Forecasting Cost and Yield

Pranab K. Nag
Wojciech Maly

Dept. of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Hermann Jacobs

Siemens, AG
Munich, Germany

Key Technologies:

Defect Related Yield Learning, Semicondcutor Economics, Startegic Planning, Virtual Semiconductor Enterprise Simulation

At a Glance:

This article describes a prototype of a discrete event simulator - Y4 (Yield Forecaster) - capable of simulating defect related yield loss as a function of time, for a multi-product IC manufacturing line. The methodology of estimating yield and cost is based on mimicking the operation and characteristics of a manufacturing line in the time domain. The effect of particles introduced during wafer processing as well as changes in their densities due to process improvements are taken into account. A spectrum of results are presented for a manufacturing scenario to demonstrate the usefulness of the simulator in formulating IC manufacturing strategies.

Introduction

Improving productivity and cost effectiveness of a semiconductor industry has always been a high priority and is more so in the light of increasing complexity and competition in the market. Manufacturing strategies to deal with ever increasing dimensions to this problem is often at best ad hoc in nature. It was only in 1992 when published articles on usefulness of particle monitors [1] and in just 5 years these equipment are considered almost indispensable in todays semiconductor industry. And yet there are no models to indicate how far useful they had been in improving productivity and cost effectiveness - only anecdotal evidences like the recently published article [2]. Such examples are plenty in semiconductor industries and no systematic approach has been employed to deal with them.

The most systematic attempt till date perhaps is embodied in the National Semiconductor technology Roadmap (NTRS) [3] which identifies many bottlenecks and suggests strategies for maintaining/improving the historical productivity and cost curves. However, the vastness and the complexity of this problem necessitates, on one hand, dividing into individual and

more tractable focus areas like design, test, packaging, fabrication, etc. On the other hand, these domains, sub-domains, etc., are also tightly coupled to such an extent that overall productivity cannot be realized without the capability to capture and model inter-domain dependencies.

A case in point is the projected requirement of achieving high yield learning rate. The NTRS is very terse on this topic and does not adequately expound on it. The reason could be as simple as noting that high yield learning rate can be achieved if one meets all the criterion set forth by roadmaps for design, test, fabrication, etc. But is it really that straight-forward? One has to first ensure that the design is least sensitive to fabrication uncertainties. Then fabrication must be very stable and as much as possible contamination free. Design must be characterized well so that one test and diagnose causes for all possible failures. And then one has to depend on a variety of time consuming failure analysis to pin-point the failure. The ability to pin-point failures depends on the design, the test methodologies, fabrication characteristics and effectiveness of in-line and historical data available. Rapid increase in density and number of interconnect layer compounded with heavy reliance on traditional (optical inspection) techniques during failure analysis, have only made the matter worse. It is clear that to be able to explore the spectrum of possibilities, one cannot ignore the inter-domain dependencies.

Optimum exploration of cost-revenue trade-offs is thus difficult, involving yield forecasts, and cannot be realized unless it is based on adequate experimental or simulation models. A few researchers have investigated yield learning in a semiconductor manufacturing line [4,5,6,7], but the models applied do not capture the mechanics of yield learning itself. As a result, methodologies to perform cost versus yield trade-off analysis over time do not exist at present.

To address this need, we have developed a new methodology to predict defect-related yield which takes into consideration not only the operational aspects of manufacturing, but also the process of yield learning. Models have been developed to estimate yield and cost as a function of time. The goals of this article are to present a tool -Y4 (**Y**ield **Fore**caster) - which implements this methodology, and to illustrate Y4's use in developing manufacturing strategies.

Modeling Methodology [8,9,10,11]

For the purpose of modeling yield learning curve, a manufacturing process can be viewed as consisting of two components: product fabrication and failure analysis. In order to capture the essence of the mechanism of yield learning, it is necessary to take a closer look at the key events in each of these components. In order to describe the yield as a function of time, let us first concentrate on a single product manufacturing line. Let us also assume that there exists only one type of defect originating from a single source (a piece of equipment) of particles. This simple case suffices to capture the essence of the yield learning process.

The hypothetical yield versus time curve for the above scenario resembles the staircase function shown in Figure 1. Here, T_f is the time required for analysis and detection of the failure mechanism leading to process intervention. T_e is the time needed for a process correction which decreases contamination levels, and the time required for the new process parameters to be effective. T_r is the interval between the time process correction is made and the time change in yield of the fabricated wafers is observed. Thus, the total time required for yield change to occur is $T_c = (T_f + T_e + T_r)$ and the net change in yield is Y_c. The value of Y_c is determined by the new level of contamination.

