Collaborative Platform for DFM

Wojciech Jacob Poppe

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-175
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-175.html

December 20, 2007

Collaborative Platform for DFM

by

Wojciech Jacob Poppe

B.S. (University of California, Berkeley) 2003

M.S. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering

and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Andrew R. Neureuther, Chair

Professor Tsu Jae King

Professor David Dornfeld

Fall 2007

Collaborative Platform for DFM

Abstract

Collaborative Platform for DFM

by

Wojciech Jacob Poppe

Doctor of Philosophy in Electrical Engineering

University of California, Berkeley

Professor Andrew R. Neureuther, Chair

This dissertation addresses the two biggest challenges in Design for Manufacturing (DFM), how to inject process variations into design and how to identify and quantitatively characterize the main sources of transistor performance variation so that the information that is fed into design is accurate enough to make design tradeoff decisions effectively. To address these challenges the Collaborative Platform for DFM has been built in three main parts; the Parametric Yield Simulator, which is a scripted link between process simulation, non-rectangular device modeling, and circuit simulation, a process characterization strategy that leverages a large set of process sensitive electrical test structures for extracting process conditions, and a collaborative database that serves as the glue between simulation and experiment and facilitates high volume data analysis.

The Parametric Yield Simulator (PYS) is built as a modular platform that links processing, currently lithography simulation, device modeling, and circuit analysis. This simulation flow is built around a non-rectangular transistor model that uses a set of channel position dependent slice lookup tables for fast model generation and translates a 2D geometrical gate shape into an equivalent 1D compact transistor model. The PYS can be wrapped with perl scripts to simulate layouts across the lithographic process window and hence be used for rapid prototyping of process sensitive test structures.

In order to identify the main sources of threshold voltage variation and quantify their significance a multi-student testchip has been designed with over 15,000 individually probable transistors and test structures. All test structures are electrically probable and have various sensitivities to different process parameters. Most use a novel Enhanced Transistor Electrical CD Metrology (ETEC-M) that is based on an "enhanced" transistor that is 3X more sensitive to gate length variation than a standard transistor.

In order to make sense of the high volumes of data, a relational database was built for data aggregation with structure that enables queries and flexibility to adapt as new attributes become necessary for data analysis. A high volume process extraction strategy used process sensitive and process insensitive test structures to identify a unique signature of each process parameter on the data, which it used to extract defocus and misalignment with sub-10nm accuracy. Some unintentionally sensitive designs where found to be 2X as sensitive to defocus as an isolated line.

A second testchip, that has been manufactured in a short loop single layer experiment, demonstrated that using electrical open/short data from sets of structures can be used to extract defocus with sub-10nm accuracy. Sensitivities to different layout parameters were quantified in simulation and experiment. This thesis demonstrates how high volumes of electrical data from process specific test structures can be used to accurately characterize the main sources of transistor performance variation and enable more accurate DFM tools.

_____

Professor Andrew Neureuther

Committee Chairman

Dla Mamy

The one who made me who I am

Acknowledgements

This research and thesis could not have been completed without the help and guidance from many people. First and foremost I have to thank my advisor and mentor over the last six and a half years, Professor Andrew Neureuther. His guidance in my research as well as in solving problems in general will help guide me through the rest of my career. Under his guidance I was able to grow from a wide eyed undergrad to a Doctor of Philosophy. My career at Berkeley has also been greatly influenced by Professors Tsu-Jae King, David Dornfeld, Bora Nikolic, William Oldham, Robert Meyer, Jan Rabaey, Costas Spanos, Kurt Keutzer, Fabian Pease, Alexander Liddle, Erik Anderson and Ali Javey.

Working in Prof. Neureuther's group I have had the pleasure and privilege to work and learn from (in chronological order) Mosong Chen, Costas Adam, Mike Williamson, Yasheesh Shroff, Michael Shumway, Scott Hafeman, Lei Yuan, Frank Gennari, Garth Robins, Greg McIntyre, Charlie Zang, Michael Lam, Seiji Nagahara, Dan Ceperley, Koji Kikuchi, Juliet Holwill, Eric Chin, Lynn Wang, Chric Clifford, Marshal Miller, Darshana Jayasuriya, and Patrick Au. Juliet Holwill, Lynn Wang, Patrick Au and Darshana Jayasurija have significantly contributed to the content in this thesis. Outside the TCAD and lithography group I have had the pleasure of working with or learning from Alejandro De La Fuente, Chung Sung, Zheng Guo, Liang-Teck Pang, Paul Friedberg, Louis Alarcon, Qingguo Liu, Yu Ben, Mohan Dunga, Donavan Lee, Drew Carlson, Jason Cain, Shalin Mody, and Piotr Prokop. I have also had tremendous help from the microlab staff, especially Kim Chan, Phill Guillory, Sia Parsa, Attila Horvath, and Anita Pongracz.

I had the good fortune to participate in the FLCC grant and have my research sponsored by SRC and DARPA. This has given me the opportunity to collaborate with a lot of people in industry. I owe special thanks to ASML, Cypress, and SVTC for participating in the Enhanced NMOS and short loop experiments. I especially want to thank, Anita Pici, Susan MacDonald, Mircea Dusa, Luigi Capodieci, Cyrus Tabery, Vinny Pici, David Hansen, Mary Zawadzki, Igor Polishchuk, Rosemary Gettle, Hui Brickner, Feng Dai, Sundar Narayanan, Kevin Jang, Diana Coburn, Srikanth, Govindaswamy, Nadar Shamma, and Chris Capella for helping me get my experimental and simulation work done.

I have also had the good fortune to round out my technical side with the opportunities offered by the Management of Technology Program. I want to especially thank Andrew Isaacs, as going through his classes has sparked an entrepreneurial thirst in me that I hope one day to quench. As part of the MOT program I had the opportunity to become a Mayfield Fellow and learn a lot from my Mayfield Fellow 06 comrades; Sebastien Peyen, Ryan White, Alex Ortiz, Jeff Boortz, Lawrence Chang, Melissa Ho, David Noy and RJ Honicky.

Also, for cheerful assistance with administrative and grant issues, I would like to thank Charlotte Jones, Ellen Lenzi, Vivian Kim, Ruth Gjerde, and Farah Pranawahadi.

Most importantly, I cannot express my appreciation for my family with whom I came to this country, my Mom, Dad, and brother Aleks. It has been a tough journey over the years, but it has been the strength of our family as a unit that has helped me get to where I am today. I am forever grateful for the support I have been given.

Finally, and certainly most of all, I am forever grateful for the unending love, patience and dedication of my wife Dawn and two boys Anatol and Andrzej. The three of you make all of this worthwhile.

Table of Contents

Table of Figures

1 Collaborative Platform for DFM: Overview and Introduction

The semiconductor industry has managed to stay on an exponential technological advancement cycle for over thirty years with performance and density doubling and cost decreasing by 50% every two years or so. Two major drivers of this revolution have been advancements in photolithography that have reduced the minimum printable feature size by 30% every two years and an incremental infrastructure that enabled chip designers to utilize smaller and faster transistors. On one side of this massive engine are the technologists that have consistently come up with technologies to bend nature into doing what they want to. This means printing feature sizes 1/2,000 the width of a human hair and growing layers of silicon dioxide 1/50,000 the width of a human hair with a consistency of less than 1 defect/1,000,000,000 transistors. On the other side are designers that have devised methods of weaving complex logical systems or chips consisting of millions and potentially billions of transistors that double each generation. These two worlds have grown in parallel in an incremental fashion so that design techniques and circuit designs are valid from generation to generation. Electronic Design Automation (EDA) was born from this notion and Computer Aided Design (CAD) tools have been created to automate and hence scale designs to the complexity of today's chips. As long as designers or the EDA tools they use follow a set of design rules, which

are created by technologists, their circuits are guaranteed to be manufacturable. So designers do not need to worry about the challenges of building transistors at each new generation and can concentrate on the design challenges of rebuilding their 32-bit microprocessor into a 64-bit microprocessor.

This convenient division of labor is unfortunately decreasing efficiency and productivity in current process generations so the contract between design and manufacturing needs to be expanded. Design rule manuals have exponentially increased in size and complexity and now include a gray area called recommended rules[1]. This is largely due to the sub-wavelength gap, which means technologists are printing features that are a smaller and smaller fraction of the wavelength of light used to print them. Playing with various Resolution Enhancement Techniques (RETs) such as Optical Proximity Correction, off axis illumination, and high NA immersion lithography technologists have been able to push resolution, but at the cost of strong pattern dependent effects that result in complicated design rules[2]. Not only do these rules have to guarantee manufacturability, but also consistency from exposure to exposure. Since this is inherently a frequency domain problem a set of spatial domain design rules are hard if not impossible to describe these issues. With a growing sub-wavelength gap and bigger area of influence, reproducibility not only depends on the nearest neighbor in a layout, but the next next nearest neighbor. Hence a new methodology, one that leverages the comprehensiveness of process simulators needs to be implemented during the design stage to maximize silicon area utilization while minimizing process variability. This enhanced notion of addressing interaction effects in advance as to improve manufacturability has grown into the field of Design for Manufacturing.

As we enter an era of pushing manufacturing limits, Design for Manufacturing (DFM) techniques and tools will become essential to overcome the growing yield degrading effects of process variability and uncertainty. Process information needs to be communicated into the design phase in such a way that informed design tradeoffs can be made consistently. The wall of abstraction also needs to be maintained as it has enabled the exponential growth of hardware on top of technology, systems on top of hardware, and software on top of systems. For this reason a systematic methodology needs to be developed for injecting process simulation tools into design as new sources of transistor performance variation become dominant. A systematic methodology also needs to be developed for identifying and characterizing these main sources of transistor performance variation. This is inherently a multi-disciplinary problem that requires a process/device/design framework for an optimal solution. This thesis will present a Collaborative Platform for DFM that aims to meet this challenge by joining process and circuit simulators as well as a set of process characterization experiments that are needed to enable quantitative DFM.

1.1. Current state of DFM

Currently DFM is a somewhat ambiguous field as it overlaps a lot of previously existing fields such as optical proximity correction (OPC), design for test, process characterization, and Yield Analysis among others. True DFM where process information is fed upstream to design has slowly started gaining traction in both academia as well as industry. People have shown that critical paths can change if simulated gate shapes were used for static timing analysis[3] and others have shown that using regular fabrics can help mitigate these issues[4,5]. Electronic Design Automation (EDA) companies have started

coming up with products aimed at designers that model Chemical Mechanical Polishing (CMP), lithography, and Critical Area Analysis (CAA)[6]. The first generation of tools concentrated on catastrophic hotspots such as pinching or bridging, but second-generation tools are focusing on parametric performance and translate process variations into performance variations[7,8]. First generation litho hotspot checkers have also been implemented in automated hotspot fixers that can automatically fix designs by making small changes in polygon placement that has no impact on functionality, but significant impact on manufacturability[9]. These model based tools are an evolutionary step above design rules, which are either too restrictive and lead to inefficient design or not robust enough and will allow hotspots. Model based tools on the other hand can be comprehensive without being restrictive as each instance of the layout is treated independently. As long as the processes are modeled accurately, designers can maintain relative freedom with using a smaller set of design rules and technologists can guarantee manufacturability if final designs are hotspot free. DFM hence is a combination of bringing process simulation tools further upstream and process characterization or identification and quantitative analysis of the main sources of transistor performance variation.

1.2. Current State of Process Characterization

Currently there exists a wide variety of lithography and process characterization approaches ranging from metrology for process control to calibrating process models for Optical Proximity Correction (OPC) to electrically testing memories and ring oscillators to evaluate process margins. Scatterometry can be very useful in process control as it is very accurate and easily automated, and has been used in a feedback loop to minimize

process drift[10]. Electrical Linewidth Metrology (ELM) has been used to characterize systematic CD variation from a variety of sources such as illumination non-uniformity across the slit to wafer level effects during the Post Exposure Bake (PEB) and etch steps[11]. Offering a lot more geometrical flexibility, CD-SEMs have been used extensively for process monitoring and calibrating OPC models[12]. With the proper amount of automation and image recognition, CD-SEMs can be used for very detailed and extensive studies of pattern transfer variations under process non-idealities[13]. Finally, electrical test structures in the form of ring oscillators, SRAMs, and leakage test structures have also been used to look at process variations in more detail[14,15,16]. Electrical testing is easily automated and has the benefits of looking at the finished product, which is important as the final electrical performance of a circuit is what ultimately matters.

1.3. Dissertation Research Overview

Previous methods of process characterization have been shown to be very effective, but either lack scope as they may only concentrate on one specific process or lack detail if process specific test structures are not used. A more comprehensive approach should leverage the strengths of the different techniques in such a way as to reinforce individual weaknesses and paint a clearer picture together. A multi-designer testchip has the benefit of having a wide variety of test structures, perspectives, and experimental approaches all on one chip. This shotgun approach also allows characterizing processes that were not originally targeted as some test structures are bound to have a higher sensitivities than others. This is of utmost importance as having test structures with process specific sensitivities creates a unique signature for each process parameter as that parameter is varied.

Two multi-student FLCC testchips have been designed with this philosophy in mind[17]. The Enhanced NMOS testchip has six contributing student designers and over 15,000 individually probable transistors or test structures, which have varying responses to different process non-idealities. It leverages a passively multiplexed Enhanced Transistor CD Metrology strategy that minimizes parasitic leakage by eliminating control logic and "enhanced" transistors that are hyper sensitive to gate length variation. Another short loop FLCC testchip has programmable defocus monitor that can be designed to pinch open at different defocus levels. The hope is combining a broad set of test structures from multiple students will help filter out the contributions of different process effects.

To help facilitate data comparison and collaborative data mining, a database has been designed that stores data from simulation and experiments, process parameters if not nominal, as well as descriptions of all the test structures that can be used for identifying process monitors. The database was designed to centrally store all data and enable a high volume process extraction strategy, which is necessary when dealing with transistor leakage currents that are a confluence of many process steps and process parameters. In order to peal layers off the process onion and isolate specific parameters, a strategy has been devised that uses the unique signature of each process parameter on all the transistors together to extract specific process conditions. By fitting modeled dimensions to extracted conditions and comparing them to measured dimensions it is possible to evaluate the goodness of fit by looking at the distribution of residuals. As long as there is enough discrepancy between sensitivities among different test structures, incorrectly extracted process conditions will lead to a residual plot that is very different from what

one would expect from random noise. There will be a significant component of the randomness that will stem from a difference between the extracted and the actual process conditions. So accuracy can be quantified by looking at the residuals, which is very important when dealing with unknown variations.

A key component of the database and data analysis strategy is using simulation to identify process monitors, which is implemented using the Parametric Yield Simulator(PYS). The Parametric Yield Simulator makes up the second half of the Collaborative Platform for DFM as it translates characterized sources of variation, currently the lithography process, into circuit and transistor performance variation. It links process, device, and circuit simulators and models in a modular fashion that can be improved incrementally or used for variability analysis. The Parametric Yield Simulator is built around the transistor model, which is the perfect bridge between design and manufacturing as it can translate all process effects into transistor performance. This way new modeling tools can be developed on the manufacturing side and DFM tools can be developed on the design side almost independently. This approach fits the traditional evolutionary model the IC industry has used as design techniques developed using lithography simulators and first generation DFM tools will transfer from generation to generation as more and more processes are modeled. Hence the Parametric Yield Simulator is not only a tool for pushing lithography information into design, but a platform for communicating any and all sources of transistor performance variation in a standardized method that is not disruptive to designers. To tie simulation back with experiment the PYS can be driven using the database so that simulation data can automatically be uploaded. This enables an automated strategy for running dry-lab

experiments of designed testchips and evaluating test structure response to programmed process non-idealities.

