IC Performance Prediction for Test Cost Reduction

Jungran Lee
Dept. of Computer Science
Texas A&M University
College Station, TX USA

D. M. H. Walker
Dept. of Computer Science
Texas A&M University
College Station, TX USA

Linda Milor, Yeng Peng, Gene Hill
Advanced Micro Devices
Sunnyvale, CA, USA

Abstract – This paper describes a methodology for building models predicting manufactured integrated circuit performances as a function of inline and wafer electrical test measurements. We show how these predictions can be used to predict the performance of an industrial microprocessor, and reduce the average number of speed bins that must be tested by 45%.

INTRODUCTION

As manufacturing geometries continue to shrink and circuit performances increase, statistical process variation is of increasing concern. They must be controlled to reduce parametric yield loss, and the resulting circuits tested in order to guarantee that they meet their specifications. However it is often impossible to directly measure process variation. The goal of this work is to build models predicting manufactured integrated circuit (IC) performances as a function of inline and wafer electrical test measurements, as shown in Figure 1. Such models can be used for a variety of applications. In process diagnosis, they can be used in reverse to determine the source of process variability, and identify test structures and measurements to help control that variability. In process optimization, they can be used to select process settings to optimize the tradeoff between parametric yield loss and circuit performance. In scheduling applications, performance prediction can be used to reorder the processing of wafers in assembly and final test in order to meet product demands with minimum cycle time. In test cost reduction, prediction models can be used to predict one IC performance given the measurement of another. One example is to test the circuit at one temperature and predict the performance at another temperature.

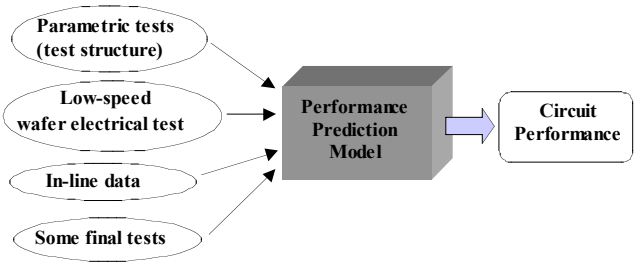

Figure 1. IC performance prediction modeling.

Given the increasing difficulty of testing high-speed microprocessors, the rapidly rising costs of testers, and increasing cost pressure in the microprocessor market, prediction models provide an attractive means of reducing test costs. In this work, we build a model predicting microprocessor clock speed as a function of wafer electrical tests, and use it to significantly reduce the number of speed bins that must be tested to bin the product.

In the sections that follow we discuss our model-building methodology, the results of building a performance prediction model for an industrial microprocessor, and experimental results on the test cost reduction application.

METHODOLOGY

The model-building methodology consists of the following steps:

1. Identification of low-cost measurements that correlate well with the circuit performances of interest.

2. Construction of response surface models using measurements and performances.

Many different types of measurements are correlated to IC performances. These include inline measurements [1][2], such as a poly CD; wafer electrical test structure measurements, such as ring oscillator frequency; and product measurements, such as the delay on a long circuit path. Test structures can even be specially designed to have high correlation to performance. Selection of the most appropriate measurements to use depends on the particular application. The decision will typically be based on the cost of the measurement (including any die area cost) versus the value of the prediction. In general, we want the lowest possible measurement costs and the highest possible prediction accuracy.

Correlation of measurements to performances can be done empirically [3] or by simulation [4]. Simulation can be done with little production data, but it requires much design data and well-calibrated models. In our experiments we had production data, but only meager design data, so we only describe the empirical approach in this paper. A full discussion of a simulation-based model-building approach can be found in [5][6].

After selecting a set of measurements sensitive to circuit performances, low-order polynomial response surface models are built, using the simulated or measured data. We use stepwise regression to build such models. When several measurements are highly correlated with one

another (e.g. several different types of ring oscillators), we discard those with lesser correlation before building the model.

