29):

- Automated in situ (i.e. sensor) data acquisition from process equipment in real time
- Real time distributed and remote data display, if desired
- Performance of SPC, real time SPC, or real time fault classification on the data
- Disabling of the machine upon alarm (alarm management)
- Data analysis and interpretation
- Central management of factory-wide process data
- Performance of arbitrary correlations across the process, such as correlation to test, WIP, or parametric data
- Display of real time data from real time database, with storage of essential information in relational database
- Building of causal models (such as FMEA) across the process based on the data
- Maintenance of 2000+ charts across a typical fabrication
- Keeping track of alarm explanations given by operators and engineers

- Provision of a versatile I/O-to-equipment interface, such as a scan-module-to-SECSII-protocol interface, which most process equipment has
- Ease of application of various diagnostic algorithms (diagnostic tool box) for monitoring the data in question
- Provision of feedforward and feedback control based on the real time monitoring (can be done after the above are done)

The functionality listed above is generally missing from the fabrication, but is becoming more and more important for rapid yield learning and process/equipment diagnosis. Process faults need to be detected as early as possible. The time of final testing is too late. In this way, the testing data can also be correlated to real time process sensor data. Current systems with some of the functionalities listed above include MonitorPRO 97 and ControlPRO 97 from Real-time Performance Inc., and a distributed measurement and control platform called Vantera from Hewlett-Packard.

MONITORING/DIAGNOSIS AT THE PROCESS FLOW LEVEL

While the current practice is to use equipment-level monitoring whenever possible for early process diagnosis, in-line and end-of-line physical and electrical measurements are being used within the process flow as monitors for process diagnosis.

Three computer systems that provide process diagnosis at the process flow level are described in detail below. The characteristics of these systems are:

- the use of technology computer-aided design (TCAD) process and device simulation tools in order to correlate process parameters and in-line physical and end-of-line electrical measurements (TCAD simulators provide the physical models that describe the process and devices)
- the use of expert systems, statistical methods, and/or neural network techniques to facilitate process diagnosis, especially when the physical simulators cannot fully model the actual process

With much information correlated at the process flow level, these systems are especially useful during early process yield learning, such as determining unstable unit processes or equipment. However, compared with equipment-level monitoring/diagnosis, significant time delays can occur if problems are left undetected and are not diagnosed until end-of-line measurements.

The Expert System Approach

Expert systems, also known as knowledge-based (KB) systems, were developed as an early attempt to provide diagnosis at the process flow level. Whereas the early systems usually employed more heuristic artificial intelligence (AI) qualitative techniques (30), later systems have integrated quantitive physics-based process/device models with qualitative knowledge. AESOP (31), developed at Stanford University, is such an integrated system. This system, which has been successfully transferred and deployed in an industrial environment, is described in the following sections.

AESOP System Overview. The goal of the AESOP system is to automatically diagnose process problems on the basis of the readily available end-of-line electrical test (e-test) data used extensively in semiconductor manufacturing. These e-test data contain measurements on specially designed electrical test structures.

The AESOP system is developed in three stages: knowledge generation, knowledge representation, and diagnostic reasoning. The development is done in a expert system engineering environment called HyperClass.

AESOP Knowledge Generation. AESOP's general methodology for generating a process diagnostic knowledge base has been applied to a 2-μm CMOS process at Stanford University. The prototype AESOP system restricts its knowledge base to basic transistor data only. The process variables used in the knowledge base are the effective channel length (L_{eff}), p-substrate concentration, n-well ion implant dose, and gate oxide thickness. Each of these four parameters is varied over a range of fixed values or percentages of its nominal value. The selected electrical measurements included in the knowledge base are extrapolated threshold voltage (V_{T}), maximum transconductance (G_{m}), saturation current (IDSAT), an intermediate 3 V gate voltage current (IDS35), and the subthreshold currents at a voltage of 0.4 V for both n and p long-channel and minimum-length transistors.

TCAD process and device simulation tools are used to capture the physical relationships between the process deviations and the resultant electrical measurements on the test structures. AESOP is one of the first diagnostic systems that generated its knowledge base using quantitative physical process and device simulators. The simulator SUPREM III is used for process simulation. SUPREM III is a one-dimensional (1-D) process simulator that provides the necessary doping concentration profiles for the device simulator PISCES IIB. PISCES IIB is a two-dimensional (2-D) device simulator developed for the selected e-test measurements described above.

Given the complexity of the process and device simulators, a very long simulation time is required to fully characterize the entire CMOS process flow. As a result, special analytical response surface models (RSMs), which correlate the process variables directly to the end-of-line electrical test measurements of the devices, are built. Statistical design of experiment (DOE) technique (32) is used to pick the simulation runs efficiently in order to construct these RSMs. A standard fractional factorial experimental matrix, which contains rows with the appropriate simulation levels for the input process variables, is generated. The experimental matrix is augmented with additional simulation levels at twice the variable settings, as well as settings that accommodate quadratic terms in the RSMs. For each row in the experimental matrix, a SUPREM process simulation run is performed. Four relevant regions (pn gates and source–drain regions) of the CMOS transistor are simulated by SUPREM. Simulated doping profiles based on the input process variables are then used by PISCES to compute the e-test measurements under the several bias conditions described above.