Estimating T_r is equivalent to estimating the cycle time for a process, albeit partially, starting from an intermediate process step where the correction is made until the last step of the process. Thus, it is the sum of the raw processing time (RPT) and the queuing time that

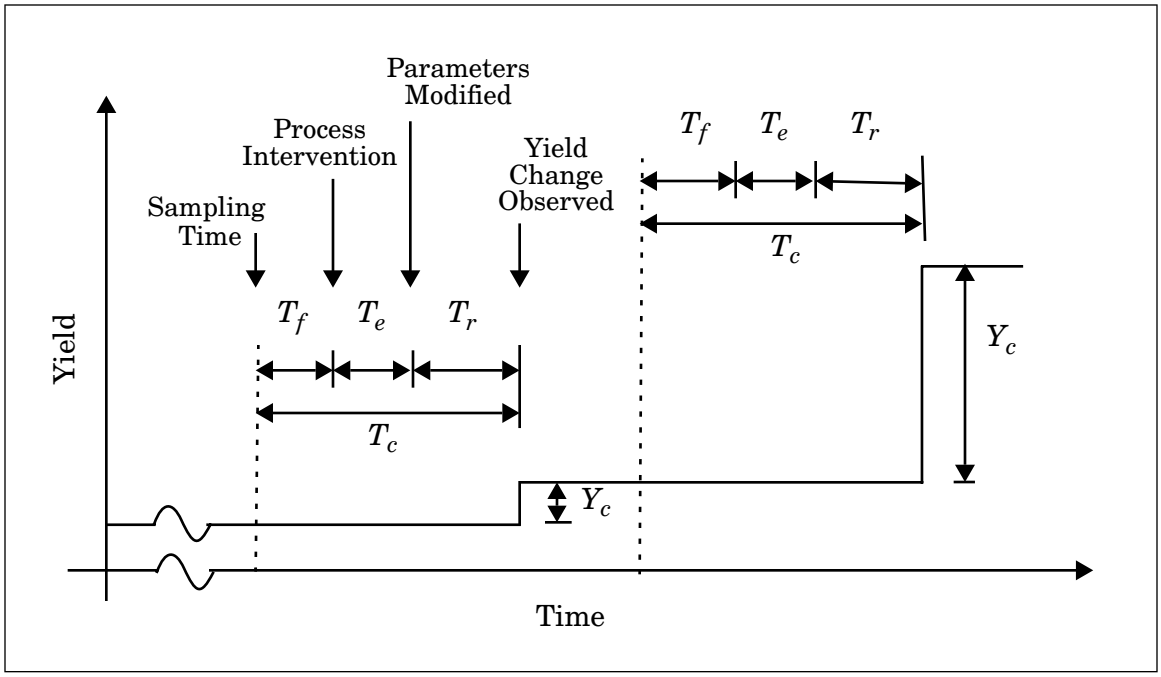

Figure 1. Key events in yield learning process.

results when wafers must wait between process steps. One of the major contributors to the queuing time is the downtime of the equipment. Note that the time factor T_e may also contribute to the equipment downtime depending on the outcome of failure analysis. T_f, the time needed to detect and localize the defect, depends on a number of attributes associated with IC design, defect and failure analysis process. The change in yield, Y_c, on the other hand, depends on the correctness of the diagnosis and the efficiency with which the contamination rate can be reduced as a result of the corrective actions.

From the above short summary, it is evident that the yield learning process should be described as a sequence of events starting with the introduction of particles, followed by detection of defects and identification of their source, and concluding with completely (or partially) eliminating the source of particles. The rate of yield learning, therefore, depends on:

1. The relationship between particles, defects and faults;
2. Ease of defect localization which in turn depends on:
   a. size, layer and type of defect,
   b. level of "diagnosability" of the IC design and,
   c. probability of occurrence of catastrophic defects;
3. Effectiveness of the corrective actions performed;
4. The timing of each of the events mentioned above;
5. Rate of wafer movement through the process.

All of the above factors must be modeled in order to build an yield learning simulator.

From this basic model of the yield learning process, it is clear that the primary capability of the simulator must be to keep track of the sequence of events in a factory. The second requirement for the simulator is the ability to simulate the movement of the wafers in a fabrication line, and representing such entities as product, process recipes, equipment, person-

nel and operating rules. In [9,10] the specific modeling aspects including cost models have been dealt with in detail.

Structure and Implementation of Y4

The methodology and the models for yield learning described in the previous section have been implemented as a software tool called Y4 (Yield Forecaster). Figure 2 shows the overall structure of the Y4 framework. The heart of the simulator is the event handler which communicates with six sub-modules: the wafer movement simulator (WSIM), the yield simulator (YSIM), the failure analysis simulator (FASIM), the in-line particle monitor simulator (PSIM), the cost simulator (COSIM) and the probe tester simulator (TSIM). The operation of the event handler and these six modules can be controlled through the simulation control unit. The user can implement different models using the toolkit of functions provided for accessing and modifying the common database for all the modules and the event handling routines. A basic user interface is available to read input files for the models, output the statistics gathered and customize the simulation control strategy.