1.4. Dissertation Organization

This thesis will describe the Collaborative Platform for DFM and the ways it has been used in the following sections. Chapter two describes how the Parametric Yield Simulator was implemented. Chapter three describes the basis and details of the non-rectangular transistor model. Chapter four describes the test structures and the passively multiplexed Enhanced Transistor Electrical CD Metrology strategy used in the Enhanced NMOS FLCC testchip. Chapter five describes the collaborative database and complementary website that serves as the glue between simulation and experiment as well as the front end for the Collaborative Platform. This chapter also includes a process characterization strategy that leverages the database and the large number of test structures that exist on multi-designer testchips. Chapter six is a detailed dry-lab analysis of the FLCC Enhanced NMOS testchip. Process monitors were identified using the methods explained in chapter five and used to extract misalignment and defocus with sub 10-nm accuracy. Chapter seven will detail an improved programmable defocus monitor and experiment results that were implemented in a second testchip. This testchip was designed for a short loop single layer process that was feasible to manufacture after the original Enhanced NMOS testchip process was scrapped when Cypress spun off SVTC. Chapter eight shows some initial ideas that will hopefully inspire future work on the Collaborative Platform for DFM. It proposes using the PYS for evaluating testchip results that showed pattern dependent ring oscillator frequencies and a revised test structure approach that aims at a small footprint and electronic testing. This approach can be

utilized in a production setting for accurately characterizing the process window, which is very important for maximizing DFM tool utility. Finally the conclusion will summarize the results and impacts.

1.5. Main Contributions

The Collaborative Platform, which consists of the Parametric Yield Simulator and the collaborative database, and the high volume process characterization strategy are the two biggest contributions of this thesis. The Parametric Yield Simulator is one of the first automated flows that links process and circuit simulation. It breaks down the increasingly incorrect approximation that all transistors are created equal and enables process aware analysis of transistor variations. This analysis can systematically explain what traditionally has been characterized as just random noise. The second main contribution of this thesis are the test structures in two multi-student testchips and database enabled process characterization strategy. It is shown that with a set of process specific test structures and a data analysis strategy that can identify the signature of each process parameter in the large set of test structures can be used to extract defocus and misalignment with sub-10nm accuracy. Together these two contributions can identify and characterize the main sources of variation and systematically reduce the amount of transistor variation attributed to random noise.

Furthermore, the Parametric Yield Simulator lead to the non-rectangular transistor model, the third major contribution, that translates a physical 2-dimensional geometrical gate shape into a 1 dimensional compact transistor model that is simple enough to efficiently simulate circuits in HSPICE. It does not require new transistor models in HSPICE or increase circuit complexity while accounting for complex short channel and

narrow width effects. It stands as an example of what should be the new bridge between design and manufacturing, a process aware transistor model. The transistor model can standardize all sources of process variations and translate them into transistor performance variation, which can then be used as a consistent metric for making design tradeoff when dealing with multiple process modeling tools.

The implementation of the high volume process characterization strategy lead to the fourth major contribution, two testchips with dozens of process sensitive test structures and monitors and a novel electrical linewidth metrology technique. Enhanced Transistor Electrical CD Metrology (ETEC-M) is based on a destabilized transistor that is 3X more sensitive to gate length variation and only requires one extra extension implant step. A stronger correlation between gate length and leakage current enables electrically measuring leakage current and more accurately extracting an effective gate length. With a transistor based metrology method it is possible to design 2D structures are tuned to various process parameters as well as achieving a very high packing density, which enables high volumes of data at a relatively low silicon area cost. This flexibility has been utilized in the Enhanced NMOS testchip with process sensitive structures designed to look at misalignment, defocus, dose, LER (line edge roughness), random dopant fluctuations (RDF) , corner rounding, and many other systematic effects.

The Enhanced NMOS testchip has been run though a simulation dry-lab experiment with results from the PYS automatically uploaded to the collaborative database. Simulation results with programmed process conditions were used to characterize each test structure through dose, focus, and misalignment and simulated leakage currents under randomized process conditions were uploaded to the experimental portion of the

database to test the accuracy of the high-volume process extraction strategy. Test structure evaluation identified a broad range of responses or sensitivities to dose, misalignment, and focus. This is a key ingredients to process characterization as each process parameter has a unique signature encoded into sets of different test structures. Some test structures even showed a 2X higher sensitivity to defocus than an isolated line (the traditionally most defocus sensitive feature) and others were insensitive to defocus. Using this set of test structures and an iterative approach inside the database that compared modeled and measured dimensions, defocus and misalignment were extracted with sub-10nm error and dose with sub-1% error. Not only was process extraction accurate, but metrics were found that could quantify the goodness of fit, so these results should extend to silicon experiments. As long as a set of test structures has a unique response to a particular process parameter, that process parameter can be extracted as redundancy can be used to eliminate random noise and systematic variation can be captured in programmed experiments.

A second testchip designed for a short loop single layer process was also designed that leveraged the attenuated phase shift mask with a 90 degree phase etch technology that was developed for the Enhanced NMOS testchip. The 90-degree phase etch proved to be critical in the dry-lab simulation study as it resulted in asymmetrical response through focus and a strong CD response at low defocus values. A 90 degree phase shift was originally implemented in a defocus monitor that was designed by Juliet Holwill, but in the second mask the defocus monitor was programmed by varying four layout parameters. Programming the defocus monitor with linewidth, phase shifting probe size, number of rings, and offset between probe and line changes the sensitivity and pinch-

point for the monitors. Simulation showed with 454 unique combinations of those four parameters it is possible to pinch open a target every nanometer in a range from -120nm to 60nm. Experimental results with 600 targets showed 20 targets pinching open every 10nm in a range from –90nm to +20nm. The database showed its value during data analysis as sub-10nm defocus extraction was only made possible by looking at sets of targets instead of looking at individual monitors.

The collaborative database and website hence make up the fifth contribution of this thesis. Inspired by the large number of test structures and design contributors, the collaborative database combines simulation and experiment results with a web interface that also facilitates collaborative data mining. The database itself was custom built outside the traditional database mold to have structure with flexibility and expand as it is used for different test chips. It was also optimized for statistical analysis along with organized data storage, the primary function of databases.

In summary the single biggest contribution is the high volume process extraction strategy that was validated in simulation and experiment with demonstrated sub-10nm defocus and misalignment extraction accuracy. The second biggest contribution is the Parametric Yield Simulator and collaborative database, which stand as a platform for rapidly prototyping test structures, a platform for collaborative and/or high volume data analysis, and a platform for injecting process variations into design. The non-rectangular transistor, which is the key enabler to the PYS is the third major contribution. The passively multiplexed Enhanced Transistor Electrical CD Metrology experiment strategy and complementary test structures make up the fourth main contribution. Finally the collaborative database is the fifth major contribution as it is a key enabler for multi-

student testchips as well as for high volume data analysis. Together these contributions have summed up to a capability for pushing variations upstream to circuit simulation and a process extraction strategy that can be used to characterize the lithography process with nanometer accuracy. This better understanding of how and why circuits vary is the basis for DFM where this information can be used to make process aware design decisions effectively.

2 The Parametric Yield Simulator

The Parametric Yield Simulator (PYS) is a comprehensive simulation flow from processing to circuit simulation that can keep up with processing technology, package process variations into a standard transistor model, and enable circuit designers to analyze circuit robustness. Process variations are translated into a language that the circuit designers can understand; so minimal understanding of lithography and other processing steps is necessary to understand their repercussions. The PYS is an open-ended platform that allows for improvements on the processing side (ex. new CMP models, etch models, litho models) and on the circuit side (ex. Process aware circuit analysis or yield optimization), so it can serve as a platform for creating new DFM strategies that break down the traditional wall of abstraction between process and design.

The automated flow from gds to HSPICE circuit was implemented using Mentor Graphics Calibre Work Bench, HSPICE, and a set of Perl and TCL scripts. Jie Yang has created a similar closed loop flow, but without non-rectangular transistor models, for advanced timing analysis using simulated CD values from Calibre Work Bench[18]. The non-rectangular transistor model is the basis of the PYS as it translates a 2 dimensional physical shape that is the output of Calibre Work Bench into an equivalent 1 dimensional BSIM model that serves as an input into HSPICE. The flow has been implemented in a

modular fashion with distinct breaks between poly gate extraction, transistor model generation, and circuit implementation (Figure 1). The modular fashion allows for improvements and changes to be made independent of the other modules. The processing module was built based on Calibre Printimage to model the lithography step at specific process conditions. The device module uses the simulated geometrical gate shape and builds an equivalent BSIM transistor model. Finally, the third module injects all the calculated transistor lengths into an HSPICE netlist for circuit simulation. The whole flow can be wrapped in Perl scripts for simulating circuits across a process window or with different process parameters. The key benefit or goal of the Parametric Yield simulator is to shift the statistics in statistical timing and power analysis upstream from gate level CD distributions to a set of characterized processes. In order to be able to systematically predict a majority of transistor performance variation the PYS is implemented as a platform that can adapt and expand as new sources of variation become more prominent. This chapter will give an overview of the Parametric Yield Simulator.

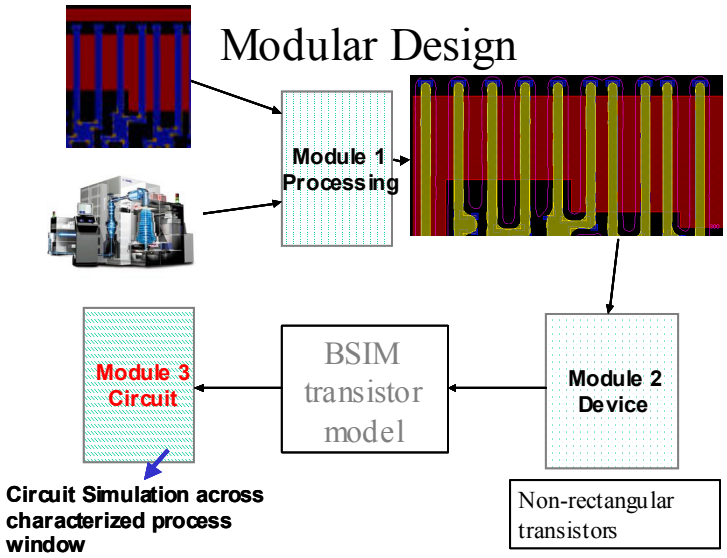


Figure 1 Parametric Yield Simulator modular implementation. Three distinct modules allow for independent improvements from the processing, device, or circuit side.

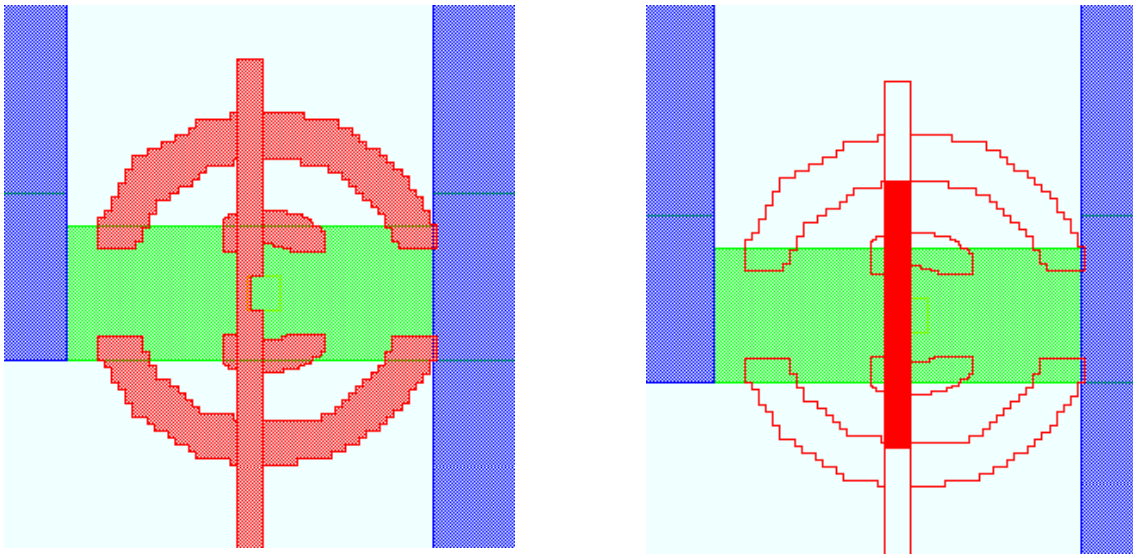
2.1. Processing Module

The processing module of the Parametric Yield Simulator has been built using Calibre WorkBench and Perl scripts for data manipulation. Calibre has a very robust TCL interface as well as their proprietary SVRF scripting language for manipulating layouts, applying model and rule based OPC, and running aerial image simulations. The processing module, as well as all scripts/modules in the PYS, takes as input a setup file that contains a list of file pointers to pertinent files, such as the layout gds file and transistor gate location file, and a set of parameters such as sampling frequency and range of gate lengths. Since each script has one or more files that will be used by subsequent scripts, a setup file saves the hassle of keeping track of appropriate filenames in each script.

The processing module consists of four scripts that eventually spit out a file that has all the slice dimension information as well as image parameters such as Image Log Slop (ILS) and contrast. The database stores the ILS, Imax (maximum intensity), Imin (minimum intensity), min CD, max CD, and stdev CD for each transistor, parameters that are calculated in slice_calc script. The first tcl script, get_gate_widths.tcl, in the processing module simulates the active region and the second perl script, update_widths.pl, adjusts the widths of each transistor to the simulated width. The third script, gen_m_run.tcl is the most time intensive step that simulates each gate in at least ten places. The ten edge slices are simulated at a sampling frequency of once every 10nm and the center slices are simulated at a set frequency in the setup file, default is 60nm. A sampling frequency needs to be chosen below the resolvable feature size so that all bumps and protrusions can be captured. Calibre spits out measurement results in a Calibre specific output file that is then parsed with a perl script, slice_calc.pl, that finds all the gate image statistics and reformats them into a format the device module can understand. The most important setup file parameter for the processing module is the Calibre setup filename, which contains optical and resist model information. Both the setup file and the gate locations file can be generated from the database, so the database can be used for simulating transistors of interest under specified process conditions.

An auxiliary script called fix_gds.tcl has also been created that sanitizes the gds file so that Calibre will not crash. Calibre Work bench first looks at the drawn layer to figure out where to simulate the image based on locations provided in an input file and then uses the post-OPC layer to run aerial image simulation. If the drawn layer is complicated, such as in a defocus monitor(Figure 2), or has a corner close to the point of interest, Calibre

crashes and complains it cannot find an appropriate edge. Hence fix_gds erases all polygons off the drawn layer and replaces them with perfect simple rectangles. This way Calibre does not get confused figuring out where to simulate, but uses the complex polygons on the post-OPC layer for simulation.



(a) (b)

Figure 2 Calibre tries to interpret the drawn layer and crashes with complex designs as seen in (a). The gds needs to be sanitized as in (b) where Calibre only sees the solid rectangle when looking at the drawn layer (which is NOT the layer used to calculate the aerial image)

The processing module is the key to the Parametric Yield Simulator as it allows for discriminately evaluating each transistor. The capability of simulating pattern dependent phenomena can show that a CD distribution that looks statistically more random then another can be deterministically predicted. Figure 3 shows a histogram of a CD distribution of four standard cells with 160 gates. Each gate was sampled at a frequency of at least one CD measurement per 60nm and the average CD was calculated. As the aerial image is simulated with no defocus and with 120 nm defocus, the standard deviation of the CD distribution doubles from 1.2nm to 2.6nm. CD measurements of one

sample might look more "random" than another, but if the process can be characterized and modeled accurately, then a portion of random CD noise should be explained systematically.

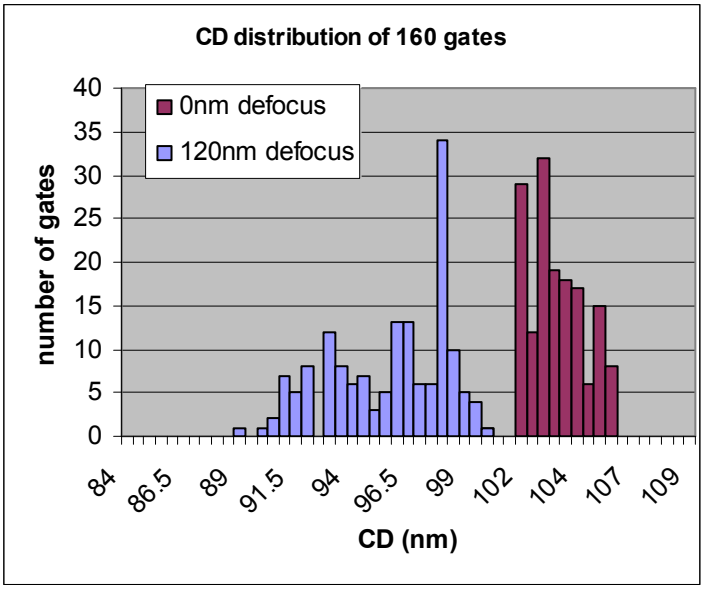


Figure 3 CD distribution of 160 gates at 0nm defocus and 120nm defocus. The stdev of the CD distribution doubles from 1.2nm to 2.6nm, yet this "extra" randomness can be systematically predicted.