The resulting RSMs are quadratic regression models. These models provide the direct cause-and-effect relationships between the process variables and the final e-test measurements. In addition, these analytical models allow characterization of the process flow to be performed easily, due to the short computation time as compared with the full process/device simulators.

The last step in the knowledge generation methodology involves the use of RSMs to characterize the process excursions in the process flow. Given the format of the system knowledge representation, the quantitative results of running the RSMs have to convert into qualitative relationships. For example, a *low* value of oxidation time will give a *low* value of oxide thickness. To convert the quantitative relationships to the internal qualitative representations, various *very low, low, nominal, high,* and *very high* values are chosen appropriately for the input process variables. For example, a *high* description was 12.5% of the nominal implant dose, and a *very high* description is twice the *high* value. Similarly, windows are also chosen for the qualitative descriptors for the e-test measurements. Once the rules for the qualitative descriptors are determined for both the process variables and e-test measurements, Monte Carlo simulations are then performed, based on the RSMs. The causal relationships that represent the specific process flow is then constructed after converting the quantitative RSM simulation results into qualitative descriptors.

AESOP Knowledge Representation. Process diagnostic knowledge in AESOP is based upon the knowledge of a typical semiconductor process engineer. Many primitive units for two basic categories—fault concept and causal link—make up the AESOP knowledge base. These primitive units are modeled as software objects in the object-oriented (OO) environment HyperPIES. Each of these objects has attributes that describe the object properties.

Fault Concepts. In AESOP, a fault refers to an anomalous process condition induced by failures/problems within the semiconductor process. For example, the process fault of very thin gate oxide can be caused by a failure/problem in the oxidation step.

Faults are represented by a fault concept object in the AESOP OO environment. Faults are organized according to semiconductor knowledge into four related conceptual levels called causal levels:

- *Root Fault.* The root faults usually happen at the equipment level, human and/or environment levels.
- *Process Fault.* Root faults manifest themselves as process faults. One example of a process fault is an out-of-control oxidation temperature due to a furnace temperature control problem.
- *Physical Fault.* Process faults manifest themselves as physical faults on the wafer. For example, a high oxidation temperature results in a thick oxide.
- *Measurement Fault.* Physical faults on wafer device structures manifest themselves as test measurement faults, where the electrical or physical device measurements have abnormal results.

A full configuration of these causal levels is known as a fault taxonomy in AESOP, as shown in Fig. 6.

Causal Links. *Causality,* represented as *causal links,* describes the *cause-and-effect* relationships between the fault objects in the fault taxonomy. The casual links are mapped into *causal link objects* within the AESOP HyperPIES development environment.

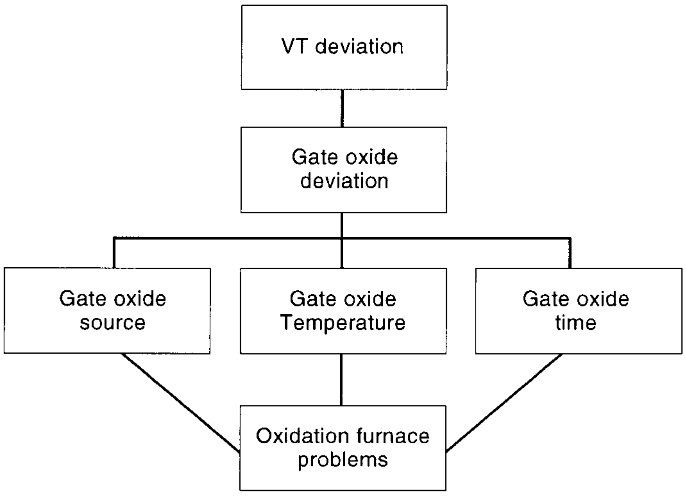

Figure 6. Fault taxonomy: fault concept and causal links.

A qualitative attribute, *association strength,* is used to "measure" the causal correlation between fault concepts. This attribute has five qualitative levels: *must, very likely, likely, probably, maybe.*

Two types of causal links are possible:

- *Interlevel* causal links represent causality between fault concepts in two consecutive causal levels. This type of causal link captures most of the relationships for process diagnosis. For example, a long oxidation time at the process fault level causes a thick gate oxide at the physical fault level.
- *Intralevel* causal links represent causality between fault concepts in the same causal levels. This type of causal link captures the more detailed relationships within the same causal level. For example, an *n*-well implant process fault can cause both low substrate concentration and high well concentration.

User Interface. In order to facilitate knowledge creation and maintenance by the process engineers themselves, easy-to-use user interfaces were developed for AESOP. Specifically, two kinds of knowledge editors are used:

1. *Fault Concept Editor.* This editor allows users to add or delete new fault concepts, as well as their respective cause-and-effect lists.
2. *Causal Link Editor.* This editor allows users to define the cause, causal level, effect, effect level, link type, association strength, etc. for the causal links.

