From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

Neural Networks in CMOS Manufacturing: Some Examples

Edward A. Rietman

Bell Laboratories

600 Mountain Ave., Murray Hill, NJ 07974, ear@allwise.att.com

Abstract

The focus of this paper will be on two neural network models for plasma aided CMOS manufacturing. Both models were developed with strict statistical cross-validation and applied to real world applications. A plasma neural network gate etch controller has shown a 20% improvement in throughput in wafer processing by eliminating a set-up step, and has reduced the variance of thickness of an etched film by 40%.

In a second example, a multistep system model was built that enabled us to Pareto rank the various processing steps and their impact on a yield metric. A similar model allows us to predict yield for the multistep process prior to completion of manufacturing. With these large scale system models we can essentially do feedback and feedforward control of a manufacturing line with engineers in the control loop. In summary we show the results for an on-line neural network controller and we show results for decision support tools for the engineering staff.

1. Introduction

One of the objectives of modeling a process is to learn something about the process or to make real world predictions based on the behavior of the model. The traditional methods of statistics, expert systems and first principle models are excellent for specific tasks, but they are limited.

Statistical models are not capable of dealing effectively with more than a few variables (cf. Jenkins et al., 1986) and these are usually dealt with, by some linear assumptions. For a complex system with dozens of variables all coupled to each other the statistical approach fails. Expert systems are excellent advisors if the system is not too complex. There is no fundamental impediment to the application of expert systems in all fields. There is, however, a bottleneck to their employment. Transferring the experts' knowledge into the computer code can be extremely difficult and expert system models built from a knowledge base often are too brittle because of over-specialization of the rules generated by induction tree pruning. Dynamic models based on differential/difference equations would be very difficult for large nonlinear systems with many variables. Like the statistical models, the difficulty lies in the linearity assumptions. Furthermore, the calculus is often inappropriate because of boundary condition assumptions and infinities. First principle models require very detailed knowledge of the chemistry, physics and the dynamics to generate a model. These conditions are rarely met for manufacturing processes. Nonlinear neural network models are being increasingly utilized as complex system models. This nonlinear modeling technique is an excellent tool to allow us to circumvent some of the problems of conventional modeling and the difficulties of reintegrating these systems.

Although neural networks are quite useful for complex system modeling and classification, they do have their limits. For example, neural network models do not provide any insight into the physics of the phenomena. Specifically, it would be difficult to deduce subsystem dynamics without building separate subsystem models. Also, like all modeling techniques, including analytical models, when they are expected to perform beyond their bounds, the results may be questionable.

Typical process control, as practiced in integrated circuit manufacturing, consists of open-loop control with statistical analysis to observe changes in some process variable or process signature compared with a mean and variance from a short historical record. This approach to process control does not allow wafer-to-wafer control of the type needed for developing submicron structures with a good yield at the end of the process. For a high yield, some wafer attribute should be correlated to a control variable(s) or an *in situ* process signature for real time wafer-to-wafer feedback control.

2. Introduction to Plasma Processing

Plasmas are ionized gases used for chemical etching and deposition of films on silicon wafers. In the case of plasma deposition, the molecular species in the plasma is ionized (not 100%). As a result of this ionization "free electrons" in the plasma increase the effective temperature of the gas, and the plasma is hotter than the substrate. The plasma, in affect, provides a thermodynamic advantage by keeping the reactant molecules at a high temperature and the

Copyright © 1998, American Association for Artificial Intelligence.
All rights reserved.

From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

substrate at a much lower temperature.

Plasmas are also used for etching. In this case the etching gas is ionized. The plasma induces a dc bias (or one can be added) to the substrate. This bias, on the order of a couple of hundred volts, will cause the ions in the plasma to accelerate in a perpendicular direction (with respect to the substrate). This perpendicular acceleration results in the etching gas, etching in anisotropic directions. This anisotropic etching is one of the key processing techniques enabling the manufacture of submicron devices. Without anisotropic etching we would not be able to manufacture devices with high aspect ratio and small line width geometries. An extensive review of plasma physics and chemistry models is beyond the scope of the present work. The subject has been reviewed by Chapman (1980), van Roosmalen et al. (1991), and Manos and Flamm (1989).

3. Neural Network Computing

Neural networks are massively parallel networks of automata interconnected by adaptive connections. These adaptive connections, in conjunction with the nonlinear nodes, allow one to construct arbitrary adaptive nonlinear computing networks (c.f. Hornik et al., 1990). These networks have been well described in the literature (c.f. Rumelhart et al., 1986, and Weiss and Kulikowski, 1991). They have also been called parallel distributed processing networks, connectionist networks and neural networks. Early work on this computing paradigm was done by McCulloch and Pitts (1943), Woodrow and Hoff (1960, 1962) and Minsky and Papert, (1969), among others.