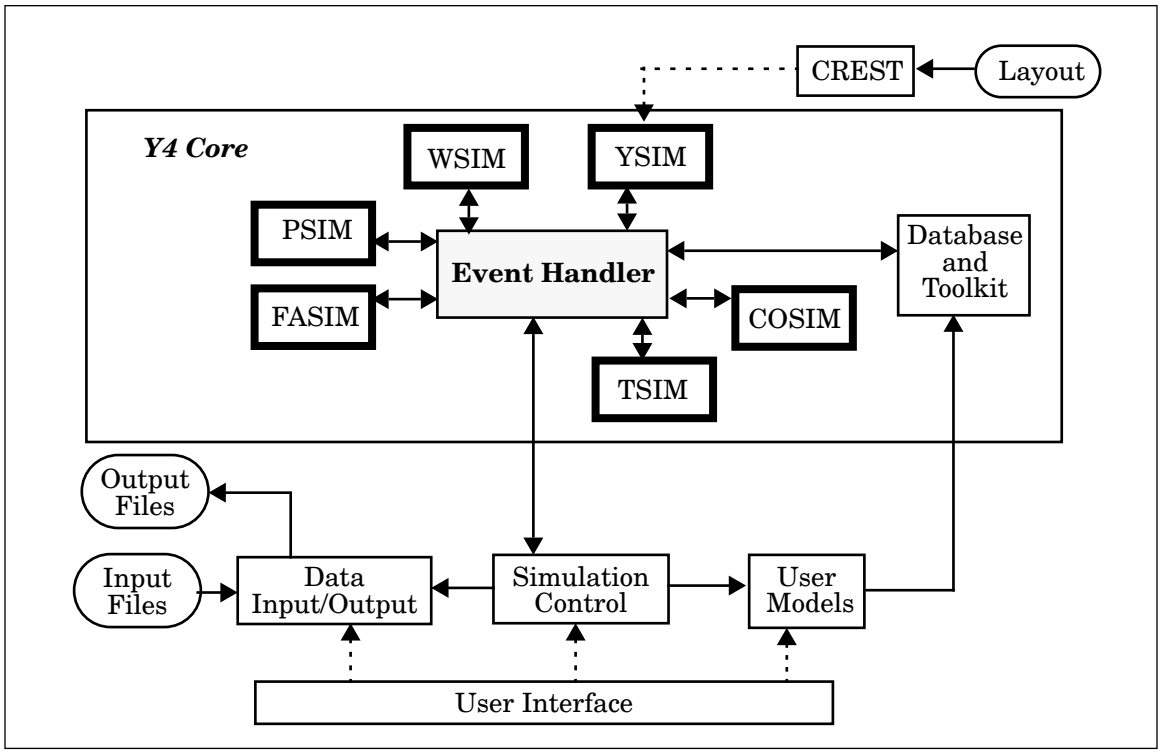

Figure 2. Top level structure of the Y4 framework.

The models described in [9,10] have been implemented as internal models of the submodules (WSIM, etc.). WSIM is similar to the commercial fabrication line simulator ManSim [12] although the current implementation models only a subset of ManSim's operating rules and conditions. On the other hand, the number of external events that can be defined in ManSim is limited. Thus, it was considered necessary to implement Y4 with the ability to define events for particle introduction (YSIM), failure analysis (FASIM), particle monitoring (PSIM), corrective actions (YSIM), and testing (TSIM).

Simulation Experiments

In this section, results of a spectrum of simulations which will demonstrate the capabilities of Y4 in being able to model interactions between various design, fabrication, in-line monitoring, and failure analysis attributes will be presented. First, the assumptions and some aspects of model parameters setup will be discussed in order to establish the premise. Then, the simulation results for various relevant scenarios will presented and compared from productivity and cost effectiveness perspectives.

In the examples that follow, we will use a 0.5 micron 3 metal CMOS process recipe. Due to its proprietary nature, data pertaining to cost of equipment, etc., has been scaled appropriately. The process recipes had to be modified for the same reason. The modified recipe consists of 145 steps using 183 pieces of equipment for a 2496 wafer starts per week (WSPW) capacity factory (a medium sized factory). The lot size is 24 wafers and thus the line capacity is 104 lots per week. The raw processing time is 302 hrs. The above process and cost data are good approximations of medium size real life manufacturing operations.

In order to validate WSIM (the core of Y4), cycle time simulations were also conducted using ManSim, and the two were found to be within 1% of each other. Cost simulations were also performed to confirm the dependence of cost of wafer on several factors including product mix, start rate, number of metal layers, etc. Details of these experiments alongwith basic simulations of cycle times, test costs and yield distributions are presented in [9].

Yield Learning Analysis

It was assumed that wafers are 6 inches in diameter which can accommodate 110 chips of 1.4 cm^2 size each. To keep it simple, only 4 types of particles were considered resulting in either shorts in metal and poly. The defect sensitivities (measured by critical areas, as a function of defect size), for each defect type was derived by scaling results obtained from several CMOS designs in order to mimic a microprocessor like product [13].

Wafers were sampled for failure analysis when there were more than 30 defective die on a wafer and when there are fewer than 3 wafers waiting to be analyzed. The failure analysis was simulated as comprising five steps: observation under microscope, observation with SEM, stripping layers (if required), cross section analysis and spectroscopic analysis. The model parameters are set such that the maximum time required to analyze 30 defects in the top metal layer was about 2 weeks (not considering the queueing time).