On the other hand, systematic variation that is not predicted by an optimistic optical model, which does not account for across field variation, can also be accounted for in the simulator. Systematic intra die and inter die CD variation has been thoroughly studied and causes, such as dose non-uniformity across the slit, have been assigned (mostly of grating type structures that present little 2D proximity dependent variation)[19,20,21]. These variations are currently not modeled in any aerial image simulator as a field position dependent optical model would have to be employed, which can add significant computational complexity. A possible workaround is to overlay empirical data and add

an appropriate CD bias. This comprehensive functionality can be built into the processing module as long as the format of the output file, or input file for the device module, is maintained. This strategy can hence be used to improve the parametric yield simulator as whole without necessarily understanding the inner workings of the rest of the modules. Since final gate CD is a function of not only litho, but Post Exposure Bake, ashing, and etch, the processing module can be augmented with more simulators. Fortunately since wafer processing is a serial process, process simulators and models can be implemented serially without the need of building new simulators from scratch. As long as a list of slices is generated for the device module, more processes can be implemented for more accurate circuit simulation.

2.2. Device Module

The device module is the second phase of the Parametric Yield Simulator that acts as the critical link or bridge between process and circuit simulation. The goal of building the device module was to come up with a transparent wrapper for BSIM that takes as input a poly gate profile and translate it into an equivalent rectangular model that can be used in HSPICE. This is a critical link that can translate any simulated or measured gate shape into a model that can be used in HSPICE for electrical simulation. As the transistor model can be modified to account for pretty much all process non-idealities it serves as a perfect bridge between process and circuit simulation. In fact this strategy is the cornerstone of a sustainable DFM strategy as DFM needs to be evolutionary vs revolutionary in order to be economically viable. One of the biggest reasons for the exponential advancement in the IC industry over the last three decades is the layer of abstraction that exists between design and manufacturing that in turn leads to design

techniques and strategies that transfer from generation to generation. Designers develop a gut feeling for making designs based on experience of using the same or similar tools from generation to generation. DFM tools need to be built in a similar fashion where DFM aware design techniques do not change significantly when new processes are modeled. Designers need to build a gut instinct for making tradeoffs between performance and yield and that instinct should not have to change from generation to generation. Hence the metrics used for evaluating the susceptibility of a design to variations should stay the same independent of if the CD of a gate is changing or the strain in the channel is. Fortunately the transistor model can translate any and all transistor parameters into a set of currents and voltages that can then be used by design side tools for DFM aware layout optimization. Building this bridge between non-rectangular gate and equivalent BSIM transistor model is one of the defining strategies discovered during the implementation of the Parametric Yield Simulator.

2.3. Non-Rectangular Transistor Model

The device module translates a gate shape into an equivalent rectangular BSIM model by using an Equivalent Gate Length model and extraction technique (Figure 4). This method accounts for the short channel effect, punch through, the narrow width effect, and all other channel length dependencies that are modeled by BSIM. A critical assumption, which has been proved accurate in previous studies, is that a transistor can be split into independent rectangular slices[22,23]. This is valid as long as the spatial frequency of the modulation is low enough so that enhanced 2D diffusion effects do not come into play. This is the case if the correlation length of the modulation is greater then 50nm, not a problem when the minimum printable feature is greater than 50nm. Since all the

rectangular slices of the non-rectangular transistor are independent, it can be shown that the non-rectangular transistor will behave just like a rectangular transistor with an appropriate effective length.

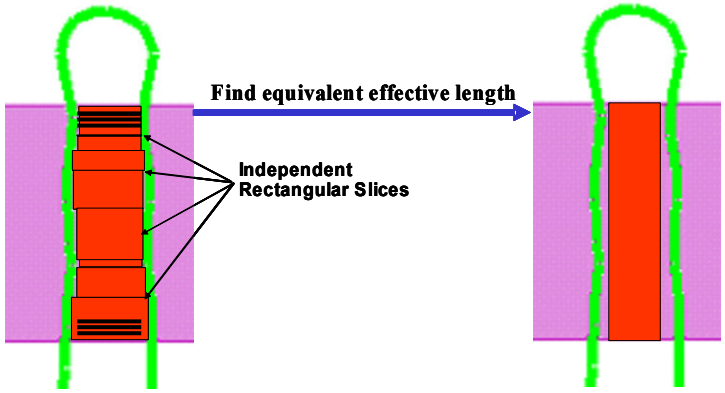


Figure 4 The Equivalent Gate Length (EGL) method approximates a nonrectangular transistor or set of independent rectangular slices with one rectangular transistor. The effective length of this rectangular transistor is a weighted average of all the slices.

The effective length is basically a weighted average of all the slices that make up the transistor, weighted by current values that are stored in a current lookup table. The current lookup table can be generated from a BSIM model, MEDICI, or silicon data. The current lookup table serves as another hook in the platform as processes can be modeled by building different lookup tables. For example strain could be modeled by building different lookup tables for channels with different levels of strain. If the process module can spit out strain values in the channel, the device module could be directed to use appropriate lookup tables and hence account for strain.

The details of this step are fully described in the next chapter. Two important points to note are that there will be a different effective length for delay and static power analysis

and that there needs to be a capacitance offset when looking at delay analysis. Once an effective length is found, it can be plugged back into the standard BSIM model to make it account for across gate CD variation. As there are many transistors in a circuit, a table of only the BSIM parameters that need to be changed is passed to the circuit module (currently the capacitance length offset parameter DLC and the gate length).

2.4. Circuit Module

Once an effective length is calculated for each simulated gate in the layout, it is plugged back into an HSPICE netlist in the circuit module. If the effective length is different then the average length for delay analysis, the capacitance length offset parameter DLC is also reset in the HSPICE netlist. The link between layout and netlist can be extracted from the placement DEF file, which maps HSPICE sub circuit names into standard cells in the layout. The current implementation of the PYS has each calculated effective length annotated with a corresponding line in the spice deck. Several perl scripts have been written that break apart a spice netlist into a list of transistor locations and corresponding spice deck lines. Sub circuits need to be copied and renamed as different instances of the same cell can have different gate dimensions. This is not always the case as sometimes many instances of the same cell have the same gate lengths, depending on the neighboring cells, cell size, as well as the illumination conditions. The primary job of the circuit module is to read in the original netlist and update the appropriate transistor BSIM parameters according to the output from the device module.

```
M0 gnd 7 Z gnd N L=60n W=5.85e-07 $X=485 $Y=1155
M1 8 A gnd gnd N L=60n W=3.2e-07 $X=1055 $Y=1260
M2 7 B 8 gnd N L=60n W=3.2e-07 $X=1435 $Y=1260
M3 vdd 7 Z vdd P L=60n W=1.05e-06 $X=475 $Y=3275
M4 7 A vdd vdd P L=60n W=3.8e-07 $X=935 $Y=3275
M5 vdd B 7 vdd P L=60n W=3.8e-07 $X=1505 $Y=3275
```

```
M0 gnd 7 Z gnd N L=62n W=5.85e-07 $X=485 $Y=1155
M1 8 A gnd gnd N L=60n W=3.2e-07 $X=1055 $Y=1260
M2 7 B 8 gnd N L=61n W=3.2e-07 $X=1435 $Y=1260
M3 vdd 7 Z vdd P L=60n W=1.05e-06 $X=475 $Y=3275
M4 7 A vdd vdd P L=62n W=3.8e-07 $X=935 $Y=3275
M5 vdd B 7 vdd P L=57n W=3.8e-07 $X=1505 $Y=3275
```

Figure 5 Example of an original spice deck and the subsequent process aware version. Notice the L parameter was changed in the bottom version.

2.5. Wrapper Scripts and Links to Database

The above described Parametric Yield Simulation flow starts at process simulation and ends with a process aware HSPICE netlist. In order to evaluate how the netlist or circuit performance changes across a process window a set of wrapper scripts were written to automatically re-run the simulation scripts under different process conditions. This way a new set of effective length can be generated for each process condition. Run_batch.pl reads in a process_conditions input file one line at a time and runs the PYS with given process conditions, one set per line. Each script is checked for completion and if the PYS crashes in one section the whole program exits. This is necessary as the PYS

can continue running even if one of the sub-scripts fails. Simulation results are automatically uploaded to the database through a script, convert_tab_delim_to_sql.pl, that converts tab delimitated output from the PYS into an SQL input file. Since the database resides on a relatively slow server, there is a benefit to running the PYS on a separate faster server. For this reason a simulation monitoring script, monitor_folder.pl, on the database server was written that checks for output files from the PYS and uploads them to the database. This hand shaking through output files allows for multiple instances of the PYS to run at once on separate servers and the database on the central server can be updated automatically. Hence the PYS is in a sense parallelizable as long as the job is intelligently split up by putting either different process condition settings or transistor locations in different input files. As aerial image simulation is computationally intensive this is a handy feature.

2.6. Discussion

The Parametric Yield Simulator has been built as an initial DFM tool for analyzing how circuit performance changes at different process conditions and as a platform for developing new DFM tools or capabilities. It was built to address the fact that not all transistors are created equal and that each transistor needs to be evaluated with a process model or simulator. Figure 6 shows 160 gates in four standard cells simulated through focus. Some gates vary a lot through focus and some don't. The wide spread of gate variation is an indication of how a set of design rules cannot capture all process information and hence large guard bands need to be used to guarantee that everything will work. This inefficiency is addressed by using a process simulator to evaluate

transistor performance variation as a function of the layout that surrounds it and a set of characterized processes.

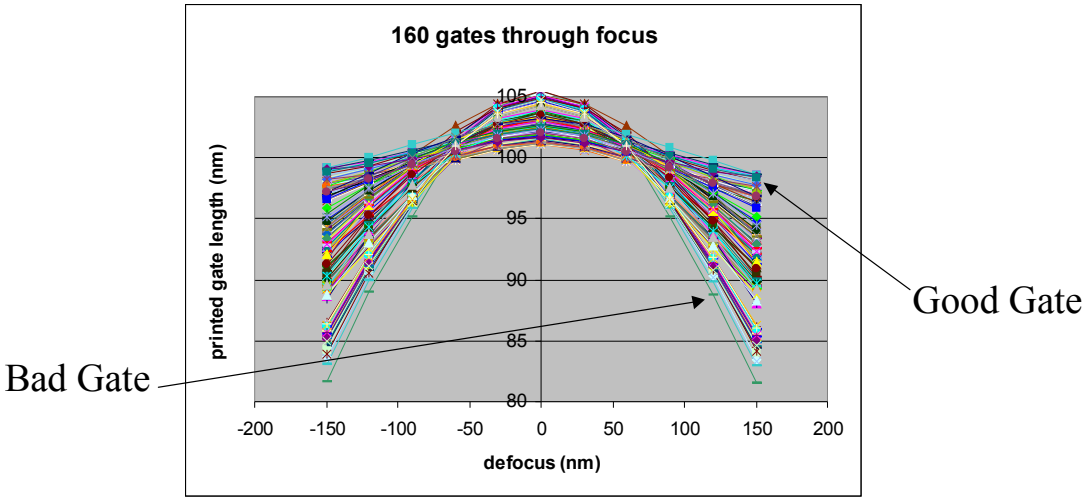


Figure 6 Not all transistors should be treated equally. Some gates are stable through focus while other have 4X more variation. These are simulation results of 160 gates in four different standard cells.

Linking litho simulation to transistor performance enables process aware circuit simulation, analyzing experimental results, rapid prototyping and testing of process sensitive test structures, and running dry-lab experiments of planned testchips. These uses of the Parametric Yield Simulator will be described in detail in subsequent chapters. The heaviest use of the PYS was for a massive simulation dry-lab experiment of over 15,000 test structures. As a platform the PYS stands to become a center for collaboration between circuit designers and technologists that together can make the PYS more robust, more efficient, and have higher impact.

The process side can be expanded to become either more efficient by using faster lithography simulators or more robust by modeling more process steps. Currently the process module uses a production grade lithography simulator that is slow and can get

slower if more accurate empirically calibrated optical models are used. Running small sets of test structures may take a couple minutes per process condition, but given a chip that takes several hours for OPC at nominal process conditions, it would take days to simulate it across an entire process window. Using fast pattern matching techniques or other fast litho simulation methods could greatly increase the speed of the PYS. A potential solution would be to use a high speed simulator to identify hotspots or areas of interest and then only simulate those areas with Calibre WB. Another dimension in which the process module can be expanded is to model etch. The output from the litho module could be fed into an etch simulator that would model pattern dependent etch biases. As long as the output from the fast litho module or etch module are consistent with the standard input to the device module, the PYS can be expanded independent of the device and circuit modules.

Modeling other process effects such as strain or Rapid Thermal Anneal (RTA) non-uniformity will require adjustment to the process and device modules. Both strain variation and RTA variation have been shown to exhibit significant pattern dependent effects and hence would be valuable additions to the PYS[24,25]. The most convenient, although probably not the most elegant, implementation would be to build a massive set of slice lookup tables for effective length extraction. These lookup tables could be build from HSPICE simulation with intelligent modification of BSIM parameters or from a 2D or 3D process/device simulator. Slice lookup tables could then be selected based on simulated or modeled process parameter values for each transistor. This would work for RTA which generally has a fairly large correlation length, but strain variation may prove a challenge as strain may vary across the gate. In which case either different slice lookup

tables would need to be referenced for different slices or a completely different transistor modeling strategy would need to be employed. In either case, the output from the device module is always going to stay consistent as a BSIM transistor model. This is the key to enabling advancement on the circuit side.

Since the input into the circuit module is always going to be a list of transistor BSIM models, it is possible to develop DFM tools independent of how the process and device modules evolve. The circuit module is the central place where process aware transistor models can be used to evaluate circuit performance, which can then be used to generate a robustness metric. A robustness metric that is based on performance variation across a set process window could be used for quantifying parametric yield and be used for making design tradeoffs. The current standard metrics in design are power, delay, and area. A robustness metric could be used as a lever to influence a process aware synthesis flow. A combined area/yield metric has already been proposed for technology mapping[26]. A similar approach could be taken in placement algorithms where nets could be weighted not only by distance and load capacitance, but the amount of variability in the driving cell. This way cells that have significant drive current variation could be neutralized by minimizing wire loading and resistance so that total delay does not vary as much. In any case, as long as improvements in the circuit module use process aware transistor models, it can evolve almost independently of the rest of the Parametric Yield Simulator.

2.7. Conclusion

The Parametric Yield Simulator has been built as a scripted link between process and circuit simulation. It stands as a platform for collaboration and linking process models and simulators with design based DFM tools. A key concept of this simulation flow is to

shift the statistics in statistical timing and power analysis upstream from gate level CD distributions to individual process steps that can be characterized with the right set of test structures. Statistical timing analysis can prevent overly pessimistic assumptions made by using a corner model, but still does not use all possible information when treating each transistor equally[27]. This leaves a lot of margin on the table as not all transistors are created equally. Pattern dependent variations can be either mitigated in critical areas or avoided in non-critical areas. The results of which are tighter margins that allow more aggressive designs.

Instead of assuming a statistical distribution of a billion gates in a circuit, one can characterize a small set of process steps and see how the gates print at various processing conditions. These processing steps can be characterized by special characterization structures or tests. Then predicted CD distributions will have defined sources of variation and solutions to potential hotspots in the layout can be attacked through redesign, improved OPC, or better process control. The key to understanding what CD values lead to actual hotspots and yield degradation is to use them in circuit simulation. The simulated leakage power and delay values can be used to evaluate the robustness of a circuit or the severity of simulated hotspots. It is important to evaluate a simulated layout in circuit terms, as large CD variations do not necessarily lead to large leakage power and delay variations. Leakage power and delay variations largely depend on the load capacitance and the amount of stacked transistors between the supply rails, factors that are accounted for during circuit simulation. With enough information fed upstream to designers, circuit designs in the variation heavy nanometer era can be more efficiently implemented.