4. Example 1: Gate Etch Control

Neural networks have been described as universal mappers. They can perform any mapping that is not a random process (cf. Hornik et al. 1990). It is clear that such a relation exists between the remaining oxide thickness in the source and drain regions, of transistors on CMOS chips and some process signatures. For example, by monitoring one wavelength in the optical emission spectrum (OES), while the etch is in progress, it is possible to collect an emission trace of the process. This emission trace has embedded in it information about the plasma, the effect of the plasma on the wafer, etch rate, uniformity of the etch, and for example, properties of the masking pattern density (i.e. wafer attributes). If an autozero step occurs to correct for fogging of the optical windows, than the emission trace also has embedded in it the history of the cleaning of the reaction chamber or optical emission system drift.

The optical emission trace, from the plasma glow, is an excellent monitor for our controller. Many plasma etch processes use an emission trace to determine the end point which usually means the clearing of a film in exposed regions. Our gate etch process consists of three steps. The first step in the etch process is a timed etch where the optical emission is simply monitored. The second step is a slower etch chemistry with greater selectivity and the endpoint is called when the OES signal reaches a threshold. In the usual procedure the over etch process is a timed process with the operator selecting the over etch time on a cassette-by-cassette (batch) basis. This time estimate is based on inspection and measurements of the previous cassette. The adaptive controller also uses the same type of information from the previous cassette.

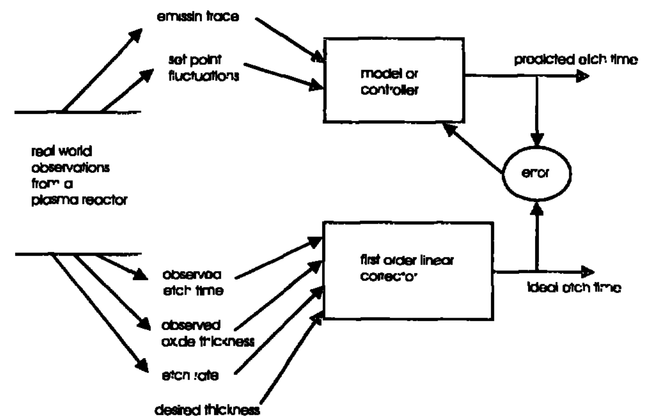

Figure 1 Block diagram to show how to train the neural network controller.

Our controller computes in real-time the ideal over etch time on a wafer-by-wafer basis, to give a desired remaining oxide thickness between the gate and the source (or drain). The controller computes this time based on information from the previous cassette and current, *in situ*, process signatures. Since there is a target oxide thickness, the controller must be developed with that information somehow embedded in its connection matrices. Figure 1 is a block diagram of the method we used to develop the connection matrices for the (37-6-1) neural network. In the nontraining phase the first order linear corrector and error correction are not present. The error correction algorithm was the back propagation of error method discussed in the literature, and consists of a gradient descent minimization of the error from the neural network. During this minimization process we are mapping a vector of inputs (process signatures) to the ideal etch time for a desired oxide thickness. The ideal etch time is computed from a simple linear corrector by the equation.

From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

$$t_{ideal} = t_{obs} + \frac{(T_{obs} + T_{des})}{E_{rate}} \quad \text{(EQ 1)}$$

Where t is the time, T is the thickness, *obs* is observed value, *des* is the desired value and E is the etch rate (cf. Rietman and Patel, 1994). In all of the examples in the database, the observed etch time is close to the ideal one, and the first-order linear corrections are small. It is important to keep in mind, however, that this kind of linear corrector is only valid for making these small corrections near the end of the etch process. The earlier steps in the etch sequence, which encounter different layers and transient effects at turn-on time, are highly nonlinear and less well behaved. These steps are accounted for in the actual production database.

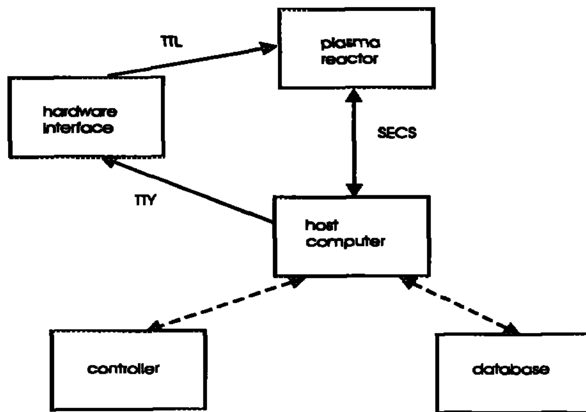

Figure 2. High-level block diagram of the entire system for controlling the plasma reactor.