The weekly average of the yield trend plot is shown in Figure 3(a) along with the yield of the polysilicon and the metal 3 layers. Observe that the yield of the metal 3 layer starts to increase almost right after failure analysis is initiated (after the 10th week). The polysilicon layer yield, on the other hand, starts to increase only after another 15 weeks (around 25th week). This reflects the fact that polysilicon defects are more difficult to detect than metal 3 defects which are nearer to the surface of the chip. Further, the yield of metal 3 is low enough during the first few weeks of failure analysis that the resources are kept busy analyzing samples for metal defects. Polysilicon defects are, in effect, ignored until the metal 3 yield reaches about 0.65. However the rate of yield learning for the polysilicon layer is higher than metal 3 since the increased availability of samples with polysilicon defects compensates for the decreased diagnosability of these defects.

Figure 3(b) shows the results of yield simulation when the number of each type of failure analysis equipment is doubled. In addition to the obvious increase in the yield learning, two more effects are apparent. First, the polysilicon layer yield starts to increase around the 20th week, which is about 5 weeks sooner than in the previous case. Secondly, at this point, the metal 3 yield is higher than that in the earlier case (0.73 instead of 0.65). There is enough

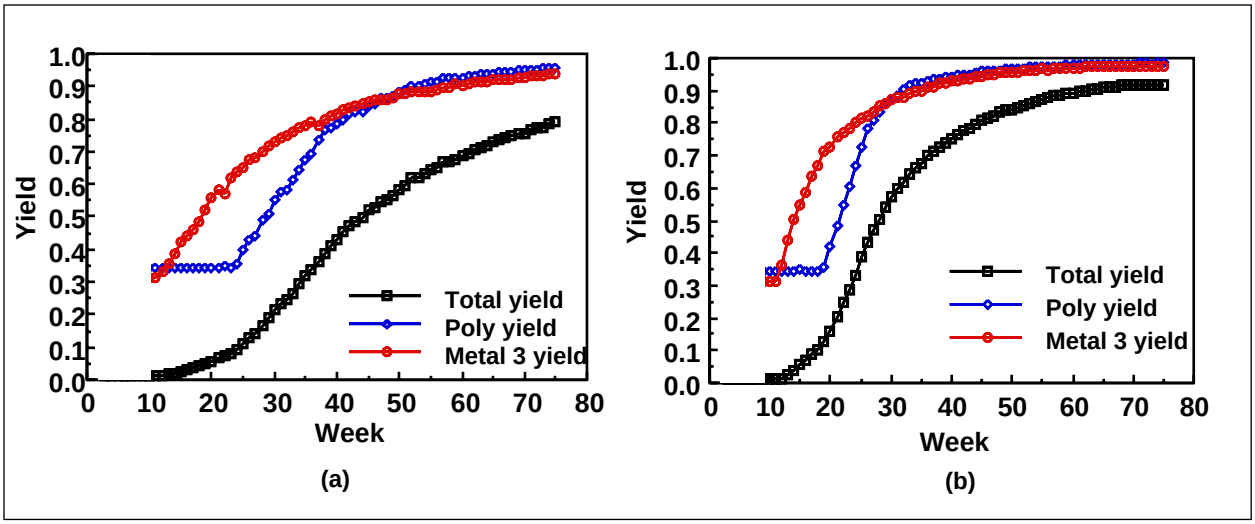

Figure 3. Average Yield learning curves for CMOS product.

leftover capacity to allow for allocation of resources to the detection of polysilicon defects while the metal defects are being analyzed. Availability of more resources enables metal defects to be diagnosed more quickly.

Impact of Sudden Change in Yield on Learning Rate and Cost

In the previous section, we have implicitly assumed that changes occurring in particle rates and size distributions due to cleaning the corresponding equipment causes an *improvement* in yield. However, they may possibly change in such a way as to *degrade* the yield. Specifically, at the end of 30th week, the mean of the particle number distribution for one of the seven sputtering tools is assumed to increase by a factor of five.

Figure 4(a) shows the result of the simulation illustrating the yield trend plots. Observe that the net yield learning rate has decreased compared to the result shown in Figure 3. The increase in metal defects causes metal yield to drop first. After a certain delay, failure analysis catches up with the increased number of defective die with metal defects, and metal yield starts to increase again. But at the same time, the polysilicon yield learning rate drops because failure analysis resources are mostly consumed in detecting metal defects.

Figure 4(b) illustrates a similar situation but with double the failure analysis capacity. As expected, the yield learning rate is higher than in the simulation shown in Figure 3(b). But there is an important difference between the two sets of yield learning curves. In the latter case, the yield learning rate of polysilicon layer remains essentially unaffected. This result again illustrates that the extra capacity helps to perform analysis on polysilicon defects in spite of higher occurrence of defective die with metal defects. Also, at the time the yield problem occurs, the metal yield is high enough that the number of defective die sampled for analysis is already low. Thus, the failure analysis facility has little trouble absorbing the relatively small increase in the number of defective die with metal defects.