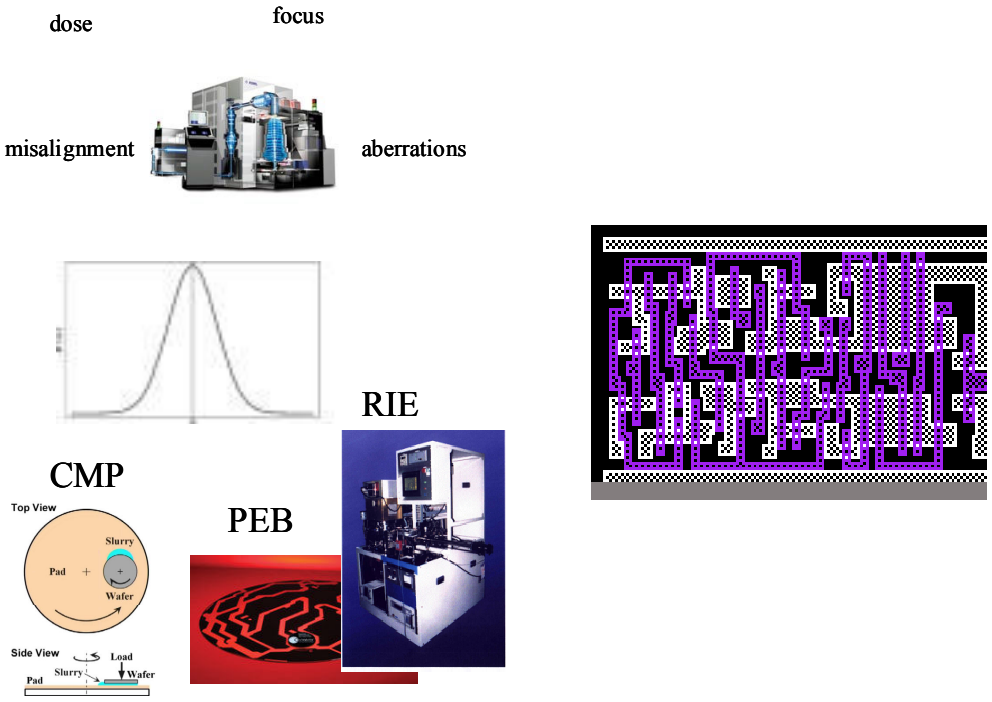


Figure 7 Inject variation upstream. Instead of assuming a generic statistical distribution for all 100 million transistors, statistics can be applied to a handful of process steps that can be characterized with high volume process characterization.

3 Non-Rectangular Transistor Model

As the k1 technology factor goes below 0.5, 2D proximity effects become more pronounced and across gate CD variation starts becoming significant. A basic principal behind Design For Manufacturing or DFM is to communicate processing knowledge into the design phase, hence being able to characterize a non-rectangular 2-dimensional transistor in a compact 1-dimensional transistor model is critical[28]. This is a complex problem as performance parameters such as threshold voltage and leakage current have a very complex non-linear relationship with gate length. The challenge with correcting the previously accurate assumption that the gate is rectangular, is that a comprehensive new model would necessitate new standards as well as changes to circuit simulators, which would have to start solving new sets of equations. The approach presented in this chapter leverages the channel length dependent information stored in the BSIM model to translate a non-rectangular transistor into an equivalent rectangular transistor that does not require any new transistor models or HSPICE simulators.

3.1. Background and Motivation

To address this problem, several authors have proposed breaking up each non-rectangular transistor gate into a set of parallel transistors [Figure 8][29,30,31,32]. The basic premise behind this approach is that a non-rectangular transistor can be broken up into independent rectangular slices, which has been shown to be true for low frequency

variation by Shiying Xiong[33,34]. This would account for across gate CD variation and would not necessitate a complex gate profile description in the transistor model. Although this approach can be effective, it suffers from two significant drawbacks. First of, it increases circuit complexity. Each transistor could potentially be replaced with multiple transistors in parallel, which would increase the transistor count significantly. Secondly, there is no compact model for transistor slices. This is mainly due to the narrow width effect, which is accounted for in the BSIM model. The narrow width effect with shallow trench isolation (STI) is an effective lowering of the threshold voltage from fringing electric fields between non-overlapping poly and the channel[35]. Since different slices will be a different distance from the edges of the transistor and will have widths less than the minimum width allowed in BSIM, the standard transistor model would not work for slices. Hence a new strategy is proposed.

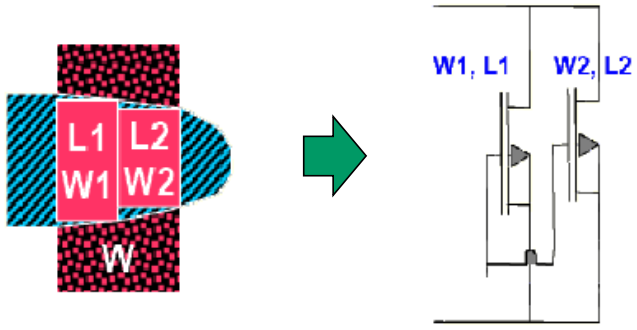


Figure 8 One proposed method of modeling non-rectangular transistors is by modeling them with a set of parallel rectangular slices or transistors.

This chapter proposes a new solution that translates a list of lengths and widths of rectangular slices that make up the gate into one equivalent rectangular transistor with an

effective length[Figure 9]. The effective length is in essence a weighted average of all the slices where each slice weight is based on a slice lookup table that is either built from HSPICE or device simulation or empirically. The calculated effective length can then be used as the gate length in the standard BSIM model for more accurate circuit simulation. The method proposed accounts for the narrow width effect, short channel effect, punch through, and all other channel length dependencies that are modeled by BSIM. The only tools necessary for effective length extraction are a circuit simulator such as HSPICE, a fitted model for the narrow width effect for slices at the edges of the gate, and an industry standard transistor model such as BSIM (note: BSIM is only one industry standard, but this approach should work with other models). The only caveat is that there will be a different effective length when looking at leakage current, Ioff, for static power analysis and when looking at drive current, Ion, for delay analysis. There is also a capacitance offset that needs to be accounted for in the case of delay analysis. This approach does not increase circuit complexity and is extremely fast as it is based Perl scripts zipping through lookup tables and comparing values.

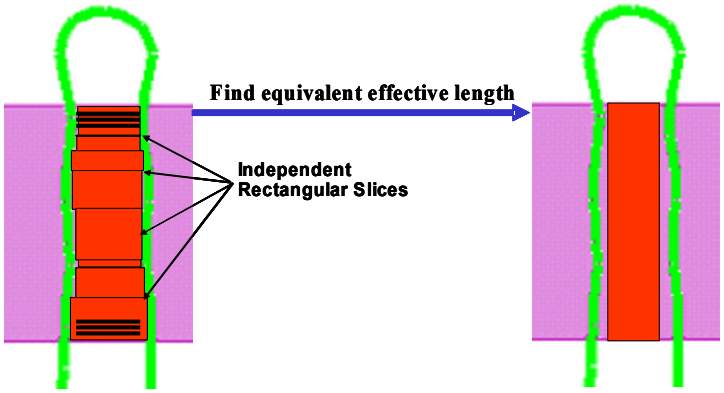


Figure 9 The proposed method approximates a nonrectangular transistor or set of independent rectangular slices with one rectangular transistor. The effective length of this rectangular transistor is a weighted average of all the slices.

This paper will first describe how slice currents are extracted from a wide transistor and then used to calculate the current through a non-rectangular transistor, which can then be used to find an equivalent rectangular transistor with an appropriate effective length. Then a set of simulation results will show how all the complex length dependent effects are conserved when comparing a non-rectangular transistor to an equivalent rectangular transistor. Finally a metric for evaluating the need for using a non-rectangular transistor model will be presented.

3.2. Modeling Approach

The basic idea behind building a BSIM model for a non-rectangular transistor is adjusting the gate length to an effective length or a weighted average of the rectangular slices that make up the transistor gate. Since the relationship between threshold voltage and channel length is complex and non-linear, an analytical solution would be very complicated if not impossible to find. Circuit simulators, like HSPICE, that use industry standard transistor models already account for all the complex effects and can be used to assign an appropriate weight to each slice. Effective length extraction consists of building a lookup table of currents for rectangular transistors and for transistor slices and then matching the total current through the non-rectangular to an equivalent rectangular transistor. Since transistor slices are independent, their currents are additive. Calculating a non-rectangular transistor's drain current is a matter of adding up the currents going through all the slices. Ioff or leakage current is used for finding an effective length for static power analysis and Ion or drive current should be used for finding an effective length for delay analysis. The difference in effective lengths should not be a problem as different tools are used for static power and delay analysis. All the currents for slices as

well as rectangular transistors should be extracted from the same pre-compiled current lookup table so that slices are weighted the same way. Once an effective length is found, it can be plugged into each transistor definition independently in the HSPICE netlist. This method employs the transistor model's capability of modeling any length dependent effects that make it very hard for finding an analytical solution for Leff. Hence this approach should be accurate as long as the model is valid.

3.2.1. Build a slice current lookup table

The first part of finding currents through rectangular slices and transistors and eventually non-rectangular transistors is building a lookup table of currents for center and edge slices. Slice currents are calculated by simulating a transistor and a slightly wider transistor (W_{TRAN}+W_{SLICE}) and looking at the current difference. Slice currents are then calculated for the middle of the transistor as the difference in the currents of the two transistors. Edge slice currents, that only make up the outside 40nm-50nm of the gate, are calculated based on an exponentially decaying edge slice model that is fit to match BSIM model currents. BSIM model currents are calculated in HSPICE and a slice lookup table is built for a range of gate lengths.

Since there is significantly more current at the edges of the transistor due to the inverse narrow width effect (INWE)[36,37], edge slices need to be treated differently than center slices, which have the same current density. The INWE is primarily due to the fringing electric field from non-overlapping poly in an STI process. LOCOS isolation technology has a different effect, called the Narrow Width Effect (NWE), that increases the threshold with a narrower transistor, but will not be considered as current aggressive lithography technologies require the planarity of STI[38]. The surface potential and hence

threshold voltage is a complicated function of the distance from the edge, but can be approximated with a linear fit. Since the threshold voltage offset is proportional to the distance from the edge, the extra leakage current falls off exponentially, which means a majority of the contribution will be in the first couple slices. This effect attenuates drastically over a distance approximated by $[e_s\ t_{ox}\ x_{dep} / e_{ox}]^{1/2}$, ($e_s$ permittivity of silicon, t_{ox} is the oxide thickness, x_{dep} is the depletion depth, and e_{ox} is the permittivity of oxide) which is around 25nm for a 65nm process. At double this distance the INWE becomes insignificant[39]. Hence the outside 50nm of the gate is sampled at a higher frequency, nominally 1/10nm, and the rest of the transistor is sampled at a specified frequency, nominally 1/60nm as that is below the resolution limit of a 193nm scanner so no deviation will be missed at this resolution. Each slice length is calculated as the average of the two outside points of the slice. Given this discrete sampling of slices it is easy to fit a single variable exponential to the extra current through the edge slices.

$$\Delta V_T = -a(w-x)$$

$$V_{T\_EDGE}(x) = V_{T\_center} - a(w-x)$$

$$I_{OFF\_EDGE}(x) = I_0\, e^{\frac{V_T + \Delta V_T}{n\,kT/q}} = I_{OFF\_Center}\, e^{\frac{a}{n\,kT/q}(w-x)} = I_{OFF\_Center}\, e^{b(w-x)}$$

Where b is the fitting constant, w is the distance of influence, 50nm, I_{OFF_Center} is the current through a 10nm slice in the center of the transistor (calculated as the difference in current between two wide transistors that have a width difference of 10nm), n is the swing curve constant[35], and x is the distance of the center of the slice from the edge. The fitting constant can then be fit by solving the following equations

$$I_{OFF\_EDGE}(x) = I_{OFF\_Center}\, e^{b(w-x)}$$

$$I_{OFF\_TOTAL} = \underbrace{2 * \int_0^w I_{OFF\_Center}\, e^{b(w-x)}}_{\text{Current through ten edge slices}} + \underbrace{\frac{(W_{TOTAL} - 2 * w)}{W_{SLICE}} * I_{OFF\_Center}}_{\text{Current through center}}$$

w=50nm

W_{SLICE} *= 10nm*

W_{EDGES} *= 100nm*

W_{TOTAL}*= Width of transistor*

I_{OFF_TOTAL} *= Simulated current of wide transistor*

I_{OFF_Center} *= Simulated current through one 10nm slice*

Once the fitting constant, b, is found then a slice current lookup table can be built for center slices and the five edge slices.

3.2.2. Calculate the current through the non-rectangular transistor

Once the current lookup table is built, each transistor gate needs to be broken up into rectangular slices and the total current can then be calculated. The poly gate profile is provided by the Process module of the Parametric Yield Simulator and can then be broken up into either slices of a specific width or slices of different widths, as long as all the modulation across the gate is captured [Figure 10]. Sampling at a set slice width can guarantee that all the modulation is captured if the width of the slices is less then the minimum resolvable feature in the scanner. In general, a good strategy is to take two measurements per slice and take the average, if the two measurements vary a lot, the slice

can be broken down into two slices. On the other hand, slice extraction can be practically free if fragmentation used for OPC (Optical Proximity Correction) is used to extract slices. This allows for using EPE (Edge Placement Error) data from OPC for slice extraction, which is supposed to capture any deviations if the fragmentation algorithm is adequate. Finding the current through the non-rectangular transistor is then a matter of summing the currents through all the extracted slices (Figure 10).

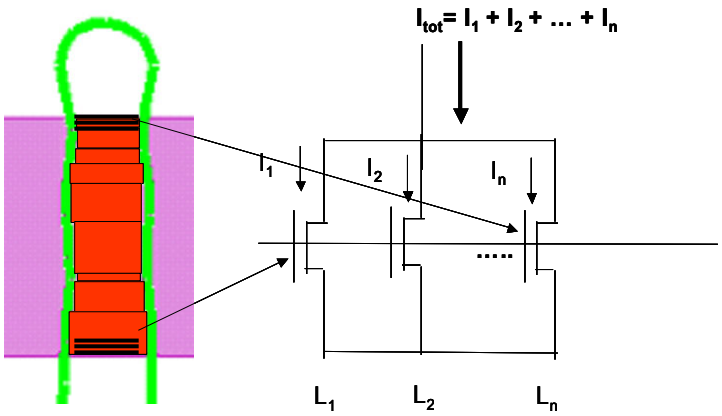


Figure 10 The current through the non-rectangular transistor is the sum of the currents through all the slices.

3.2.3. Find an equivalent rectangular transistor with the same current

Once the current through the non-rectangular transistor is extracted, an equivalent rectangular transistor can be found by using the same approach to calculate currents through rectangular transistors and matching currents. To build a current lookup table for rectangular transistors, sum up all the edge slice currents, multiply them by two, and scale the center slice current appropriately so the sum of all slices is the width of the rectangular transistor. Use the I_{OFF} current lookup table for static power analysis and Ion

lookup table for delay analysis. Once the closest current value is found, the effective length can be interpolated between the two closest points. Since the current lookup table captures all the short channel effects that are capture in the transistor model, all the slices are weighted appropriately. Making the BSIM model accurate for a non-rectangular transistor now is a matter of replacing the gate length parameter with the calculated effective length and adjusting the capacitance offset parameter in the case of delay analysis. As drive current has a much weaker relationship with gate length using the average length is accurate at least down to the 65nm node, but if an effective length extraction leads to a different length there will be a capacitance offset between the non-rectangular transistor and the equivalent gate length transistor. If this is the case the BSIM parameter DLC parameter needs to be offset by the difference between the average length and the effective length. This can be done in the HSPICE netlist where any BSIM parameter can be redefined for each instance of any transistor.