For accurate performance prediction, we must take into account spatial variation in process parameters, particularly variation in L_{eff} and interconnect parasitics [7]. If we make measurements on every die, we can estimate the local variation by looking at neighboring measurements. One estimate is the average of the neighboring measurements. Another estimate is the gradient from the plane equation $Ax+By+C = 0$ fit to the neighboring measurements. In addition to a local gradient, random and systematic local variation in process parameters is also present within a die. This may be averaged out over long circuit paths, but can reduce the correlation between a small electrical test structure and a long path [8].

MICROPROCESSOR MODELING

We apply our methodology to predicting the critical path delay of an industrial microprocessor design as a function of wafer electrical test structure measurements. This design has a set of 10 transistor and 7 ring oscillator test structures in a corner of the chip. I_{dsat} measurements are taken from devices of various sizes, orientations, and densities. The ring oscillators have a variety of frontend and backend loading, including NAND gates, and metal plate capacitance.

The performance of interest is *fmax*, the worst delay among several hundred critical paths. This test is difficult and noisy at the wafer level. It also contains outliers due to spot defects. For the model-building procedure we used a production sample of 2,780 chips over 11 lots. Outliers were discarded, but it is likely that chips with smaller delays due to spot defects remained. Half of the data was used for model building and the other half for testing.

The results of evaluating a number of different models are summarized in Table 1. Measurements were either from test structures on the chip (local), the local plus the average of the neighbors, or the local measurements and the gradient fit through the neighboring measurements. Not all chips had 4 neighboring measurements, so the gradient was fit using the measurements available. All 2^{nd} order models used quadratic (quad.) terms, while all 1^{st} order models used linear (linear) terms. The notations C_iv, f_iv, C_ro, and f_ro mean that the averaged (C) and gradient-fit (f) neighboring transistor (iv) and ring oscillator (ro) measurements were used in the model, in addition to the local measurements.

The best results came for a 2^{nd} order model using local and gradient data, with an R^2 of 0.8. To avoid overfitting, variables with low fmax correlation were discarded. One major limiter on the correlation is the poor visibility of the test structures into variations in interconnect parasitics. The interconnect-loaded ring oscillators used only lower metal layers, and did not have sufficient loading to achieve high sensitivity. In addition, test structures sensitive to intra-die variation would also be required to achieve higher correlation [5]. As can be seen in Figure 3, the prediction error is within ±5%. The model slightly underpredicts performance due to a difference between the training and testing data means.

Table 1. Summary of prediction model results.

Measurements	Model variables	Model order	Model R^2
Local test structure measurements	I_{dsat} & RO freq.	2^{nd} (quad.)	0.7303
		1^{st} (linear)	0.5326
	I_{dsat}	2^{nd} (quad.)	0.6795
		1^{st} (linear)	0.5156
	RO freq.	2^{nd} (quad.)	0.6100
		1^{st} (linear)	0.4908
Local & Averaged test structure measurements	I_{dsat} (C_iv)	2^{nd} (quad.)	0.7341
		1^{st} (linear)	0.5000
	RO freq. (C_ro)	2^{nd} (quad.)	0.6933
		1^{st} (linear)	0.5008
Local & Gradient of test structure measurements	I_{dsat} (f_iv)	2^{nd} (quad.)	0.7193
		1^{st} (linear)	0.5229
	RO freq. (f_ro)	2^{nd} (quad.)	0.6792
		1^{st} (linear)	0.5247
	I_{dsat} & RO freq.	2^{nd} (quad.)	0.8011

APPLICATION

We demonstrate our model on the problem of reducing the cost of speed binning the microprocessor. The bins are bin A, bin B, and bin C, where the speed of A is less than B, which is less than C. The mean frequency of the population is approximately $(B+2C)/3$ MHz as shown in Figure 2.