It is interesting to compare the two manufacturing lines - one with a normal capacity and the other with doubled capacity of failure analysis - from the perspective of sensitivity towards yield degradation. Table 1 summarizes the results for the two manufacturing lines. The cumulative number of good die for the simulation period and the average cost are compared. Notice that, as it should be expected, the manufacturing line with more failure analy-

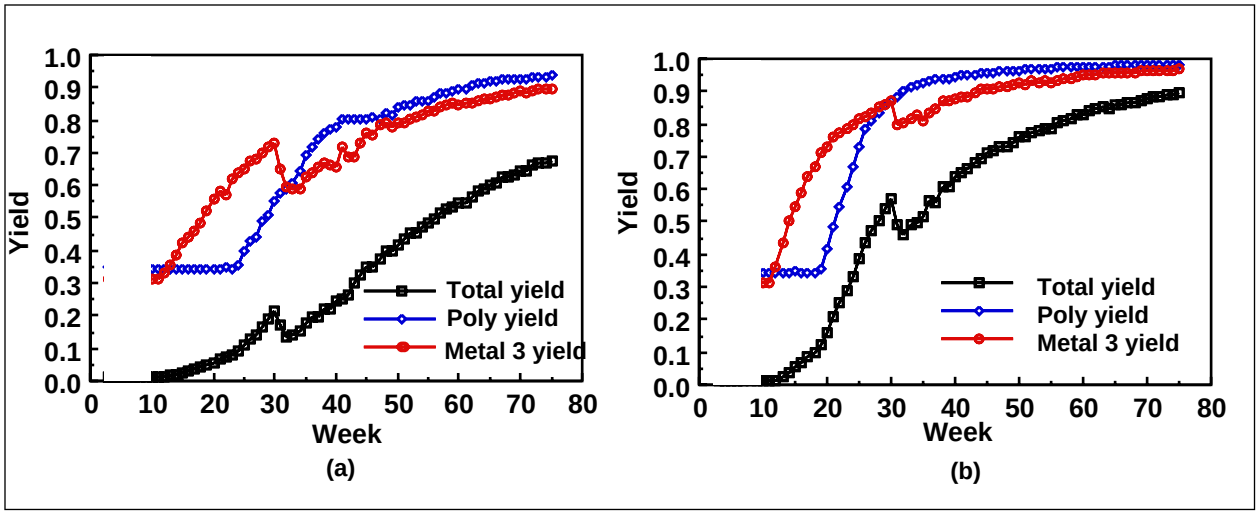

Figure 4. Average Yield learning with sudden increase in defect release rates.

sis capacity is much less sensitive to the yield problem. Thus, any loss incurred due the yield problem illustrated earlier is appreciably reduced in the second manufacturing line.

	Normal capacity			Double capacity		
	Undis-turbed fab	With yield degra-dation	% change	Undis-turbed fab	With yield degra-dation	% change
Number of good die (in million $)	7.62	5.81	-23.75	11.54	10.49	-9.1
Cost of die ($)	72.52	94.92	+29.92	51.13	56.90	+11.28
% of cost from failure analysis	5.47	5.32	-2.74	11.5	12.44	+8.17
Profit (in mil-lion $)	209	30	-85.6	564	452	-19.9
Profit (% of investment)	37.8	5.4	-	95.6	75.7	-

Table 1 Cost comparison.

Finally, for argument's sake, assume that all the ICs produced can be sold at $100 each. The last two rows of Table 1 show the estimated profit in absolute value and as a percentage of the total investment, respectively. Comparing the case where there are no yield disturbances, one can see that an extra investment of $38M in failure analysis facility increases the profit by $355M.

Impact of Particle Monitors on Yield Learning

Particle monitors are employed with the expectation that a substantial fraction of defects would be detected as and when they occur during fabrication leading to perhaps considerably faster yield ramp. In the least, particle monitors can help identify critical steps where particles may cause yield loss and detect out-of-control situations. But this technology has its limitations which include high equipment cost, questionable resolution below 1 micron particle size and wafer throughput. In fact, one can push the limits of resolution to certain extent at the expense of throughput rate. To date, the impact of particle monitors on cost and yield (an thus productivity) has not been studied.

Off-line failure analysis, however, has been traditionally the main backbone of yield ramping. It is slow, can be expensive but at the same time can be very accurately pin-point the source of yield loss. So, on one hand, particle monitors can respond quickly to yield loss problems but can be inaccurate and, on the other hand, failure analysis has long cycle time but can be very effective in yield ramping. In this section, we will introduce simple models and results which explores this scenario.