3.3. Results and Analysis

To evaluate the necessity for a non-rectangular transistor model a small set of standard cells that represented a broad sampling of different types of layouts were selected from a 65nm AMD standard cell library. A non-optimized OPC algorithm was used to increase the amount of across gate CD variation and increase the significance of the non-rectangular model. HSPICE v2004.09-SP1 and the BSIM SOI v 3.2 model were used to build a lookup table of drain currents for transistors ranging from 50nm up to 140nm in gate length. With the lookup tables built and transistor slices simulated, Leff can be extracted almost instantaneously. Over 99% of the runtime is spent running aerial

image simulation, so runtime for equivalent gate length extraction is negligible. All the table lookups and current comparisons were implemented in Perl scripts.

Looking at leakage current, or static power analysis, a plot of leakage current vs Vds, or the drain to source voltage can be seen in Figure 11a. This is for a simulated transistor profile with a large amount of across gate CD variation. To help quantify the amount of across gate CD variation, the ratio of the STDEV of slice lengths to the average slice length, STDEV/AVE, is used. This plot is for a transistor with an average length of 65.12nm and a STDEV/ave of 5.5%. This translates to an effective length of 62.25nm or a shift of 3nm. There are three curves corresponding to slice currents summed up at each operating condition (the actual current through the non-rectangular transistor), an equivalent gate length transistor and a transistor with a gate length equal to a simple average of all the slice lengths. Taking a simple average leads to a 50% error in the leakage current, but the equivalent rectangular transistor fits very well. The model fit gets worse at lower Vds values due to a decreased short channel effect. The short channel effect is a function of the electric field between the source and the drain, if the Vds voltage decreases, then the short channel effect becomes less prominent and the shorter slices no longer have such a heavy weight, so the effective length changes slightly. This is not an issue as all the transistors that contribute to leakage current have either a full VDD across them or a large fraction of VDD in the case of stacked transistors. Hence this model is accurate for static power analysis. Figure 11b shows the same three curves for a hypothetical transistor with an extreme amount of across gate CD variation, 25% STDEV/AVE. The model still works, which indicates that this approach should be accurate for future generations when across gate CD variation gets much worse.

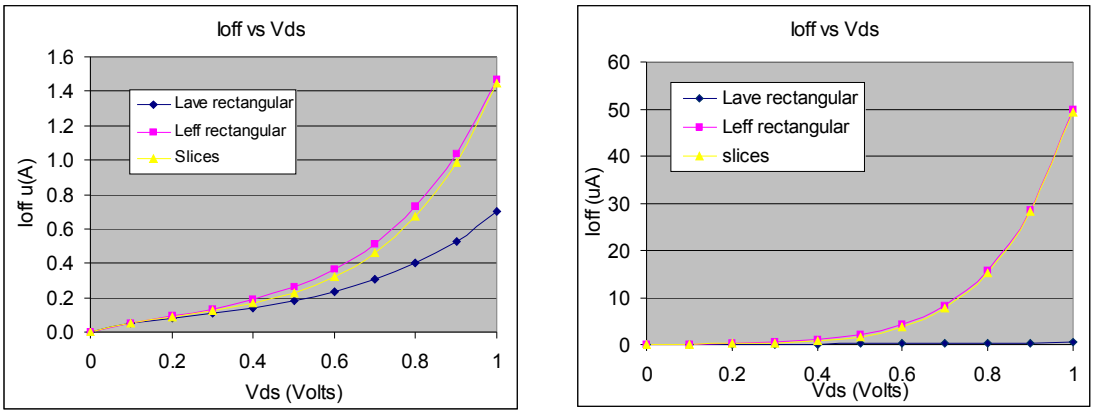


Figure 11 Plots of leakage current vs Vds (drain to source voltage) for three different transistors. One curve is for a non-rectangular transistor, one is for the equivalent gate length transistor, and one is for a rectangular transistor with a simple average as the gate length. (a) Lave = 65.12nm Leff = 62.25nm STDEV/AVE = 5.5% (b) This is a hypothetical transistor with extreme across gate CD variation Lave = 65nm Leff = 54.2nm STDEV/AVE = 25%

A set of 2066 65nm poly gates were simulated and their effective lengths were extracted [Figure 12]. The OPC algorithm was not optimized in an effort to create a lot of across gate CD variation. A set of hypothetical Gaussian slice distributions was also created to create a standard reference. Most of the simulated gates fall below this Gaussian standard plot. This is due to the fact that the standard deviation of most poly profiles is driven by several wide slices at the end of the transistor that correspond to corner rounding. Since wide slices have a less significant effect than short slices, they boost up the STDEV/AVE, but do not affect the shift as much. In either case, a significant amount of variation, ~4%, needs to exist to get even a 1nm shift for a 65nm gate length. A 4% STDEV/AVE translates to a 7.5nm 3σ edge placement error, which is much higher than predicted by the ITRS roadmap for 65nm gate[40]. This is somewhat

counterintuitive as leakage current has a very strong exponential relationship on channel length, which would indicate a larger shift. This begs the question of when a non-rectangular transistor model is needed.

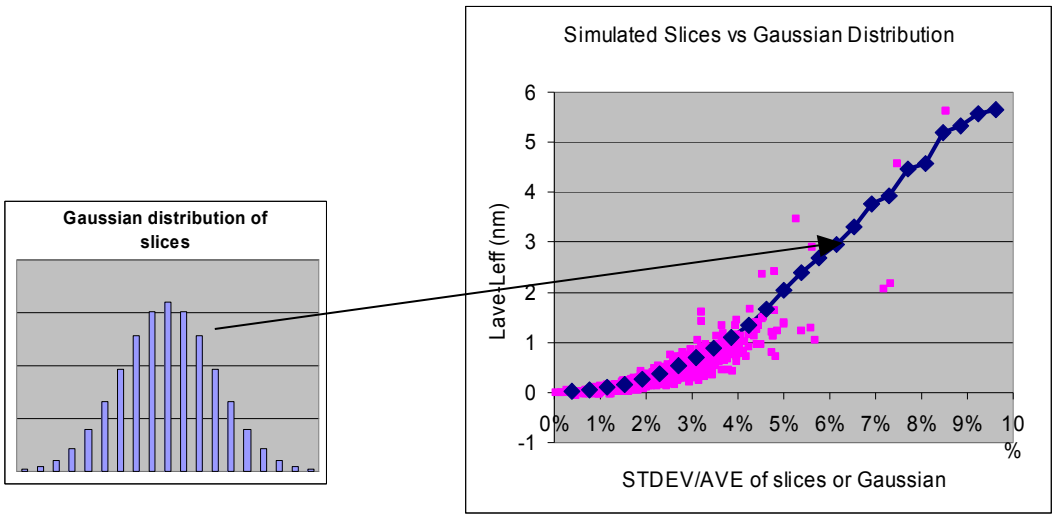


Figure 12 Plot of Leff-Lave for 2066 simulated poly gate profiles and for a set of hypothetical transistors with a Gaussian distribution of slices. The curve is from Gaussian transistors and the points are from simulated poly images in Calibre PrintImage.

3.4. Discussion

The need for using a non-rectangular transistor model can be evaluated based on the difference between the effective length and a simple average length. The difference between the effective length and the average length was found to be dependent on two major factors: the level of across gate CD variation and the transistor design or the relationship between leakage current and gate length. Unfortunately both of these factors will vary widely depending on the quality of OPC, device design, process flow, as well as model fit, so setting a specific threshold like a specific technology node would be

ungrounded. Each process or technology will have to be evaluated based on the amount of across gate CD variation and the transistor design. The most important indication is how much the curve of leakage current vs gate length deviates from a line over the range of CD values that are expected. If leakage current scaled linearly with the channel length, or the curve was a perfect line, then the effective length would be a simple average of the measurements and a non-rectangular model would not be necessary. This can be seen from some simple math shown below.

$I_{non\text{-}rect} * w_{non\text{-}rect} = I_1 * w_1 + I_2 * w_2 + \ldots I_n * w_n$

Assume I is linear with L so $I = a * L + b$

$\mathbf{I_{non\text{-}rect} = a * L_{eff} + b}$ $\qquad$ $\mathbf{I_m = a * L_m + b}$

$(a * L_{eff} + b) w_{non\text{-}rect} = (a * L_1 + b) * w_1 + (a * L_2 + b) * w_2 + \ldots + (a * L_n + b) * w_n$

$a * L_{eff} * w_{non\text{-}rect} + b * w_{non\text{-}rect} = a * (L_1 * w_1 + L_2 * w_2 + \ldots L_n * w_n) + b * (w_1 + w_2 + \ldots w_n)$

$a * L_{eff} * w_{non\text{-}rect} = a * (L_1 * w_1 + L_2 * w_2 + \ldots L_n * w_n)$

$L_{eff} = \mathbf{(L_1 * w_1 + L_2 * w_2 + \ldots L_n * w_n) / w_{non\text{-}rect}} = L_{ave}$

If we approximate the leakage current vs CD as a line, then the effective length will always be a simple average of the currents. Any curve becomes a line if the input range is very small, so the significance of using a non-rectangular transistor will depend on the level of across gate CD variation and the transistor design. Depending on how much the curve deviates from a line over the range of CD values that you expect, then it will be

equivalent to creating a different slope value, 'a_1', 'a_2',...,'a_n' or weight, for each slice. Since the curve is exponential, smaller slices will have a higher slope or bigger weight and the wider slices will have a shallower slope or a smaller weight. Although most Vt rolloff curves look very non-linear, the x-axis generally spans over a micron. If you zoom into a 15nm section, then the curve looks a lot more like a line [Figure 13]. Hence the need for a non-rectangular transistor model is questionable in the case of AMD's 65nm transistors, which where the subject of this study.

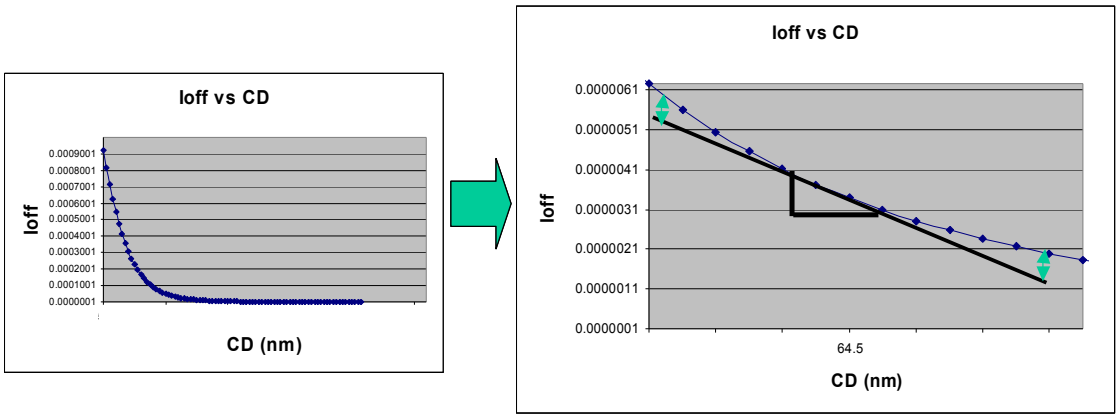


Figure 13 The significance of the non-rectangular transistor model can be evaluated based on how much the Ioff vs CD plot deviates from a line across a range of CD values that are expected.

Looking at delay analysis, the need for a non-rectangular model is even less urgent. This is due to the much weaker non-exponential relationship between drive current and channel length. Drive current will vary 5X while leakage current will vary 5000X over the same range. Nonetheless, to test the accuracy of our model for delay analysis, an 11-stage ring oscillator was created with three separate types of inverter stages. The error in delay was calculated by comparing ring oscillators built with very wide parallel flat slices/transistors and ring oscillators built with rectangular transistors with either an

equivalent gate length or an average gate length. For equivalent gate length transistors the DLC BSIM parameter(Channel-length offset parameter for CV model)[35] had to be adjusted by the difference between the area of the non-rectangular transistor and the equivalent gate length transistor (L_{ave} * W – L_{eff} * W). The model proved to be very accurate for large amounts of across gate CD variation with an error less than a fraction of a percent. It should be noted that even using an average length yields only a 7.5% error for a STDEV/AVE of 15% or 3σ edge placement error of 29nm for a 65nm gate. This brings about the point of using a non-rectangular transistor model for active power analysis. Active power dissipation is a function of load capacitances and supply voltages and to a lesser degree delay, as faster transistors can lead to a higher operating frequency, which will increase active power dissipation. Hence a non-rectangular transistor should be even less important for active power dissipation analysis than for delay analysis and is therefore not addressed. To analyze total power dissipation, only the static power portion of total power analysis will have to be changed to account for non-rectangular transistors.

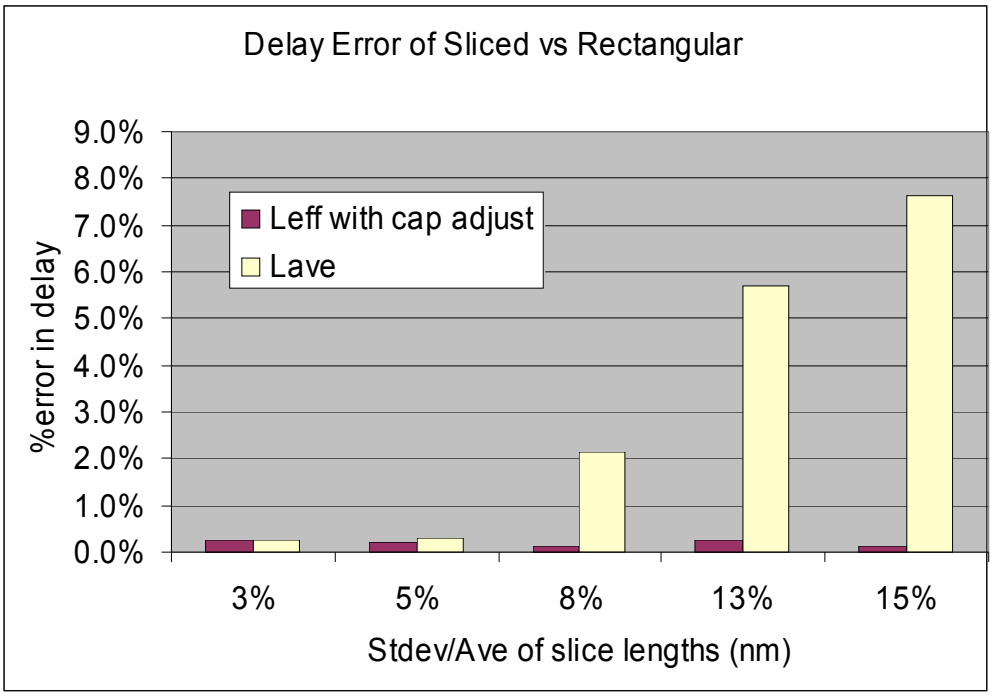


Figure 14 Error in delay when taking a simple average of the gate length instead of an equivalent gate length that weighs all slices based on simulated currents. Notice 15% Stdev/Ave for a 65nm transistor means a 3 sigma LER value of 29nm, hence a simple average of slices should be accurate for delay analysis.

As far as the limits of this model, it is expected to be accurate as long as the assumptions made result in second order or insignificant effects. The primary assumption of all slices being parallel is mostly valid for acceptable levels of LER and reasonable layout geometries. Most non-rectangular gates will have most deviations from corner rounding, necking, and pinching from proximity effects, which are all symmetric effects or will effect both edges in roughly the same place. In all these cases the shortest distance between two points in the channel is directly across the channel. If another source of non-rectangularity is present where a diagonal line would create the shortest distance then the independent slice assumption will fail. If this diagonal distance leads to the smallest effective length, then this current could dominate in the real transistor and lead to an error

in this model. This type of potential error could be found by looking at the edges independently, such as with edge placement error (EPE) data generated by optical rule checking.

3.5. Conclusion

The equivalent gate length approximation of a non-rectangular transistor using the model proposed in this chapter is accurate to within 1% for large amounts of across gate CD variation, >20% STDEV/AVE. The approach described accounts for short channel effects and the narrow width effect and does not require new models to be generated or circuit simulators to be built. This method does not increase circuit complexity and is based on text comparison using Perl scripts, which is extremely fast. The approach calls for approximating a non-rectangular transistor with an effective length that is a weighted average of all the corresponding slices. The weights of the slices are proportional to slice currents that can be extracted from HSPICE simulation or silicon data. Since BSIM models are fit to silicon data, all complex short channel effects are accounted for. Given a good transistor design that is on a relatively linear portion of the VT rolloff curve and acceptable amounts of across gate non-uniformity, the non-rectangular transistor model may not be necessary. The need for moving to a non-rectangular transistor model can be evaluated by the linearity of the leakage current vs gate length plot for the range of gate lengths that can be expected across one non-rectangular gate.