AESOP Diagnostic Reasoning. AESOP uses a *backward chaining* (33) strategy for process diagnosis. The diagnosis starts from the measurement fault level and ends in the root fault level:

1. *Measurement Fault.* The user selects measurement data sets for analysis. Electrical measurement deviations serve as initial symptoms to infer device physical faults.

2. *Physical Fault.* Physical structure deviations are used to hypothesize potential process faults.
3. *Process Fault.* A set of likely process anomalies are then used to search for the root causes.
4. *Root Fault.* Root faults are then identified for the diagnosis.

A strategy of *hypothesis and verification* is used to isolate failures at each causal level:

1. At each causal level, a set of initial symptoms is identified, based on either the test data or the diagnostic reasoning from the previous causal level.
2. During the hypothesis phase, the candidates that possess the strongest correlations to the symptoms are identified.
3. During the verification phase, the candidates are then matched against the expected symptoms in the knowledge base.
4. The candidates are then sorted and clustered according to their matching scores. The cluster with the best matching score is then passed to the lower causal level. The reasoning process is repeated until the root cause for the symptom(s) is diagnosed.

Result and Summary. An expert system, AESOP, for semiconductor process manufacturing has been described. The system was developed in three stages: knowledge generation, knowledge representation, and diagnostic reasoning. The knowledge base was generated with the use of TCAD process/device simulators, whose simulation values were determined through the use of statistical design of experiment (DOE). Both qualitative and quantitative knowledge are represented in a fault taxonomy with causality links and fault concept objects. The diagnostic reasoning was done using backward chaining and the hypothesis-and-verification approach.

At the time of publication (31) AESOP was able to diagnose single-fault test cases successfully. These test cases were artificially generated using the RSMs described in the previous subsections. The AESOP system was later extended and deployed in a major semiconductor manufacturing company, where it successfully diagnosed real life manufacturing problems.

However, diagnosing multiple faults, e.g. when both gate oxide and channel length exceed their normal ranges, still presents a challenge to the AESOP system.

Statistical-Based Systems

With a better understanding of the process technologies, as well as improved TCAD numerical process/device simulators, quantitative-based process diagnosis systems were developed. These systems combined statistical techniques with information from the process/device models to perform the automated diagnosis. As a result, the approaches are more systematic and rigorous, with better diagnostic results, than the qualitative heuristic approaches used in AI expert systems.

A very successful statistical-based system (2) from Carnegie-Mellon University (CMU) is described below. This system was later commercialized as PDFAB, a product of PDF Solutions, which is now widely used in the industry.

System Overview. The statistically based process diagnosis system (2) described here was developed in the early 1990s at CMU by Kibarian et al. This system is novel in its combination of powerful statistical techniques and numerical simulators. Like AESOP, the system performs diagnosis at the process step level. The system was developed in the following stages:

- *Feature Selection.* Principal component analysis (PCA), a direct statistical method of data analysis, is used to extract specific features from the raw measurement data. This step also reduces the dimension of the measurement data. This step has two substeps: computation of eigenvalues and eigenvectors from the measurement correlation matrix, and dimension reduction.
- *Feature Interpretation.* The interpretation step consists of two major substeps:
  1. *Process Sensitivity Analysis.* TCAD simulators were used to provide the process sensitivity information, especially when historical data were lacking. The process sensitivity information is expressed as sensitivity vectors, which are used in the feature matching phase. This is similar to the AESOP knowledge representation stage.
  2. *Feature Matching.* The diagnostic reasoning stage is performed here. The system tries to match the selected features (eigenvectors) from the measurement data with the sensitivity vectors from the simulations. Once a match or multiple matches are found, the underlying process steps that caused the process problems can be identified. Using this technique, the CMU system can handle multiple faults.

A flowchart outlining the process diagnostic steps is shown in Fig. 7.

The CMU system is developed with the capability to diagnose intrawafer process problems, which are becoming dominant in those process technologies using larger wafer sizes.

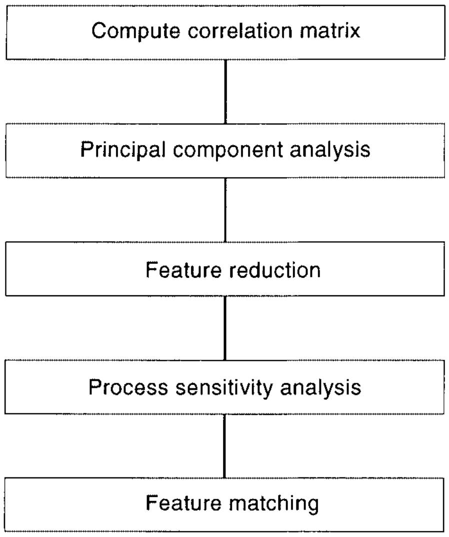
Compute correlation matrix
Principal component analysis
Feature reduction
Process sensitivity analysis
Feature matching

Figure 7. CMU process diagnostic flow.