Unlike the previously presented examples, a few changes were made to the simulation setup to reflect a more realistic scenario. It was assumed that 13 particle sources resulted in 12 unique defect types leading to faults. Process recipes were modified to include the particle monitoring step after each step where particles are introduced. It was assumed that 4 chips on 3 wafers are scanned on *all* wafers. This required 16 scanners operating at 90% utilization level. It was also assumed that defects in lower layers like active area and poly are more difficult to detect than defects in metal layers. A simple rule to initiate equipment cleaning as a result of particle monitoring activity was implemented to simulate yield learning. Equipment cleaning is initiated when average number of particles per die in a lot exceeds a given threshold (20). The excess downtime of equipment due to such activity was limited to maximum of 5%. Under these assumptions the simulation resulted in the yield learning curve shown in Figure 5 and expected the yield learning rate is slow and the highest yield achievable is also low.

If we turn our attention to off-line failure analysis alone, then the yield learning curve is as shown in Figure 6. The learning rate and final yield are obviously higher and this is due to some important differences in assumptions. It was assumed that defect reduction as a result of failure analysis is much more effective, by as much as factor of 5 at times (depending on the correctness of diagnosis), than due to particle monitoring. However, the feedback cycle time of failure analysis is usually considerably longer compared to the particle monitors.

Figure 7 shows the yield as a function of time when *both* particle monitors and failure analysis are used to ramp up the yield. For comparison, the curves in Figure 5 and Figure 6 are also shown. In this example, we assumed that these two yield ramping methods are essentially independent of each other under the constraint that the net downtime of any equipment due to equipment cleaning must not exceed the predefined threshold of 5%.

Let us now turn our attention to productivity and cost comparisons as shown in Table 2. As it can be seen, for the case where both techniques are employed for yield ramping productivity is highest and cost of good die is lowest. From a strategic point of view timing is important especially for ASIC products. The last two rows compares the time to reach a specified number of good dies (2 million) and number of good dies produced in a given period of time (20 weeks

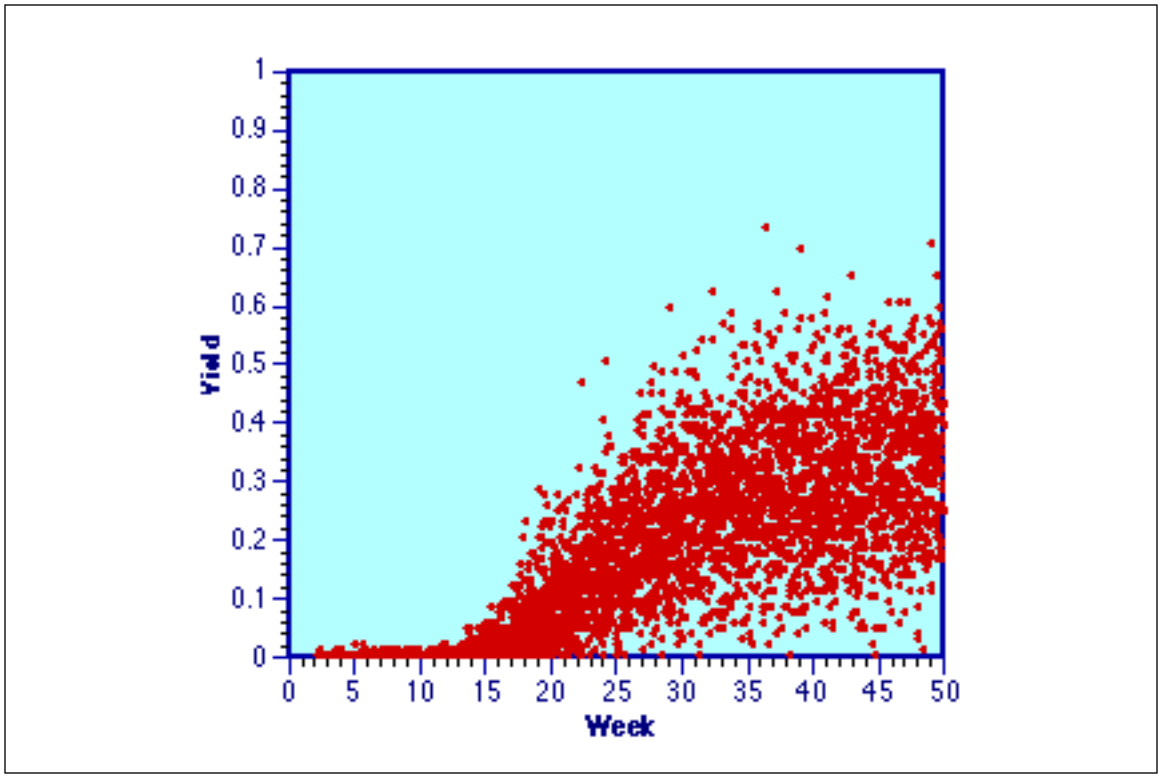

Figure 5. Particle monitoring initiated yield learning.