4 FLCC Enhanced NMOS Testchip

To complement the Parametric Yield Simulator, the second portion of the Collaborative Platform for DFM is a set of multi-disciplinary multi-student experiments aimed at identifying and characterizing the main sources of Vt variation. Six students and three industrial partners contributed designs that made up over 15,000 individually probable test structures. The main strategy behind the Enhanced NMOS testchip was to include a wide array of test structures as to capture any and all sources of transistor performance variation. The goal was to have six students contributing six sets of test structures and forming six sets of conclusions all on one chip. Conclusions would have to reinforce each other and hence become stronger. Confounding effects can be filtered out as different test structures are designed to have varying sensitivities to different processes. As long as the response of all test structures is unique to a given process, that process can then be characterized.

4.1. Experiment Description and Design Strategy

The multi-student test chip has over 15,000 individually probable transistors and test structures that were designed for an Enhanced NMOS Process Flow at SVTC (Silicon Valley Technology Center). Each structure is electrically probable via an automatic probe station using either standard transistors, Enhanced Transistor Electrical CD Metrology or ELM. The ELM structures are for characterizing systematic CD variation and Enhanced

Transistor based test structures are designed to study defocus, dose, misalignment, mask errors, corner rounding, Random Dopant Fluctuation (RDF), Line Edge Roughness (LER), and other random effects. A standard 30-pad cell was used for all test structures for probing simplicity (Figure 1). It is worth noting that this standardization of designs proved to be very useful in creating a multi-designer test chip as scripts could be easily created for combining data. The 30-pad cell served as the basic unit of real estate and as long as each designer followed a specific naming convention, a tcl script could automatically combine all the designs into one chip. In addition to measuring standard and enhanced transistors, cryogenic testing could be used to study and filter out any dopant related threshold voltage variations. This shotgun approach allows for exploring the unknown and identifying and quantifying the effects of different levels of process non-idealities.

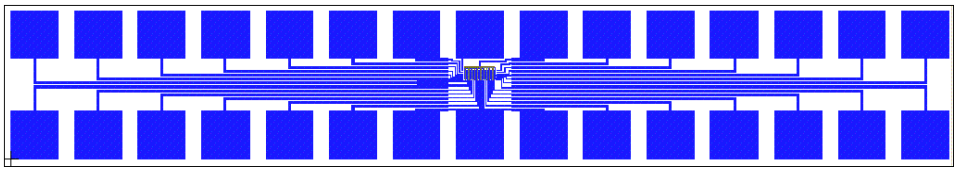


Figure 15 Standard 30-pad cell. All test structures have been hooked into a 30 pad cell for easy probing.

4.1.1. Passive Multiplexing Strategy

In order to fit over 15,000 individually probable transistors a multiplexing strategy had to be devised that enabled individual addressing up to 196 with only 30 pads and no logic. There are two basic strategies, the series transistor method uses 14 drain pads, 14 gate pads, 1 source pad, and 1 body pad and the parallel transistor method uses one gate

pad, 14 drain pads, 14 source pads, and one body pad. Essentially each intersection of a source/drain pair with a gate becomes a transistor that can have a completely unique proximity. Since none of the signals are routed through logic there is no need for high Vt transistors that would be necessary for minimizing parasitic leakage.

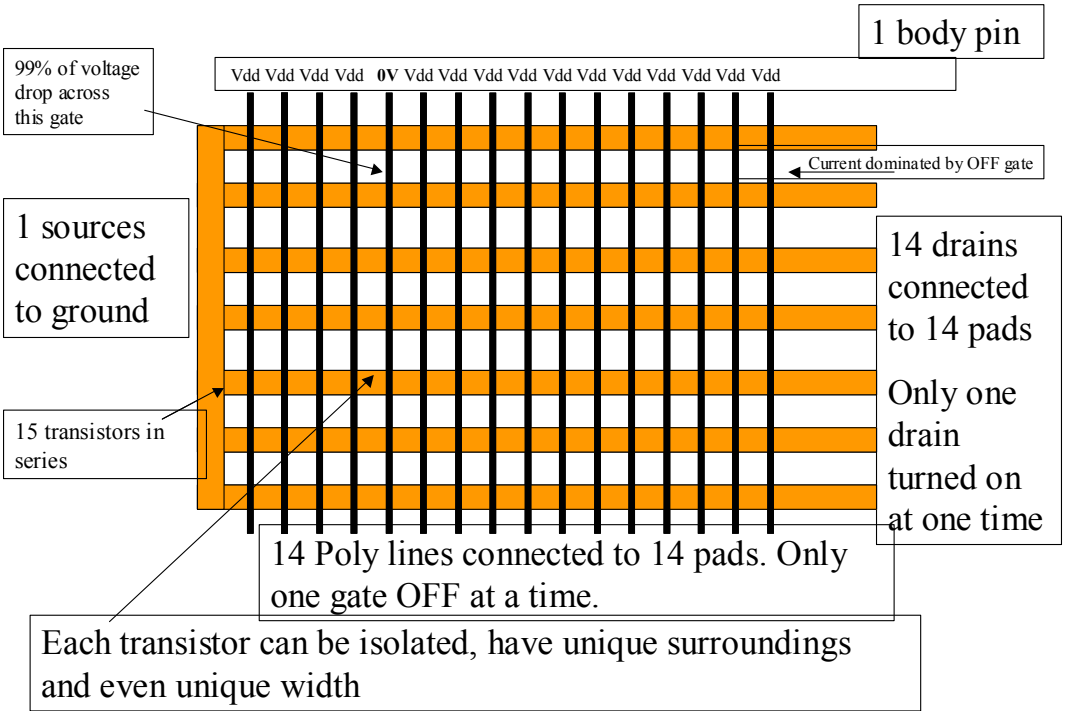


Figure 16 Passive multiplexing strategy for individually addressing 196 transistors with only 30 pads. Each transistor is isolated by applying a bias across only one set of transistors and turning all gates ON except the one you want to measure.

The series transistor multiplexing approach seen in Figure 16 has the most layout flexibility, but has the drawback of only providing leakage current measurements. Dense pitches are easy to create and do not require a contact to be placed in between lines as the transistors measured are in series anyways. This multiplexing strategy was used in creating test structures with varying pitches and standard cells. The way to isolate one

transistor with the HP 4145B probe station is to connect SMU1 (Source/Measurement unit) to the drain pin of the transistor and set it to a low voltage, SMU2 to the single source pin, body pin, and the rest of the drain pins and set it to 0V, SMU3 to gate of the transistor and set it to 0V and one SMU4 for the rest of the gates and set it to 1V. The leakage current measured from SMU1 should be the leakage current through the one OFF transistor where almost all of the voltage is dropped. It is important to use a low drain voltage to ensure a Vt drop from gate to source across all transistors. Since the leakage current of these transistors is expected to be in the tens to hundreds of nano-amp range, the source and drain resistance would have to be greater than a mega-ohm in order for a significant voltage drop to exist between the drain pin and the drain of the measured transistor.

That said, any voltage drop between the source pin and the source can lead to significant changes in leakage current as the gate to source voltage would drop. If this becomes an issue the proper extraction techniques will need to be used and each transistor will need to be modeled with a significant source and drain side resistance. If this is the case it might be necessary to probe transistors with different source side resistances at slightly different source voltages. The goal being to replicate the same gate to source voltage for all measured transistors. Another potential source of noise in this structure is junction leakage. This is greatly reduced by using low drain voltages. As the total area of the active region is in the range of 10s of microns in the worst cases, junction leakage current may significantly add to the measured leakage current. If this is the case then the junction leakage current will need to be quantified by measuring the current if both the source and the drain are at high and all the gates are ON. If this current is more

than 1% of the measured leakage current, then it will need to be included in the calculations. Fortunately this current scales linearly with active area and should be easy to filter out. It can also be measured for almost all gates as transistors in series can serve as stacked transistors and if more than two or three transistors are turned off, the leakage current will fall drastically and the junction current will dominate. This structure hence suffers the problem of potentially requiring significant amounts of post processing instead of just using the raw data, hence it is wise to avoid it if possible. As the SVTC process was a single metal layer process with relatively large contacts, this multiplexing strategy was necessary for a lot of dense pitch designs and was a lot simpler to implement.

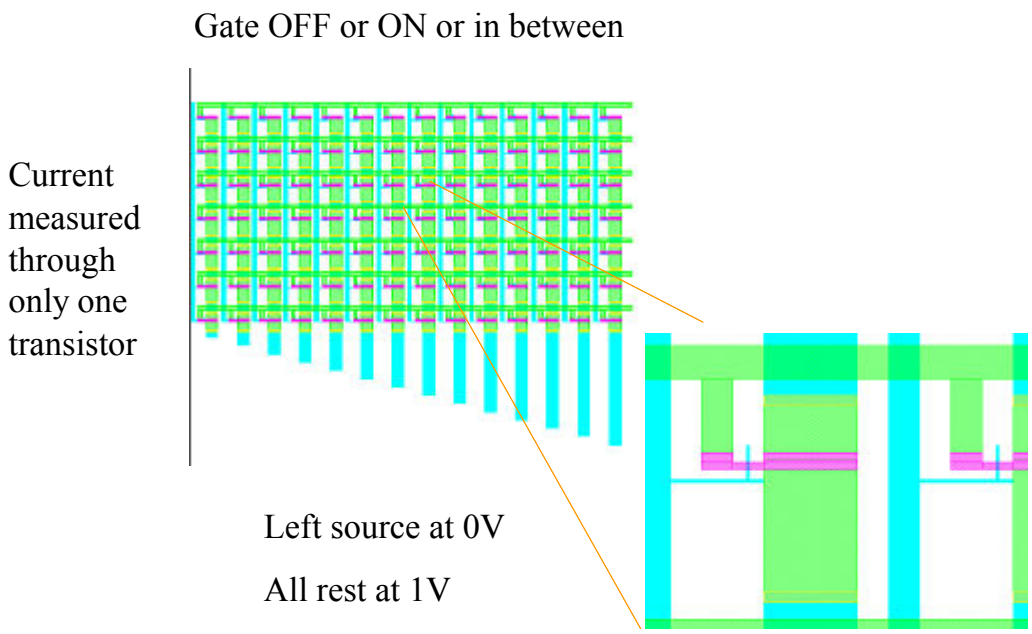


Figure 17 This is the ON/OFF structure with 196 passively multiplexed transistors. This array is automatically generated by scripts that can modify the dimensions and proximities of each transistor.

The ON/OFF test structure (Figure 17) uses a slightly different approach that enables multiplexing 30 pads into a 196 transistors that can be measured for both ON and OFF

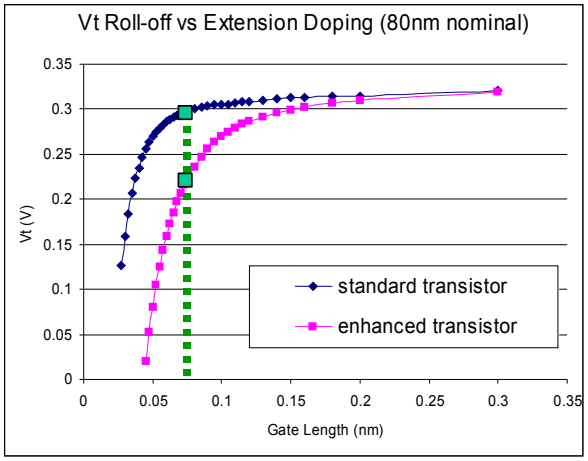
current. Instead of 14 gate pads there are 14 source pads and only one gate pad. So each transistor can be addressed individually by applying SMU1 to the drain pin at 1V, SMU2 to the source pin at 0V, SMU3 to the gate, SMU4 to the body and an extra source unit (of which there are two) to all the remaining source and drain pins and set it at 1V. This way SMU1, which is connected to the drains of 14 transistors only sees 1V across one transistor and hence the current from SMU1 will be the leakage or drive current from the one targeted transistor.

This method is preferred as one can measure leakage and drive current, but it has some geometrical limitations and can be susceptibility to junction leakage. Since there is only one gate connection, the minimum pitch allowed is a contacted pitch or currents would have to be measured through stacked transistors. There might also be some issues with source resistance as the signals had to be routed through poly. These disadvantages should be mitigated with more than one metal layer and smaller contacts (the SVTC process had 160nm contacts). A potential improvement may be in the form of a hybrid structure with two or three gate pins. This multiplexing strategy would then allow pitches that do not have a contact next to the gate under measurement, but would not suffer the restrictions of a structure with 14 gates in series. With enough metal layers this structure can be used for passively multiplexing and individually addressing very dense test structures as routing source and drain signals to appropriate pins takes up significant area in a single metal layer process. Testing transistors with a wide array of different proximities can then paint a clearer picture of the process conditions. Enhanced Transistors may be used for more process specific analysis or to increase the signal to noise ratio.

4.2. Enhanced Transistor Electrical CD Metrology

Enhanced Transistor Electrical CD Metrology uses enhanced transistors that are hyper-sensitive to gate length variations and hence have a much stronger correlation between CD and leakage current than standard transistors. These transistors can be created with the addition of one extra implant step that basically increases the short channel effect or sensitivity of threshold voltage to gate length. The key benefits of this method are automated electrical measurements, almost no geometrical restrictions, and high density test-structures. The enhanced transistor method also allows to isolate small segments of a gate, down to 224nm in our current process, which is critical for studying LWR and proximity effects. So unlike other electrical metrology techniques, that have significant geometrical limitations, like ELM, measurements can be automated to study how any standard cell or test structure design responds to process variation.

The basic idea is to increase the correlation between gate length and threshold voltage by increasing the short channel effect. The short channel effect is an effective lowering of the threshold voltage at shorter gate lengths. This can be seen in Figure 18 where the threshold voltage of the device decreases below ~80 nm for the standard transistor and ~150nm for the enhanced transistor. The Vt rolloff curves were generated from a dummy transistor built with TSUPREM4, a process simulator, and was simulated in MEDICI, a 2D device simulator. The enhanced transistor in Figure 18 was created by increasing the extension implant dose by 10X, but this enhancement will vary based on device design.



Denotes the operating point of a 80nm transistor

	Gate CD	Gate Oxide Thickness	Channel Doping
Ioff sensitivity enhancement	**2500%**	**750%**	**1125%**

Figure 18 Vt rolloff curves for standard transistor and enhanced transistor. The enhanced transistor has a 10X higher extension implant dose. The sensitivity of leakage current to gate CD is increased by 2500%.

Increasing the extension implant dose shifts the Vt rolloff curve to the right as the drain starts to have more influence on the channel with the higher doping. Looking at the 80nm operating point on both curves, the standard transistor curve is relatively flat, or will have a relatively small Vt response to a change in CD. The enhanced transistor on the other hand is in a steep portion of the curve where a small change in CD translates to a large change in threshold voltage. In fact, the sensitivity of leakage current to gate length is increased by 2500% with the enhanced transistor. The sensitivity of leakage current to gate oxide thickness and channel doping is also increased as the device is a lot less stable, but not as much as the sensitivity to gate length. This translates to a stronger correlation between CD variation and threshold voltage variation. Hence leakage current measurements, which are done automatically on an HP4142B probe station, can be more accurately translated to CD data. CD data can then be analyzed and sources of variation can be assessed.

To employ this method, the runcard needed to be amended with two extra lithography and implant steps and a characterization structure was designed to characterize the enhanced transistors. One lithography step was used to mask standard transistors and use an increased extension implant dose for enhanced transistors and the second lithography step was used to cover the enhanced transistors and use the standard extension implant recipe. It is possible to use just one litho step if enhanced transistors are implanted twice, but this type of structure may be harder to fine tune. A test structure with transistors of varying gate length from 65nm to 134nm was implemented for 224nm, 400nm, and 2000nm wide transistors in order to build a lookup table of leakage current vs gate length. A step size of one nm was chosen as the designed gate lengths are expected to vary themselves. Fitting an exponential curve to the graph of leakage current vs designed gate length should yield a very accurate mapping from leakage current to CD. So even if some gates end up being narrower or wider due to LWR, mask errors, and process non-idealities, looking at all the transistors together and averaging over many die will significantly reduce any errors.