Feature Selection. As many different electrical and physical measurements are taken from the devices on the processed wafers, the number of measurement data can become very large. In order to reduce the dimension of the data analysis, as well as to extract the salient features of the data set, the PCA technique is used in the CMU system.

The feature selection involves the following steps:

Computation of the Correlation Matrix. Many different physical and electrical measurements are taken for the device chips on silicon wafers. These measurements indicate the physical/electrical performance of the processed device chips on the silicon wafers. All these measurements are statistically correlated, since they all relate to the same common set of underlying process variables (or conditions). As a result, a correlation matrix can be computed from the measurement samples. In the CMU system, the matrix is computed using the maximum likelihood estimate of pairwise correlation.

Principal Component Analysis. Once the correlation matrix is computed, eigenvectors and their associated eigenvalues an be computed as follows:

$$R = A^{\mathrm{T}}LA$$

where R is the correlation matrix, A is the matrix of eigenvectors, and L is the diagonal matrix of eigenvalues. The PCA yields a set of independent latent variables described by the eigenvectors. These latent variables are functions of the original independent process variables. Note that PCA is a direct data analysis technique to which a priori knowledge of the process relationships is not required.

The associated eigenvalue for each eigenvector represents the sum of percentages for the variances of each kind of measurement (i.e., performance) that is accounted for by the corresponding eigenvector.

Feature Reduction. Once the independent set of eigenvectors and their values are determined, another filtering step is performed to yield the minimum set of data features to be used for process diagnosis. This minimum set of features is defined to be the set that contributes a certain amount (e.g., 95%) of the total variance of each performance measured. The problem of determining the minimum set can be cast as an integer linear programming problem:

$$\min_{\phi} \sum_{i=1}^{n} \phi_i$$

subject to $A\Phi L^{0.5} \geq \alpha$, where A is a matrix in which each row is a sample eigenvector; ϕ_i has the value of 1 if the ith feature is significant, 0 otherwise; Φ is a diagonal matrix with the ϕ_i on the diagonal; $L^{0.5}$ is a vector of the square roots of the sample eigenvalues; and α is the desired percentage of the total variance contributed by the minimum set.

Feature Interpretation. Once the important features are extracted from the measurements, the next step is to interpret the extracted features with respect to the specific process technology used to manufacture the devices. The two steps for feature interpretation are:

Process Sensitivity Analysis. Before the extracted measurement feature can be interpreted, a knowledge base of the specific manufacturing process technology must be constructed. As in the case of AESOP, such a knowledge base is constructed with TCAD numerical process/device simulators. The TCAD simulators provide a mapping from the settings of process variables (also known as disturbances) to the device measurements (also known as performances):

$$\boldsymbol{P} = \boldsymbol{F}(\boldsymbol{D})$$

where $\boldsymbol{P}$ is the vector of measured performances, $\boldsymbol{F}(\ )$ is the vector function that maps process variables to measured performances, and $\boldsymbol{D}$ is the vector of process variables.

From this relationship, the means, variances, covariances, and correlation matrix of the measured performances can be computed. Since the relationship above is not in closed analytical form, numerical simulations are needed to perturb the process variables; then the sensitivities of the device performances are estimated with respect to the process variables.

Based on further derivations using Taylor series approximations, the correlation matrix R_{p} can be computed using the simulation data:

$$R_{\mathrm{p}} = LPS^{\mathrm{T}}BD^{-1}\Sigma_{\mathrm{D}}D^{-1}BSPL$$

where L is the diagonal scaling matrix with $1/\sigma_i$ on the diagonal, in which σ_i is the standard deviation of the measured performance i; P is the matrix with the nominal values of the performances on the diagonal; S is a sensitivity matrix scaled so that the rows are of unit length; B is a diagonal matrix in which the ith diagonal element equals $\sqrt{\sum_{j=1}^{m} J_{ij}^2}$, where J is the Jacobian matrix evaluated at the nominal values of the process variables; D is the diagonal matrix with the nominal values of the process variables on the diagonal; and Σ_D is the diagonal matrix of the process-variable variances.

Note that the correlation matrix derived has a structure similar to the R_{p} derived from the PCA. Specifically, the sensitivity matrix S is scaled so that the rows are of unit length. Based on this special property of the sensitivity matrix, the process/device simulators do not have to be tuned to have the exact same variances as the actual manufacturing process. In addition, the matrix S can be used as a matching target against the eigenvector matrix derived from the measurement data PCA. In essence, the knowledge required for process diagnosis is encoded in this matrix S.

Feature Matching. As described above, the correlation matrix R_{p} can be derived both from the sample measured performance through the PCA and from the linearized model that maps the process variables (disturbances) to the device performances. Based on further matrix manipulations, the linearly independent features represented by the eigenvectors can be extracted from the PCA. Likewise, the linearly independent sensitivity vectors representing the effects of process disturbances on the device performances can be extracted from the process sensitivity analysis.