4.1 High Level Description of the Controller

Figure 2 is a block diagram of a high level description of the controller and how it interacts with the on-line production database and the plasma etch reactor. The reactor is a Drytek (now Lam) Quad reactor. and is interfaced to a SUN workstation through the Semiconductor Equipment Communication Standard (SECS) interface. This interface allows communication in both directions. Recipes are down loaded from the host computer at the beginning of the cassette of wafers and processing conditions and process signatures are up loaded during the processing of the wafers. The adaptive controller and the database reside on the same host computer and are shown in the figure as separate logical boxes with virtual links. The hardware interface box reads a TTY line from the controller. A TTL signal is then used to signal the end of the etch. By using this approach there was very little hardware modifications to the production etch reactor and no modifications to the reactors' own software.

4.2 Performance of the Controller

In training an automata network controller, as in training a statistical learning tool, the data set is partitioned into two sections. One section is used for training and the other is used for evaluation and performance studies. We collected 1600 input/output tuples from a production database. This data set consisted of production wafers with pattern density that varied by as much as 50%. The pattern density, is of course, embedded in the optical emission trace and one of the actual inputs to the network was the number of die per wafer (a weak analog of the pattern density).

This large data set was randomized and partitioned into subsets for training and testing. The training consisted of repeated random selection of examples from the training set and back-propagating the resulting errors to update the connection matrices. After the error converged the training was stopped and the performance of the network was evaluated with the test set. As the training set size increases, the standard deviation of the network error decreases. At some asymptotic limit the network will be the optimum network with a minimum complexity.

The network that was put on-line was optimized for minimum complexity and low error. We allowed it to improve further in the production environment where it achieved better than human equivalence (40% improvement in the standard deviation). A fill paper describing the controller is given by Rietman and Patel (1994).

5. Example 2: Large System Model

One of the main objectives of statistical process control is to provide feedback control with humans in the loop. This *a posteriori* approach compares the end result of the process with the target specifications. By applying the rules of statistical process control (SPC) the engineers determine if a process is within specifications and adjusts the control parameters accordingly (cf. Doty, 1991). The feedback control effected is only as good as the statistical metrics and the engineering staffs' interpretation of these metrics. The major disadvantage of this approach is that the corrective action is taken after processing. Thus, several batches may have already been processed by the time the corrective action has been taken. We can, at least, improve the decision making capability of the engineers by providing them with machine learning programs that flag out-of-bounds conditions quickly or provide improved SPC metrics thereby improving the feedback control. Moreover, with machine learning programs used for yield prediction we can provide

From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

some level of feedforward control.

A few papers have appeared in the literature focusing on machine learning and SPC (cf. Beneke et al., 1988; Guo and Dooley, 1992; Smith, 1994; Hwang and Hubele, 1991; and Hwarng and Hubele, 1993). One of the best papers is by Huarng (1995) discussing a system of several parallel neural networks for detecting cyclic data in SPC control charts. Turner (1991) has observed that there is a correlation between real-time monitored process parameters, such as current and voltage in a plasma reactor, and so called "wafer-results" such as etch rate and etch uniformity.

Boskin et al. (1994) have used regression models to predict IC performance results prior to completion of manufacturing. Data from electrical test measurements between manufacturing steps were used to predict the final electrical performance of the IC. Our work is similar to Bosikn et al. except that we are predicting a "yield metric" (not IC electrical performance) from oxide film thickness and variance measurements (not electrical measurements).

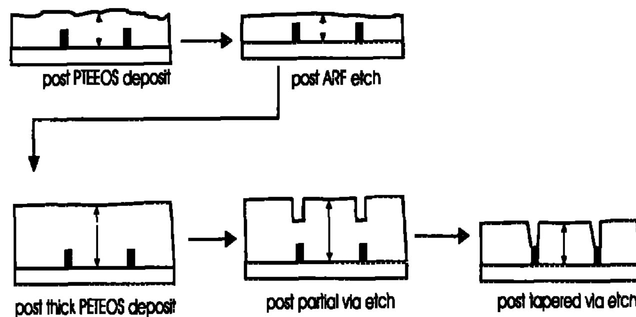

Figure 3. Schematic of the structures after each of the plasma processing steps. The arrows represent the point where film thickness measurements are made after each process. The black rectangles represent the cross sections for the metal runners on the wafer. Metal will be sputtered into the via (the contact holes), to make contact with these metal runners, in a later process.