Enhanced Transistor Electrical CD Metrology (ETEC-M) may not be useful for process control, as it requires probing of completed transistors, but it does offer extra degrees of freedom in characterizing specific processes. The accuracy, flexibility, and simplicity of this method is amenable to analyzing the response of a large variety of layouts to process variations. Using transistors directly allows you to probe real patterns that exist in real standard cells and one can pinpoint specific areas of a line by adjusting the active region appropriately. This allows for probing specially designed test structures, such as aberration monitors, that can have amplified responses to specific process

variations. All surrounding gates or transistors can be turned ON, so specific gates in a densely packed poly pattern can be addressed, with a low Vds voltage which is required to guarantee that all gates stay ON. Finally, the resolution of this technique will only be limited by other sources of Vt variation that will blur the data. Transistors of the same gate length can still have different leakage currents due to random dopant fluctuation or other process non-idealities. A significant portion of these variations can be filtered out as they are systematic and can be measured by other complementary test structures such as capacitors for oxide thickness variation and silicon resistors for long-range dopant concentration fluctuations. Isolating random dopant fluctuations can be done statistically by looking at the variation difference between enhanced transistors to standard transistors that are within a close proximity and have the same proximity effects or by looking at transistors at room temperature and then at cryogenic temperatures. At cryogenic temperatures the dopants in the channel freeze out and have a different effect on threshold voltage.

4.3. Cryogenic Testing

Testing transistors at room temperature and at 4K, where the channel is totally frozen out and channel doping has negligible effects on threshold voltage, can be used to study the components of random threshold voltage variation. Line Width Roughness (LWR) and Random Dopant Fluctuations (RDF) are two major sources of random threshold voltage variation that are very hard to segregate. The benefit of understanding the weight of each component of random variation is that design techniques can be employed that can mitigate each of these effects. LWR can be reduced by improving the Image Log Slope[41] and RDF can be improved by changing the transistor design or increasing the size

and drive strength of logical gates. Since the device physics of the device changes significantly at 4K, but all parameters like gate length and channel doping stay the same, LWR should be accurately quantified as confounding effects can be filtered out.

4.4. Lithography Test Structures

The testchip consists of 35 different types of test structures designed by six students from four different research groups in processing, devices, and circuit design. The two strategies taken in designing these litho test structures was to create high or low sensitivities to particular non-idealities such as defocus and LWR. The enhanced transistor electrical CD metrology made this feasibly as measuring a wide array of different test structures does not require any calibration or manual labor. In fact, most test structures can be measured using the same input test vectors independent of transistor dimensions or proximities. The hope is having measurements from test structures with one dimension or parameter varying incrementally from low to high will paint a clearer picture than a simple ON/OFF experiment where one value is either low or high. For example the vary pitch structure varies the pitch from min pitch to isolated lines in forty-two small increments.

4.4.1. Vary Pitch Aberration Test Structure

The vary pitch cell contains a set of five line arrays of varying pitch from below minimum pitch to isolated line. Looking at arrays with different pitches has shown predictable responses to process non-idealities such as defocus and MEEF [42,43]. The vary pitch structure consists of five line arrays, of which the center (dense) and right most (semi-dense) lines are measured, varying from below minimum pitch to isolated line in

10nm and 30nm increments. This wide array of pitches intentionally consists of preferred, forbidden, and isolated pitches. Since each pitch will sample a different part of the pupil, this structure can be used to study aberrations in the lens as well as other process non-idealities. For example, Figure 20 shows the simulated response of dense lines to defocus through pitch. One will notice certain pitches being more sensitive, hence these pitches can serve as defocus monitors if they are measured to have a higher CD error than the rest. Some wafers will intentionally be exposed with dose and defocus offsets to measure the response of this structure to dose and defocus and identify the most sensitive pitches. Different vary pitch test structures were designed at 80nm, 100nm, and 120nm as well as with varying amounts of redundancy. The current test structure is 59um x 313um, but can be easily scaled down to 20um x 72um if multiple levels of metal could be used.

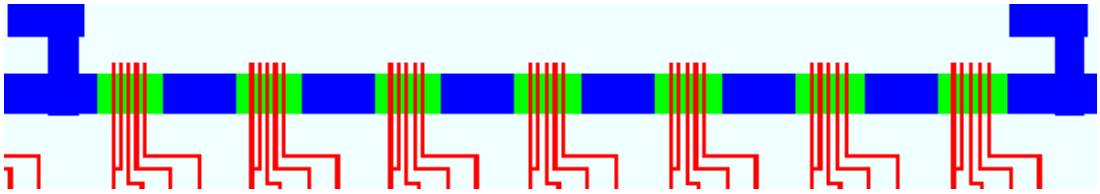


Figure 19 One row of vary pitch test structure. The middle and right most lines of each 5 line array have individual pins and the rest are tied to a permanent ON pin. This leads to dense and denso linewidth measurement for 42 different pitches inside one 30-pad cell.

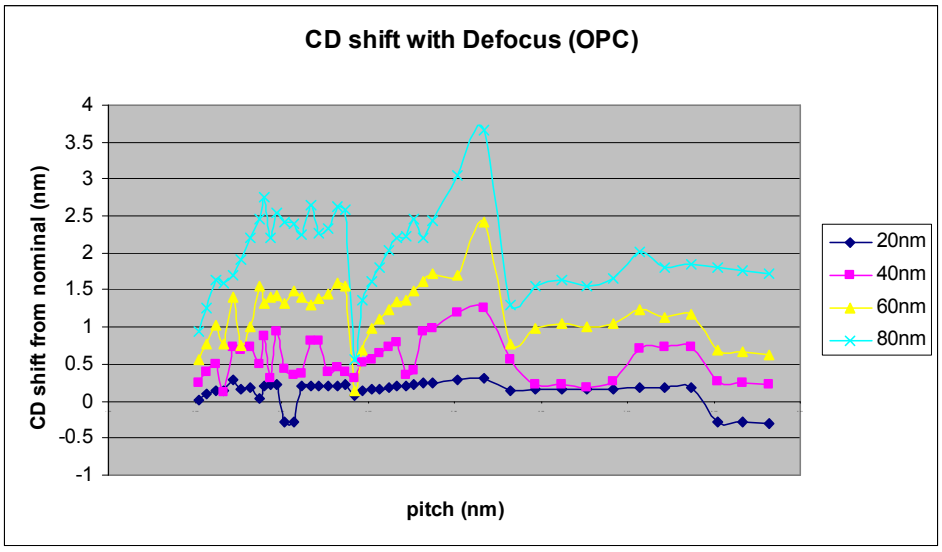


Figure 20 Simulated CD shift of 80nm line vs pitch for 4 different levels of defocus. Some pitches exhibit high sensitivity while other pitches are iso-focal or do not change through focus. NOTE: Simulation results are from a post-OPC design, which is not shown in Figure 19.

4.4.2. ELM Test Structure

Paul Friedberg added an ELM structure for studying spatial correlation as well as systematic CD variation. The basic principal of this structure is to push a constant current from one probe pad to another, while measuring the voltage drop of a segment in the middle. With thirty probe pads, a serpentine test structure can be made to measure the line width of 25 different segments. Four pads are burned on a Van Der Paw Structure to measure resistivity. This data has been used in the past to statistically characterize CD variation as die to die (dose variation), across wafer (PEB, etch, and other wafer scale process variations), and random variation (mask error)[44]. Sampling the serpentine structure at different frequencies or periods has also shown to be useful to characterize spatial correlation on a range from 0.200um to 1.0mm[45]. We will use similar techniques and test structures to help characterize this set of experiments. As a proven electrical

metrology technique it will also serve as a litmus test for accuracy of using transistors as measurement structures.

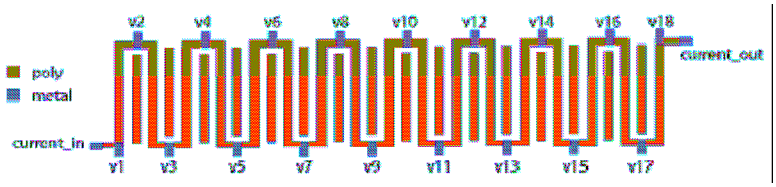


Figure 21 ELM test structure with one dummy line in between every measurable poly-silicon line. Varying the amount of dummy lines one can analyze different spatial frequencies with periods ranging from 0.200um up to 1.0mm

4.4.3. LWR Test Structure

A LWR test structure has been designed to have different amounts of Line Width Roughness(LWR) present on different width transistors where the roughness may or may not average out. Looking at the variance of the leakage current of these different transistors will help indicate the severity and characteristics of LWR, especially when RDF effects are filtered out using cryogenic testing. The basic principal behind this test structure is that LWR is a function of image quality, mainly ILS, which can be modified by changing the pitch of the structure. Figure 22 shows the LWR test structure consists of isolated lines and 5 line arrays with either preferred (good ILS) or forbidden (poor ILS) pitches strewn over active regions of different width. The idea is that on wide transistors the LWR will average out and the variation from die to die will be a lot less severe than on a narrow transistor that will only sample a small portion of the LWR.

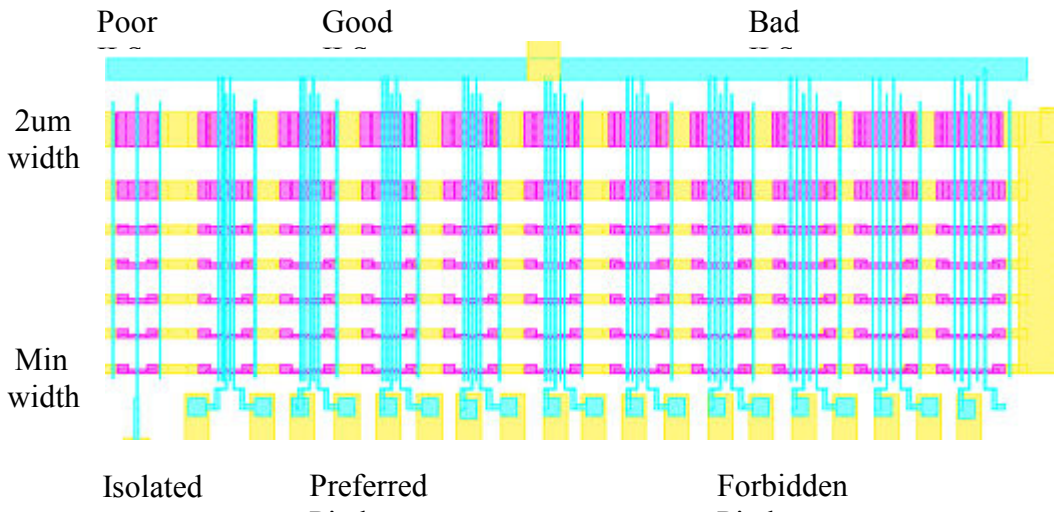
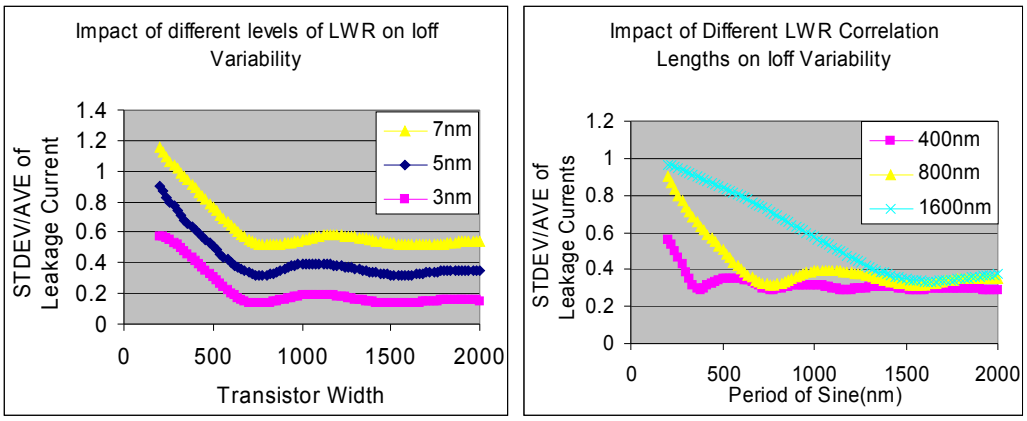


Figure 22 Electrical LWR test structure. Different pitches with different image qualities will result in different levels of LWR that will either average out on wider transistors or significantly impact variability.

Figure 23 shows the expected standard deviation of measured leakage current from 65 randomly generated transistors vs transistor width. This is a very simplified model where the LWR is assumed to be a function of two independent sine waves with a random phase and a given amplitude and period. The simulate leakage current is calculated based on the transistor shape and the non-rectangular transistor model. 65 transistor profiles were randomly generated by randomizing the phase of the sine wave on each transistor gate edge.



(a) **(b)**

Figure 23 These plots are generated looking at 65 randomly generated transistor gates with independent sinusoidal LER on both edges. (a) shows how the STDEV/AVE decreases with width for different levels of LER (3nm, 5nm, and 7nm) and (b) show the same for LER with different correlation lengths

A lot can be extracted from examining the plot of leakage current variance vs transistor width. Periodic dips indicate a specific periodicity of Line Edge Roughness (LER) or a dominant frequency in the spectral decomposition. Also, looking at how low the periodic dips are will be a sign of how correlated both edges are. If both edges are uncorrelated, as assumed in these simulations, the total variance from LWR will never go to zero as different transistors will have phase differences between edges and hence may have worse pinching depending on how close the phase shift is to 180 degrees (Figure 24). Finally the amount of LWR will also decide the ratio of STDEV / AVE, especially for minimum width transistors. Keep in mind high frequency roughness < 50nm is averaged out due to diffusion, so low frequency roughness will most likely be the main contributor to transistor to transistor variation.

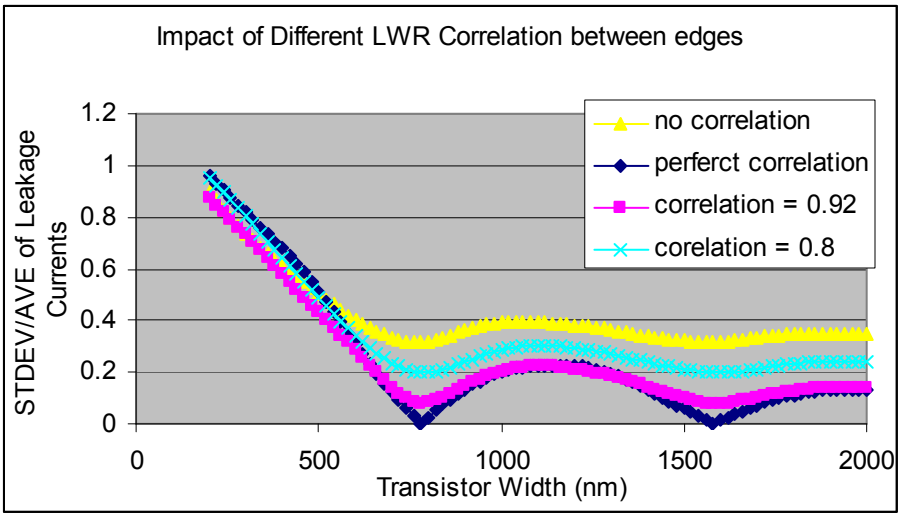


Figure 24 The minimum variance in the dips depends on how correlated LER is between edges. The more correlated and periodic the LER is, the variance of appropriate width transistors due to LER could be very low.

4.4.4. Overlay Test Structure

Since misalignment plays a critical role in circuit performance and may become more critical as double patterning techniques become more popular, an electrical misalignment test structure can help characterize overlay errors on all die and all wafers[46]. Two types of misalignment test structures have been designed; one is transistor based and the other is resistivity based. The transistor based test structure uses 192 transistors with varying amounts of poly overlap. The poly overlap is varied from –50nm to + 135nm in 1nm increments. Finding which transistors can no longer be turned OFF will signify that they don't fully overlap the active region and a misalignment value can then be estimated. It is expected that plotting leakage current vs programmed overlap will have an inflection in the curve that will help identify misalignment.