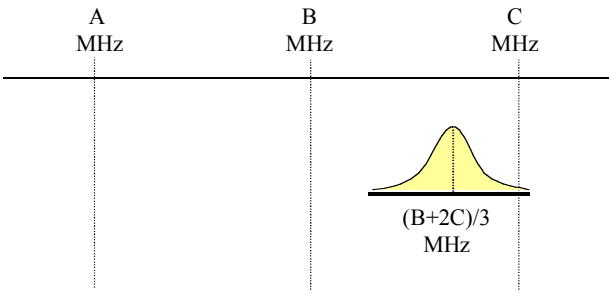

Figure 2. Microprocessor binning example.

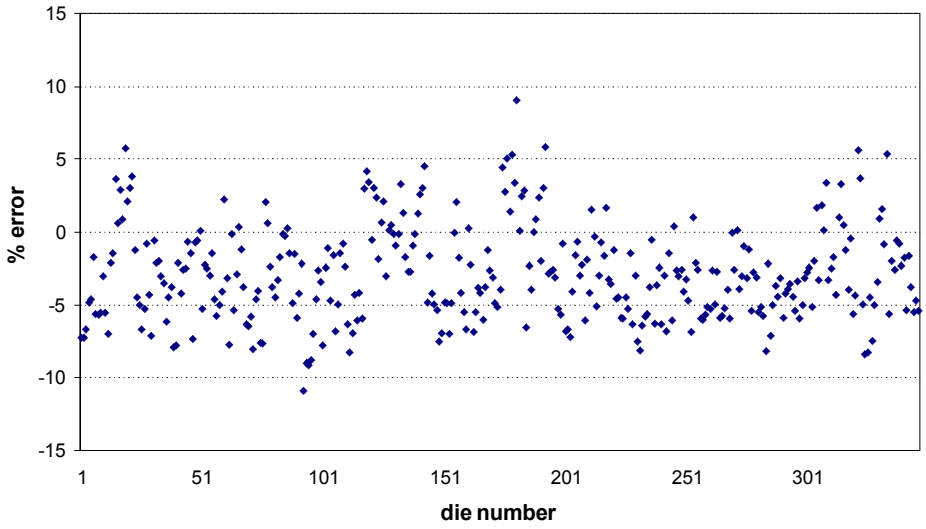

Figure 3. Delay prediction error in final model.

We analyzed three simple test strategies in Table 2. The first is to test from slowest (A) to fastest bin (C), stopping when a bin fails. The second is to test from fastest (C) to slowest (A) bin, stopping once a bin passes. The third is to test bin B first. If B passes, test bin C, otherwise test bin A. For this last case, two bins are always tested, while for the other two strategies, the number of bins tested depends on the frequency distribution of the microprocessor speed. From the table it can be seen that testing fastest to slowest results in the fewest bins tested due to the high average speed of this microprocessor.

These results depend on the yield of the microprocessor. At the start of production when parametric yield may be low, testing the slowest bin first would be a better strategy. Testinghe fastest bin first also reduces test time more than indicated by the number of bins. A chip tested beyond its speed capability will fail sooner, and the test time of the faster bins is less than that of the slower bins.

Table 2. Number of bins tested for simple test strategies.

Test Strategy	1 bin	2 bins	3 bins	Avg. # bins
A => B => C	•	5.76%	94.24%	2.94
C => B => A	16.43%	77.81%	5.76%	1.89
B => if pass, C => if fail, A	•	100%	•	2.00

In Table 3 we used the mean of the performance prediction to select the starting test bin. This is almost always bin B in our data set. We use two strategies. The first strategy is to test in the starting bin, and then go to the next higher bin if the first passes. In the second strategy, we do not test at the higher speed bin. In this case there is a risk of downbinning the product if the prediction is inaccurate. In our data 16.43% are downbinned, while the test time is cut nearly in half. This is not a reasonable economic trade-off for a microprocessor.

Table 3. Average number of test bins based on mean of predicted delay for different test strategies.