5.1 Model Description and Results

In this paper a small factory within a larger factory is a case study. The mini-factory in this example manufactures vias. Advanced via etch processes are one of the enabling technologies to increase the packing density as device sizes drop to submicron ranges. In manufacturing integrated circuits the metal layers between the dielectric layers must be connected. The holes through the dielectric layers are known as vias. These allow metal contacts between the conductive layers and thus connect the various subcomponent levels (cf. Wolf, 1990). In the case study we assume that wafers in need of vias come into our via mini-factory, are processed and leave the mini-factory with vias.

Figure 3 shows a schematic of the structures being formed on the wafer for the five plasma processing steps. The processing steps are in the correct sequence. The first process is a PETEOS oxide deposition. Due to underlaying topography the top surface of this oxide layer will be somewhat rough and need to be leveled out. Newer generations of technologies are manufactured by chemical mechanical polishing (CMP) at this stage. The older generation of technology discussed here (0.9 micron) used an argon etch-back for planarization. Once the surface is smooth a thick layer of oxide is deposited and the vias are formed in a two different reactors optimized for the processes. When wafers leave the via mini-fab they are ready for metal sputtering.

After processing at each step in the mini-fab the film thickness is measured as indicated in Figure 3. These numbers are then passed to a statistical process control (SPC) database. We show that SPC data act as ill-posed pointers to allow one to model processes. Using 2 years of data (over 3000 lots) from a production SPC database we built an integrated model of the processing steps involved in manufacturing of vias for 0.9 micron ASIC technology. This integrated study includes the five key plasma processes (etching and deposition) that culminate in the production of vias on CMOS wafers. We found that there are essentially no linear cross-correlations between the processing steps and there are no linear correlations between the individual process steps (film thicknesses) and the yield for vias, as measured by the resistance between metal-one (M1) and metal-two (M2). Our models were reactor independent and product code independent.

A whole system model was assembled with a neural network. The goal of the learning machine was to predict the M1M2 resistance one wafer-lot in advance. Keeping in mind that there are five discrete processing steps, the numerical value of M1M2 resistance can be predicted ahead of time, before completion of all five processes. This prediction can be done to an accuracy of about 1 Ohm. For comparison, a Gaussian distribution prediction would give an error of 4.2 Ohms.

A second whole system model enabled us to conduct a Pareto analysis (Figure 4) and rank the processing steps in order of their impact on the M1M2 resistance metric. Using a neural network we demonstrate that the key processing steps, to determine the M1M2 resistance are the thick oxide

From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

deposition and the anisotropic via etch. Of lesser significance are the etchback planarization, an isotropic etch and plasma enhanced tetra-ethoxy siliane (PETEOS) deposition. A full length version of this paper is given in Rietman et al. (1997).

6. Summary

The controller described in this paper uses real-time process signatures, in conjunction with a measure of the product quality from the previous batch (measured oxide thickness) to compute on a wafer-by-wafer basis the ideal overetch time for each wafer. The prototype was developed from a production database and then installed on a production machine for continued improvement. Before the controller was allowed to etch real product it had already incorporated the information from "seeing" almost 6000 wafers. The production worthy controller has been installed on two reactors in a MOS fabrication line.

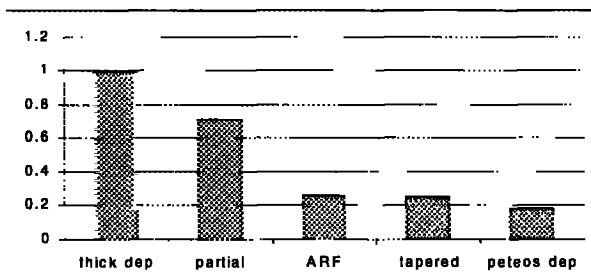

Figure 4. Pareto ranking of the via mini-fab showing the impact of the processing steps on the M1M2 resistance.

Prior to installing the neural network control on the plasma reactor the processing included a set-up step to determine the overetch time for the batch of wafers. With the neural network control on the wafer-to-wafer basis this set-up step has been eliminated. This yielded a 20% improvement in throughput in this plasma reactor. In addition the controller improved the standard deviation film thickness by 40%.

The most significant result for the large system model is, the demonstration that one can build integrated nonlinear process models for discrete manufacturing processes and conduct sensitivity analysis for a Pareto study. In a larger manufacturing system, larger than our via mini-factory, we could conduct an analysis that points directly to the key processing steps and their impact on yield. For example, we could use this method to study the processing steps that drive specific yield metrics, e.g. (L-eff, a transistor parameter in integrated circuits.)