A feature is characterized by its eigenvectors and eigenvalues:

$$f_i : l_i, e_i$$

where f_i is the ith feature, l_i are its eigenvalues, and e_i are its eigenvectors. A disturbance is characterized by its sensitivity vector:

$$d_j : n_j$$

where d_j is the jth disturbance, and n_j is the sensitivity vector of the performances to the jth disturbance.

Given that the features are linearly independent, their eigenvectors can be matched one at a time against the unit-length sensitivity vectors. The matching is done by taking the appropriate inner product between the feature and sensitivity vectors. A match occurs when the inner product is larger than a preset number, as determined by the hypothesis-testing confidence interval. Once a match occurs between a feature and a sensitivity vector representing the particular process variable or disturbance, the process variable or disturbance is a possible explanation for the process variation feature.

Result and Summary. A statistically based process diagnosis system from CMU has been described. The diagnosis was done in two stages—feature selection and feature interpretation—using powerful statistical techniques and process/device simulators. Feature selection was done using PCA to extract specific features from the raw measurement data and to reduce the dimension of the measurement data. Feature interpretation was done by performing process sensitivity analysis and feature matching. TCAD simulators were used to provide the process sensitivity information in process sensitivity analysis. In the final feature matching, the system tried to match the selected features (eigenvectors) from the measurement data with the sensitivity vectors from process sensitivity analysis. Once a match or matches were found, the underlying process steps that caused the process problems could be identified. Using this technique, the CMU system can handle multiple faults.

The system was developed to handle intrawafer process variations as described in Ref. 2. The system has been commercialized (8) and deployed in actual manufacturing environments. The commercial system is being enhanced to integrate qualitative knowledge that is not available in TCAD simulators.

Neural-Network-based System

The last system described here is a process diagnosis system based on a neural network (NN) (34). Its system function is very similar to that of the two systems studied earlier in that the process diagnosis is based on the electrical/physical test measurements. One unique feature of this system is the use of a backpropagation NN model to represent the process knowledge. All three systems use the TCAD simulators as virtual processes to generate the required knowledge base and training data.

System Overview. The NN system data flow, which outlines the relationships between different simulation/measurement data and the different systems, is shown in Fig. 8. The NN process diagnosis system uses TCAD process/device simulators to generate the simulated physical/electrical test measurements from the input process variables (disturbances). Then the same set of process disturbance and test measurement data is used to train the NN model, which represents the process knowledge. During training the data switch roles: the simulated output measurement data become the input to the NN model, and the simulation input process disturbance data become the output from the NN model.

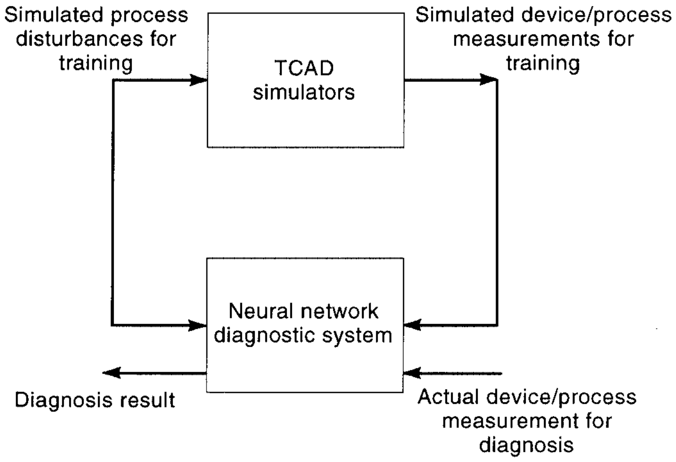
Simulated process disturbances for training
TCAD simulators
Simulated device/process measurements for training
Neural network diagnostic system
Diagnosis result
Actual device/process measurement for diagnosis

Figure 8. Neural network system data flow.

As in AESOP, a special statistical sampling technique is used to generate the process disturbance values for use as the simulation inputs. This technique is described in the next sub-subsection. Once the input samples are generated, the output measurement data are simulated by the TCAD simulators. The data are then fed into a backpropagation NN model for training.

Prior to being input to the NN, however, the data are digitized into special patterns using a special coding technique and a thresholding algorithm. The coding technique speeds up the NN training. In addition, a special fault observability algorithm is developed to select the appropriate measurement data in order to maximize fault observability. After these manipulations, the process knowledge is now represented in the NN model.

Once the NN is trained, the actual measurement data can be fed into the NN in search of faults in the process disturbance. The number of NN input nodes depends on the number of measurements. The number of NN output nodes depends on the number of process disturbances. There are several middle layers within the NN model. Once the NN is trained, diagnosis can be performed in real time.