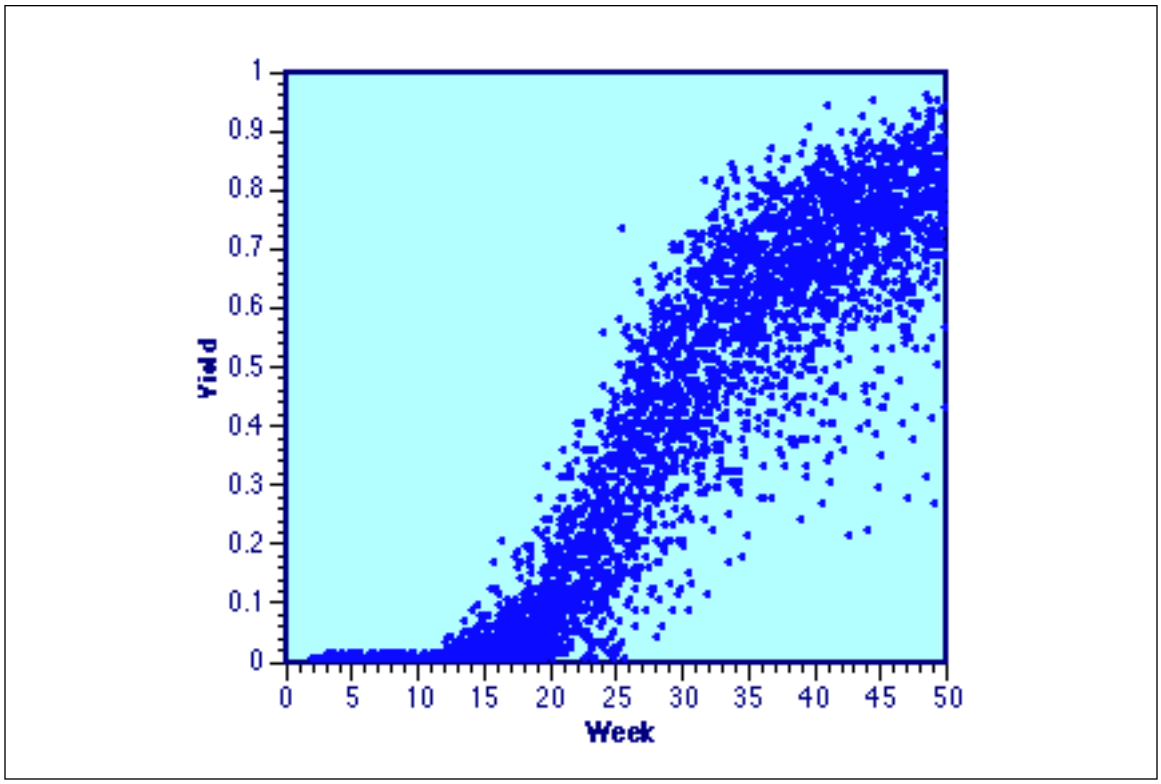

Figure 6. Failure analysis initiated yield learning.

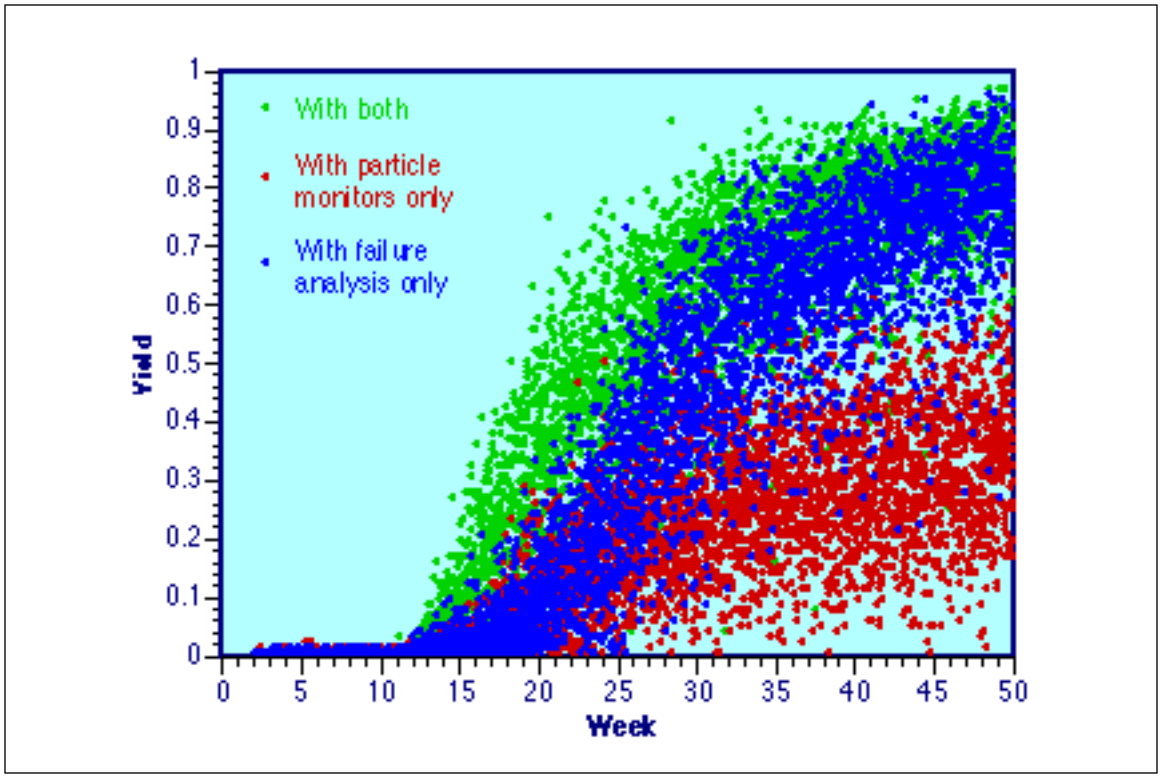

Figure 7. Combined yield vs. time curves.