Test Strategy	Mean	
	Avg. # bins	% downbin
Mean => if pass, faster bin => if fail, slower bin	1.99	•
Mean => if pass, DONE => if fail, slower bin	1.06	16.43%

In Table 4, we applied the same strategy as in Table 2, but used the 95% confidence interval (CI) rather than just the mean. When the interval crosses a bin boundary, we select either the slower or faster bin first, as shown. If we select the slower bin and the chip passes, we do not test the next faster bin. In this case there is a chance that the chip will be downbinned. The results show that the average number of bins tested does not change significantly by using the CI, but the amount of downbinning is greatly reduced. Note that using the CI actually reduces the average number of bins for the second strategy, at the same time it reduces downbinning. Using a 99% confidence interval would result in a negligible amount of downbinning at only a slight increase in the average number of bins tested. As discussed above, testing the faster bin first will result in the test time being somewhat less than indicated by the average number of bins.

Table 4. Average number of test bins based on 95% CI of predicted delay for different test strategies.

Test Strategy	95% conf. interval	
	Avg # bins	% downbin
CI => if pass, faster bin (slow) => if fail, slower bin	2.01	•
CI => if pass, DONE (fast) => if fail, slower bin	1.05	1.73%

CONCLUSIONS

We have shown that an accurate performance prediction model can be built using low-cost wafer tests. These tests are done on individual transistor and ring oscillator test structures commonly found on large integrated circuits. We have developed a performance prediction model for an industrial microprocessor and applied it to the problem of speed binning. Compared to the best simple bin test strategy, our prediction-based approach reduces the number of tested bins by 45% at the cost of a slight amount of downbinning. By using a larger confidence interval, downbinning can be made negligible. Our work also shows that the local gradient in process parameters has a significant impact on circuit performance. Including the gradient in the performance prediction increased the correlation from 0.73 to 0.80. We believe that by using test structures with better visibility of interconnect variation, it is possible to achieve higher correlations.

ACKNOWLEDGEMENTS

This research was funded in part by the National Science Foundation under grant MIP-9406946 and by the Texas Advanced Technology Program under grant 999903-100.

REFERENCES

[1] C. R. Shyamsundar, P. K. Mozumder, and A. J. Strojwas, "Statistical Control of VLSI Fabrication Processes: A Software System," IEEE Trans. Semiconductor Manufacturing, vol. 1, no. 2, May 1988, pp. 72-82.

[2] P. K. Mozumder, C. R. Shyamsundar, and A. J. Strojwas, "Statistical Control of VLSI Fabrication Processes: A Framework," IEEE Trans. Semiconductor Manufacturing, vol. 1, no. 2, May 1988, pp. 62-71.

[3] D. A. Hanson, R. J. G. Goossens, M. Redford, J. McGinty, J. K. Kibarian, and K. W. Michaels, "Analysis of Mixed-Signal Manufacturability with Statistical Technology CAD (TCAD)", IEEE Trans. Semiconductor Manufacturing, vol. 9, no. 4, pp. 478-488, Nov. 1996.

[4] C. S. Murthy and M. Gall, "Process Variation Effects on Circuit Performance: TCAD Simulation of 256-Mbit Technology," IEEE Trans. CAD, vol. 16, no. 11, pp. 1383-1389, Nov. 1997.

[5] J. R. Lee, "IC Performance Prediction for Test Cost Reduction", Ph.D. Dissertation, Dept. of Computer Science, Texas A&M University, January 1999.

[6] V. Ramakrishnan and D. M. H. Walker, "IC Performance Prediction System," IEEE International Test Conference, Washington DC, Oct. 1995, pp. 336-334.

[7] Z. Lin, C. J. Spanos, L. S. Milor, and Y. T. Lin, "Circuit Sensitivity to Interconnect Variation," IEEE Trans. Semiconductor Manufacturing, vol. 11, no. 4, Nov. 1998, pp. 557-568.

[8] L. M. Huisman, "Correlations Between Path Delays and the Accuracy of Performance Prediction", IEEE International Test Conference, Washington, DC, Oct. 1998, pp. 801-808.