Training-Data Generation

Fault Observability. Before any training data can be generated, it is necessary to determine a sufficient set of measurements that will make as many process disturbances distinguishable and observable as possible during the diagnosis process. The selection of this measurement set can be accomplished with a novel fault observability algorithm, which has the following steps:

1. *Generate a Fault Matrix.* The fault matrix is an $n \times m$ encoded matrix where each element f_{ij} represents the effect of an out-of-control process disturbance d_i on a measurement y_j:

$$f_{ij} = \begin{cases} 1, & s_{ij} > T_j \\ -1, & s_{ij} < -T_j \\ 0, & |s_{ij}| < T_j \end{cases}$$

where s_{ij} is the shift from measurement y_j's nominal value resulting from the shift from process disturbance d_i's nominal value (e.g. 3σ), and T_j is a prescribed threshold value for each measurement y_j. The computation of T_j will be described later.

2. *Check Fault Observability.* For a given set of measurements, check that each process disturbance $d_1, d_2, \ldots, d_n$ is observable. If not, try to eliminate the nonobservable process disturbance from the set or add appropriate measurements to make the process disturbance observable. A process disturbance d_i is unobservable if

$$\sum_{j=1}^{m} |f_{ij}| = 0$$

3. *Check Fault Ambiguity.* Faults d_i and d_j are not guaranteed to be uniquely diagnosed if they are in the same ambiquity group. Two faults are in the same ambiquity group when $r_{ij} = 0$ under the two conditions described below. First,

$$r_{ij} = \sum_{k=1}^{m} |f_{ik} - f_{jk}|$$

The above condition checks if in the fault matrix two rows i and j are identical. Another condition is satisfied when the two faults have large and opposite effects on the measurements:

$$r_{ij} = \sum_{k=1}^{m} |f_{ik} - (-1)f_{jk}|$$

Sampling Strategy. Once the sets of measurements and process variables are determined from the fault observability algorithm, the NN system uses the Latin hypersquare (LHS) sampling technique to generate the simulation input data. The simulation input data consist of process variables/disturbances sampled at appropriate values. TCAD simulators are then used to generate the simulation outputs, which consist of the device physical/electrical performance data. This set of input–output data is then used to train the NN model.

LHS sampling is a stratified sampling technique such that given a sampling space S of the set of random variables X,

- S can be partitioned into independent disjoint strata S_i;
- n_i random samples can be selected from each stratum S_i.

The sum of the samples taken from all the strata equals the final desired sample size N.

For a set of random variables $X = \{X_1, \ldots, X_k, \ldots, X_K\}$, it is possible to assign a stratum probability distribution function for a random variable X_k with N strata. For example, in order to ensure that X_k has values sampled uniformly across the range of all its values, it is reasonable to assign the probability $1/N$ to all its N strata. With appropriate stratum distribution functions assigned to the random variables in X, samples are then selected from each stratum, and then matched in a random fashion to form the final sampled set that consists of all the random variables.

In the NN system, the distribution functions for the input process variables are usually selected with high probabilities at the 3σ control limits. This kind of distribution function will ensure that more samples are selected near the out-of-control limits. The definition of the out-of-control limits is discussed in the next section.

Neural Network Model Representation. Before the sets of simulation input–output data, which represent the values of process variables and test measurements, can be incorporated into the neural network for training, these data sets are digitized by thresholding and coding techniques to facilitate the NN model training and construction.

Thresholding. Threshold levels are assigned to both the input process disturbances and the output test measurements. The threshold levels for the input process variables are the control limits in the process control charts (32). These control limits represent the mean value and the $\pm 3\sigma$ values of each process variable. For the output test measurements, the threshold levels are determined by the combined effects of the significant process variables. The test measurement threshold levels may or may not align with the control limits (the mean and $\pm 3\sigma$ values). Figure 9 shows the definitions and relationships of the threshold levels.

Coding. Once the threshold levels are determined, the training sets of inputs and outputs can be encoded according to the defined threshold levels. (See Fig. 10.) The code a_1 is assigned to disturbance 1, since its value falls in the range of a_1 (above the $+3\sigma$ range), while the code a_2 is assigned to disturbance 2, since its value falls in the range of a_2 (below the -3σ range). For output measurements, output 1 is encoded with c_{mn}, since its value is between thresholds t_m and t_n, while output 2 is encoded with c_{ij}, since its value falls between thresholds t_i and t_j. The unique input–output encoding pattern for the inputs and outputs can be generated as

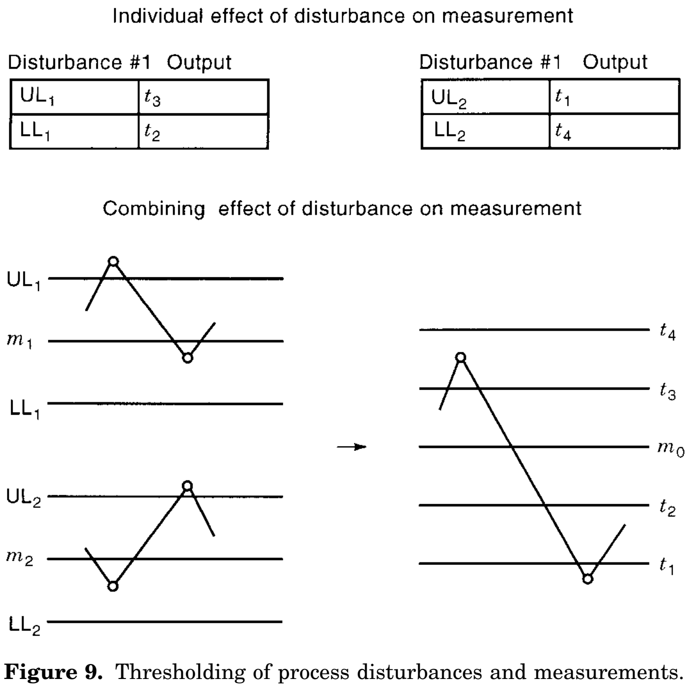

Figure 9. Thresholding of process disturbances and measurements.