in our case). As it is clear, employing both these techniques in conjunction results in a strategic advantage which is quantifiable.

	Particle monitors only	Failure analysis only	Both together
Total number of good die (in millions)	2.14	4.45	5.65
Cost of good die (in $)	167	78	65
Total product cost (in million $)	359	342	371
% cost from failure analysis	0	3.48	3.24
Time to reach 2 million good die (in weeks)	36.5	25.5	20

	Particle monitors only	Failure analysis only	Both together
Number of good die in 20 weeks (in 1000's)	654	1094	2000

Table 2 Comparison of three different yield ramping strategies.

Conclusions

We have presented a methodology to estimate both cost and yield of VLSI circuits as a function of time. The key and unique characteristic of our methodology is the **integration** of major relationships governing the kinetics of the IC manufacturing operation. Such integration provides a very powerful option for the crucial process of strategic manufacturing design and decision-making.

The methodology and the models were implemented as the software tool Y4. Through a spectrum of simulation results we have illustrated that Y4 can reasonably replicate the manufacturing line characteristics. This has been achieved after extensive tuning to semiconductor manufacturing reality.

But more importantly, we have shown that Y4 is capable of simulating scenarios which are relevant to cost-revenue trade-off studies. Such a capability in our opinion is extremely valuable if one takes into account such manufacturability-related tasks as:

a. Factory design and capacity planning,

b. Product design and analysis,

c. Designing failure analysis strategy and

Finally, it must be mentioned that the approach taken in Y4 is only a first step in modeling IC manufacturing in a manner addressing inter-disciplinary trade-offs. The methodology described here should, and hopefully will, be expanded in the future. So the results presented in this article should be viewed as an opening of a new domain of study rather than as the final results of mature research.

Acknowledgments

This research has been supported by Sematech grant MC-511 for Manufacturing Design Sciences. The authors would also like to thank Tyecin Inc., for providing the software ManSim, Alfred Kersch of Siemens AG, Munich, Steven Brown of SEMATECH, and, Darius Rohan of Texas Instruments, Dallas, for providing data, encouragement and feedback.

Further Reading

[1] L.Peters, "20 Good Reasons to Use In Situ Particle Monitors," *Semiconductor International*, pp. 52-57, Nov. 92.

[2] R. Jarvis and L. Lynn Armentrout, "Full-Fab Surface Particle Detection Improves Yield," *Semicondcutor International*, vol. 20, no.6, pp. 199-206, June 1997.

[3] 1997 National Technology Roadmap for Semiconductors.

[4] D. Dance and R. Jarvis, "Using Yield Models to Accelerate Learning Curve Progress," *IEEE Trans. on Semiconductor Manufacturing*, vol. 5, no. 1, pp. 41-46, 1992.

[5] J. A Cunningham, "Using the Learning Curve as a Management Tool," *IEEE Spectrum*, pp. 45-48, June, 1980.

[6] D. R. Latourette, "A Yield Learning Model for Integrated Circuit Manufacturing," *Semiconductor International*, pp. 163-170, July 1995.

[7] R. E. Bohn, "The Impact of Noise on VLSI Process Improvement," *IEEE Trans. on Semiconductor Manufacturing*, vol. 8, no. 3, pp. 228-238, Aug. 1995.

[8] P. K. Nag and W. Maly, "Yield Learning Simulation," *Proc. of SRC TECHCON '93*, pp. 280-282, Oct. 1993.

[9] Pranab K. Nag, *Yield Forecasting*, Ph.D. Dissertation, Carnegie Mellon University, April 1996.

[10] P. K. Nag, W. Maly, and H. Jacobs, "Simulation of Yield/Cost Learning Curves with Y4," *in Trans. on Semiconductor Manufacturing*, vol. 10, no. 2, pp. 256-266, May 1997.

[11] Y4 Project WWW Documentation @ http://www.ece.cmu.edu/afs/usr/pkn/Y4.html

[12] ManSim X, User Manual, Tyecin Systems Inc, San Jose, CA, 1995.

[13] P. K. Nag and W. Maly, "Hiearchical Extraction of Critical Area for Shorts in Very Large ICs," *Proc. of Intr Workshop on Defect and Fault Tolerance in VLSI Systems (DFT)*, pp. 19-27, Lafayette, Nov. 1995.