In conjunction with the factory-based Pareto analysis one could conduct a similar study for the individual processing-zones. From this one could deduce the optimum tool set and their impact on the yield metric for that process-zone. With the aid of neural network process-zone models, genetic algorithms and other heuristic methods it should also be possible to conduct a study to find the optimum number of tools for a given processing-zone and thus, reduce capital investment and likely extend the life-time of the current generation of processing equipment. These whole system models could be used in prediction of yield, prediction of bottle-necks and process failures one or more days in advance. These system models could form the basis of computer integrated manufacturing systems for the 21st Century.

Acknowledgments

I thank Milton Beachy and Earl Lory for collaborating on various aspects of these projects. I thank Robert Frye for friendship and many technical discussions.

References

1. Beneke, M., Leemis, L. M. Schlegel, R. E., and Foote, B. L., 1988, Spectral Analysis in Quality Control: A Control Chart Based on the Periodogram, *American Stat. Assoc. Technometrics,* 30(1), 63-70

2. Boskin, E. D., Spanos, C. J. and Korsh, G. J., 1994, A Method for Modeling the Manufacturability of IC Designs, *IEEE Trans. on Semicond. Manuf.*, 7(3), 298-305

3. Chapman, B. 1980, *Glow Discharge Processes: Sputtering and Plasma Etching*, New York, Wiley

4. Doty, L. A., 1991, *Statistical Process Control*, New York, Industrial Press Inc.

5. Guo, Y. and Dooley, K., 1992, Identification of Change Structure in Statistical Process Control, *Int. J. Prod. Res.* 30(7), 1655-1669

6. Hornik, K., Stinchcombe, M. and White, H., 1990, Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks, *Neural Networks*, 3, 551-560

7. Hwarng, H. B., 1995, Multilayer Perceptrons for Detecting Cyclic Data on Control Charts, *Int. J. Prod. Research*, 33(11), 3101-3117

8. Hwarng, H. B. and Hubele, N. F. 1991, X-Bar Chart

From: Proceedings of the Artificial Intelligence and Manufacturing Workshop. Copyright © 1998, AAAI (www.aaai.org). All rights reserved.

Pattern Recognition Using Neural Networks, *ASQC Quality Congress Transactions* - Milwaukee, 884-889

9. Hwarng, H. B. and Hubele, N. F., 1993, Back-Propagation Pattern Recognizers for X-Bar Control Charts: Methodology and Performance, *Computers Ind. Eng.* 24(2), 219-235

10. Manos, D. M. and Flamm, D. L., 1989, *Plasma Etching an Introduction,* New York, Academic Press

11. McCulloch, W. S. and Pitts, W., 1943, A logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics,* 5, 115-133

12. Minsky, M. L. and Papert, S. A., 1969, *Perceptrons: An Introduction to Computational Geometry,* Cambridge , MA, MIT Press, Cambridge, MA

13. Rietman, E. A. and Patel, S. H., 1995, 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. on Semicond. Manuf.,* 8(3), 304-308

14. Rietman, E. A. Friedman, D. J. and Lory, E. R., 1997, Pre-Production Results Demonstrating Multiple-System Models for Yield Analysis, *IEEE Trans on Semicond. Manuf.,* 10(4), 469-481

15. Rumelhart. D. E. Hinton, G. E. and Williams. R. J., 1986, Learning Internal Representations by Error Propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. Foundations,* Rumelhart and McClelland. eds., Cambridge. MA, MIT Press

16. Smith, A. E., 1994, X-bar and R Control Chart Interperation Using Neural Computing, *Int. J. Prod. Res.* 32(2), 309-320

17. Turner, T. R., 1991, Correlation of Real-Time Monitored Process Module Parameters and Wafer Results, *SPIE Vol. 1593, Dry Etching Technology,* 145-156

18. van Roosmalen, A. J., Baggerman, J. A. G. and Brader. S. J. H., 1991, *Dry Etching for VLSI,* New York, Plenum Press

19. Weiss , S. M. and Kulikowski, C. A. *1991, Computer Systems That Learn,* San Mateo, CA, Morgon Kaufmann

20. Widrow, B. and Hoff, M. E., 1960, Adaptive Switching Circuits, *IRE WESCON, Convention Record IRE,* New York, 96-104

21. Widrow, B. and Hoff, M. E., 1962, Associative Storage and Retrieval of Digital Information in Networks of Adaptive Neurons, *Biol. Prototypes Synthesis and Systems,* Vol. 1, E. E. Bernard and M. R. Kane, Eds. New York, Plenum Press

22. Wolf, S., 1990, *Silicon Processing for the VLSI Era, Volume 2 - Process Integration,* Sunset Beach, CA. Lattice Press