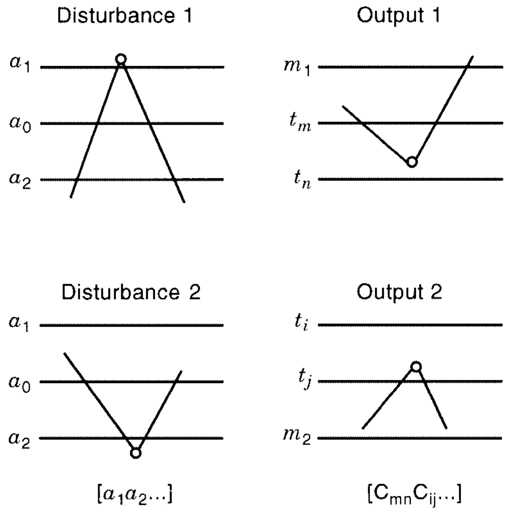

Figure 10. Encoding of process disturbances and measurements.

$[a_1a_2. . .]$ and $[c_{mn}c_{ij}. . .]$. Using encoding, the NN training is based on the digitized data instead of the actual numerical data and can thus be performed more effectively.

Diagnostic Reasoning. The NN model is trained with the encoded TCAD simulation input–output data. This training, however, is performed in an order that is the reverse of the prior simulation effort. The simulation output measurements serve as inputs to the NN model, while the simulation input process variables serve as the outputs. Once the NN model is trained and constructed, test measurements on the process can be fed as inputs to the NN, and the possible process faults are diagnosed as outputs accordingly. In contrast to AESOP's backward chaining the CMU's feature matching, the NN system relies on the instrinsic NN model structure for computing the diagnostic result, and therefore does not require a separate diagnostic reasoning engine.

Result and Summary. A neural-network-based system has been described. The knowledge generation process for the NN using TCAD process and device simulators was overviewed. A special statistical sampling technique, Latin hypersquare, was used to generate the values for the training data. A special fault observability function was used to determine a sufficient set of measurements that would make as many process disturbances distinguishable and observable as possible during the diagnostic process. The training data was effectively represented in the NN using special coding and thresholding techniques.

Testing of the system diagnostic capability was entirely simulation-based, where both the training and diagnostic data were all generated from virtual experiments. The system was tested on a simple simulated CMOS process, for which the PMOS device measurements were extracted. The NN model contains 24 input nodes, 10 output nodes, and 35 hidden units. The validation experiment produced good results. It is unknown whether this NN system was actually deployed in an actual manufacturing environment. Due to the simplicity of the approach, the NN system appears promising.

Testing and Fault Diagnosis

The systems described perform process diagnosis for the parametric deviations of the input process variables, based on mostly electrical test measurements. there is another type of system that handles diagnosis of castatrophic faults that are mostly due to particle contamination. Such systems construct relational mappings between process particle contamination, circuit physical layouts, and final electrical test measurements. The reader is referred to the Carafe (5) and CODEF (35) systems.

Future Directions

Process level diagnosis systems are moving from the use of qualitative to quantitative techniques, such as statistical, NN, and numerical TCAD simulators. Commercial systems, such as the one described in Ref. 8, have demonstrated the use of both qualitative and quantitative techniques for process diagnosis.

Equipment and unit-level process diagnosis systems are gaining importance, as they can diagnose problems much earlier than process-level diagnosis systems.

BIBLIOGRAPHY

1. R. McIvor et al., Profiting from process improvement in the new semiconductor manufacturing environment, *Technol. and Oper. Rev.,* December 1997.
2. J. Kabarian, Statistical diagnosis of IC process faults, Ph.D. Dissertation, Electrical Engineering and Computer Science, Carnegie Mellon University, 1990.
3. S. P. Cunningham, C. J. Spanos, and K. Voros, Semiconductor yield improvement: Results and best practices, *IEEE Trans. Semicond. Manuf.,* **8**: 103–109, 1995.
4. H. T. Heineken, J. Khare, and W. Maly, Yield loss forecasting in the early phases of the VLSI design process, presented at IEEE 1996 Custom Integrated Circuits Conference.
5. A. Jee and F. J. Ferguson, Carafe: An inductive fault analysis tool for CMOS VLSI circuits, presented at 11th Annu. 1993 IEEE Test Symp.
6. W. Maly, Cost of silicon viewed from design perspective, in *Proc. 31st ACM/IEEE Design Autom. Conf.* June 1994, pp. 135–142.
7. B. E. Stine, D. S. Boning, and J. E. Chung, Analysis and decomposition of spatial variation in integrated circuit processes and devices, *IEEE Trans. Semicond. Manuf.,* **10**: 24–41, 1997.
8. PDF Solutions Inc., PDFAB diagnosis module, Application Note, Winter 1994.
9. C. Yu et al., Use of short-loop electrical measurements for yield improvement, *IEEE Trans. Semicond. Manuf.,* **8**: 150–159, 1995.
10. N. Chang, Monitoring, maintenance and diagnosis in a computer-integrated environment for semiconductor manufacturing, Ph.D. Dissertation. Electrical Engineering and Computer Sciences, University of California, Berkeley, 1990.
11. N. Chang and C. Spanos, Continuous equipment diagnosis using evidence integration: An LPCVD application, *IEEE Trans. Semicond. Manuf.,* **4**: 43–51, 1990.
12. K. K. Lin and C. Spanos, Statistical equipment modeling for VLSI manufacturing: An application for LPCVD, *IEEE Trans. Semicond. Manuf.,* **3**: 216–229, 1990.
13. G. S. May and C. J. Spanos, Automated malfunction diagnosis of semiconductor fabrication equipment: A plasma etch application, *IEEE Trans. Semicond. Manuf.,* **6**: 28–40, 1993.

14. F. Nadi, A. Agogino, and D. Hodges, Use of influence diagrams and neural networks in modeling semiconductor manufacturing processes, *IEEE Trans. Semicond. Manuf.,* **4**: 52–58, 1991.
15. B. Kim and G. S. May, An optimal neural network process model for plasma etching, *IEEE Trans. Semicond. Manuf.,* **7**: 12–21, 1994.
16. C. D. Himmel and G. S. May, Advantages of plasma etch modeling using neural networks over statistical techniques, *IEEE Trans. Semicond. Manuf.,* **6**: 103–111, 1993.
17. M. T. Mocella, J. A. Bondur, and T. R. Turner, Etch process characterization using neural network methodology: A case study, *Proc. SPIE,* 1994.
18. DYM Inc., private communication, July 1997.
19. M. W. Cresswell et al., A directed-graph classifier of semiconductor wafer-test patterns, *IEEE Trans. Semicond. Manuf.,* **5**: 255–263, 1992.
20. M. E. Zaghloul et al., A machine-learning classification approach for IC manufacturing control based on test-structure measurements, *IEEE Trans. Semicond. Manuf.,* **2**: 47–53, 1989.
21. T. Hosaka, S. Arai, and H. Matsui, Vehicle control system and method, U.S. Patent No. 4,809,175, February 1989.
22. L. A. Zadeh, The role of fuzzy logic in the management of uncertainty in expert systems, *Fuzzy Sets and Syst.,* **11**: 1983.
23. R. Ramamurthi, Self-learning fuzzy logic system for *in situ,* in-process diagnostics of mass flow controller (MFC), *IEEE Trans. Semicond. Manuf.,* **7**: 42–52, 1994.
24. D. C. Montgomery, *Introduction to Statistical Quality Control,* New York: Wiley, 1985.
25. J. M. Lucas, Combined Shewhart–CUSUM quality control schemes, *J. Quality Technol.* **14** (2), 1982.
26. C. Spanos, Statistical process control in semiconductor manufacturing, *Proc. IEEE,* **80**: 819–830, 1992.
27. C. J. Spanos et al., Real-time statistical process control using tool data, *IEEE Trans. Semicond. Manuf.,* **5**: 308–318, 1992.
28. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control,* 2nd ed., San Francisco: Holden-Day, 1976.
29. C. J. Spanos, private communication.
30. J. Y.-C. Pan and J. M. Tenenbaum, PIES: An engineer's do-it-yourself knowledge system for interpretation of parametric test data, *AI Magazine,* **7** (4): 62–69, 1986.
31. J. Patrick Dishaw and J. Y.-C. Pan, AESOP: A simulation-based knowledge system for CMOS process diagnosis, *IEEE Trans. Semicond. Manuf.,* **2**: 94–103, 1989.
32. G. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters,* New York: Wiley, 1978.
33. F. Hayers-Roth, D. A. Waterman, and D. Lenat, *Building Expert Systems,* Reading, MA: Addison-Wesley, 1983.
34. W. Zhang and L. Milor, A neural network based approach for surveillance and diagnosis of statistical parameters in IC manufacturing process, in *IEEE/SEMI Int. Semicond. Manuf. Sci. Symp.,* 1993, pp. 115–125.
35. J. B. Khare and W. Maly, Yield-oriented computer-aided defect diagnosis, *IEEE Trans. Semicond. Manuf.,* **8**: 195–206, 1995.

NORMAN CHANG
Hewlett-Packard Laboratories

KUANG-KUO LIN
Intel Corporation

DIAGNOSTIC EXPERT SYSTEM.

See COMPUTERIZED MONITORING.