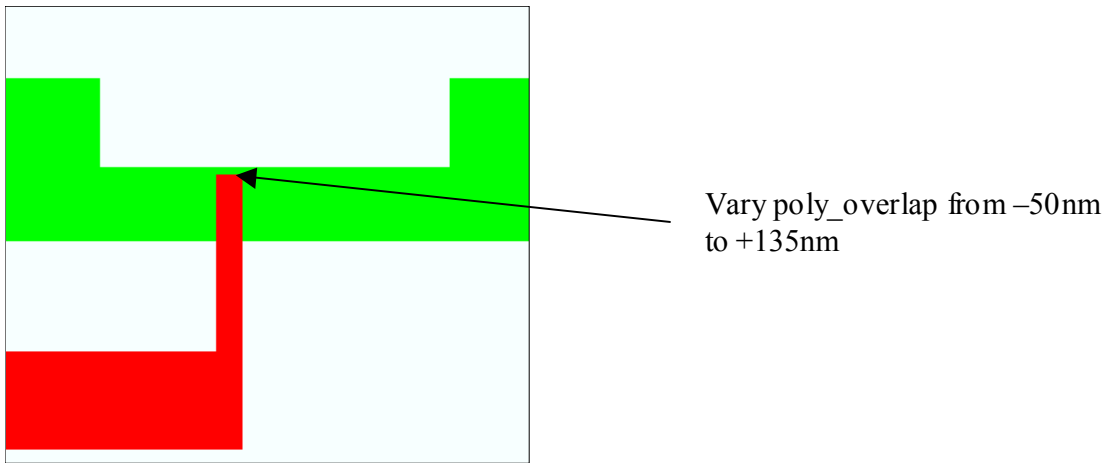


Figure 25 Transistor based overlay Test Structure. Depending on the amount of poly overlap and the misalignment error, some transistors will not be able to turn off as the gate will not fully overlap the active.

The amount of misalignment between the contact, metal, and poly layers will be measured with resistive measurements as seen in Figure 26. Since the contact area is very small, the resistivity will be highly sensitive to overlay error. Wafers programmed with varying amounts of misalignment will be used to calibrate the test structures.

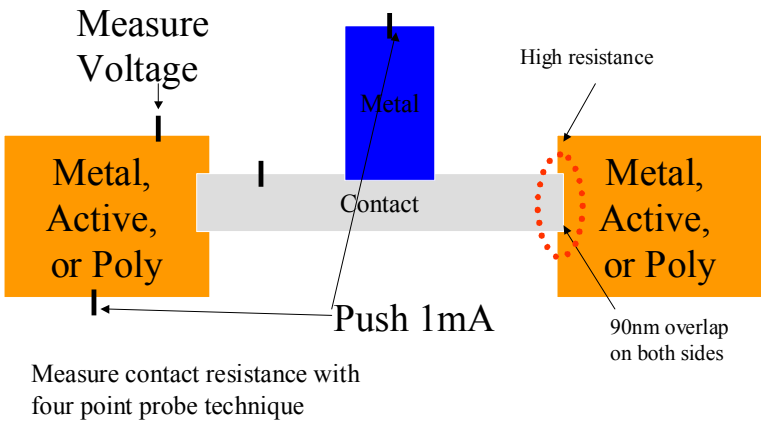


Figure 26 Resistivity based misalignment test structure. Misalignment to the contact layer is measured for poly, metal, and active as the small overlap area translates to large resistivity variation with overlay error.

4.4.5. Defocus Test Structure

Defocus plays critical role in any lithography based DFM tool as it is a source of variability that is getting worse with hyper NA lithography and needs to be fully characterized for accurate lithography simulation and hotspot detection. Probe based Zernike aberration monitors have been shown to be hyper sensitive to defocus, but require CDSEM metrology and are more digital (ON or OFF) than analog in nature[47,48]. The defocus monitor has been modified by Juliet Holwill and fashioned into a transistor to enable electrical measurement(Figure 27,Figure 28). Although sensitivity to defocus is lost as a significant portion of the aberration monitor had to be taken out, the measurement sensitivity has been greatly increased as now transistor CD will be measured. With the help of Toppan Photomask, a new type of mask has been manufactures to make these defocus sensitive monitors possible. As these small features require an attenuated phase shift mask and a center probe with a 90 degree phase etch, an attenuated phase shift mask with an extra 90 degree phase etch into glass was created.

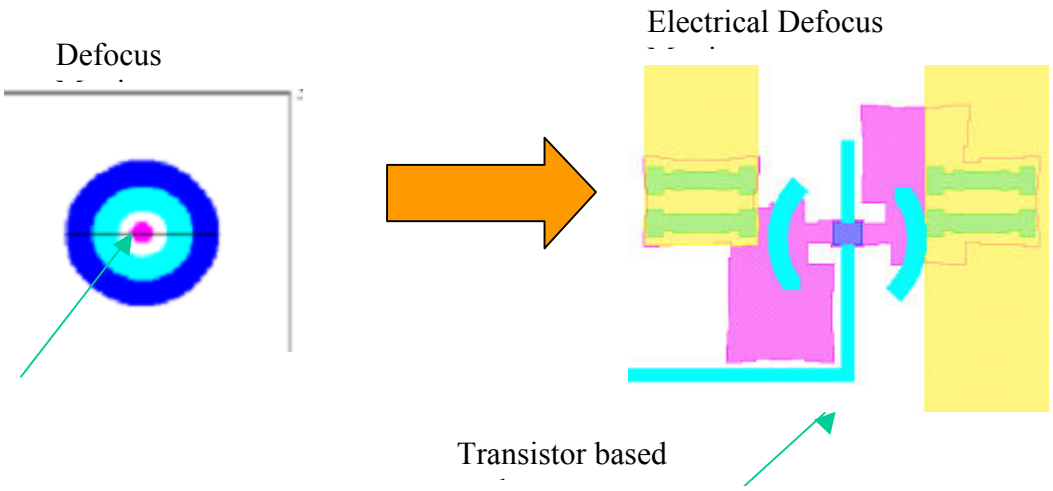


Figure 27 The defocus monitor on the left has been fashioned into an electrical defocus test structure. The center 90 degree probe region is hyper-sensitive to defocus, so the gate CD is expected to change dramatically in response to defocus in the scanner.

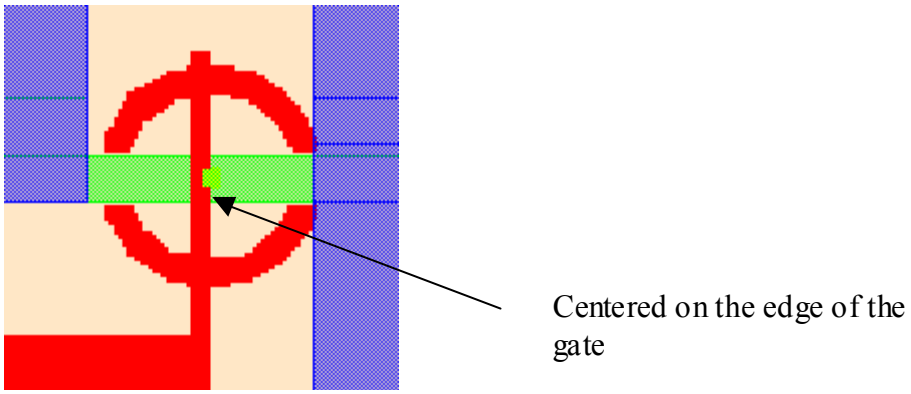


Figure 28 Second generation version has a much simpler active region and the monitor is centered on the edge of the gate to maximize edge movement through focus.

4.4.6. Corner Rounding Test Structures

Corner rounding on both poly and active set minimum active-poly space design rules and limit transistor packing density. To evaluate safe active-poly spacing rules it is important to not only evaluate the radius of the corner rounding, but also how the corner

rounding changes through process. The edge placement error can increase more dramatically as corners generally have poor Image Log Slope and are less stable. To test the sensitivity of transistor performance to corner rounding two sets of test structures vary the poly-active spacing in small increments. The poly corner rounding cell, or set of 196 transistors, varies the poly elbow below active space from 0nm to 200nm for 180nm, 224nm and 400nm wide transistors (Figure 29). Simulation results have shown that not only the corner, but a necking region beyond the corner rounding can have a very significant change through focus. Active rounding can also be a big issue leading to narrower transistors, so a similar structure has been used, but this time with the distance between the poly gate and the large step in active region width is varied.

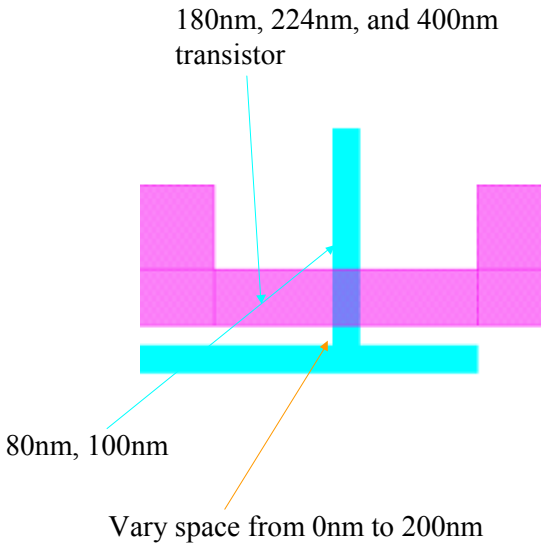


Figure 29 Poly corner rounding test structure. The space between the active region and the elbow in poly below active is varied from 0nm to 200nm.

4.4.7. Non-Rectangular Transistor

In an effort to validate the non-rectangular transistor model a set of transistors have been designed to have very non-rectangular gates. This will test the accuracy and the

limits of the non-rectangular transistor model proposed in chapter 3. One strategy is to create triangular transistors that have a wide bottom and skinny top. These transistors will have a wide array of slice widths and can double up as overlay test structures as the top slice thickness will be a strong function of overlay. Another option, that helps test the significance of slice position in the channel is to create a transistor with one wider slice or with a narrower slice. The non-rectangular test structure consists of an 80nm line with a 120nm slice in the middle The position of this slice was moved from the bottom edge to the top edge in 80nm steps in an effort to test the significance of slice position. Comparing transistors with the slice in various positions will show the significance of the slice being at the edge vs in the middle.

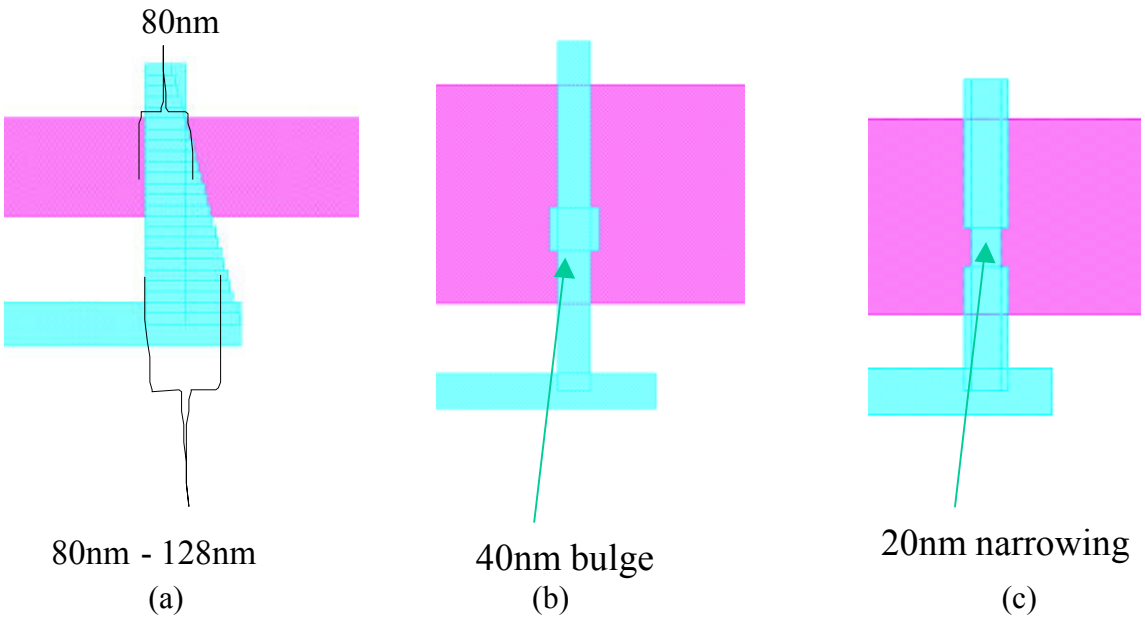


Figure 30 Three flavors of non-rectangular transistors. (a) triangular where the bottom CD is changed while the top is held constant, (b) isolated with on 40nm wider slice that varies in position from top to bottom, and (c) an isolated line with a 20nm narrowing that varies in position from top to bottom.

4.4.8. Enhanced Transistor Characterization Test Structure

In order to characterize the accuracy of Enhanced Transistor Electrical CD Metrology (ETEC-M) a structure was created that measures the same poly line in two ways. The line can be measured by Electrical Linewidth Metrology (ELM), which has shown to correlate well with CD-SEM[49]. The poly line also forms the gate of a transistor and hence can be measured by ETEC-M. Since both methods are automated a large enough sample of data can be generated to evaluate ETEC-M accuracy as it will not be limited by current measurement sensitivity, but variations in channel doping, oxide thickness, and other transistor parameters that are not associated with gate length. There are different versions of this structure with various width transistors to see if LER and RDF reduce accuracy for narrower transistors.

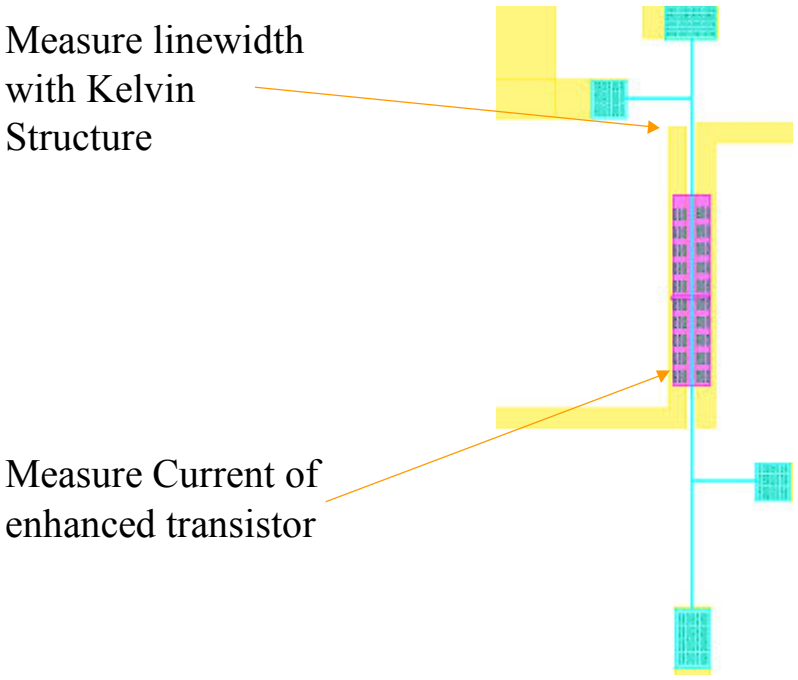


Figure 31 Enhanced transistor married to an ELM 4 point probe structure. The same poly line is measured with the proven ELM method and the Enhanced Transistor Electrical CD Metrology method.

4.4.9. BSIM model Fit Test Structure

In order to get an accurate BSIM model a BSIM model fit cell was created for both enhanced and standard transistors. As to minimize error, these transistors were not multiplexed and each drain was connected to only one pin. In order to fit a BSIM model a sampling of transistor widths and gate lengths need to be measured. The structure in Figure 32 characterizes transistors at a minimum width. There are similar structures with varying gate length on a wide transistor, varying transistor width for a short channel transistor and varying transistor width for a long channel transistor.

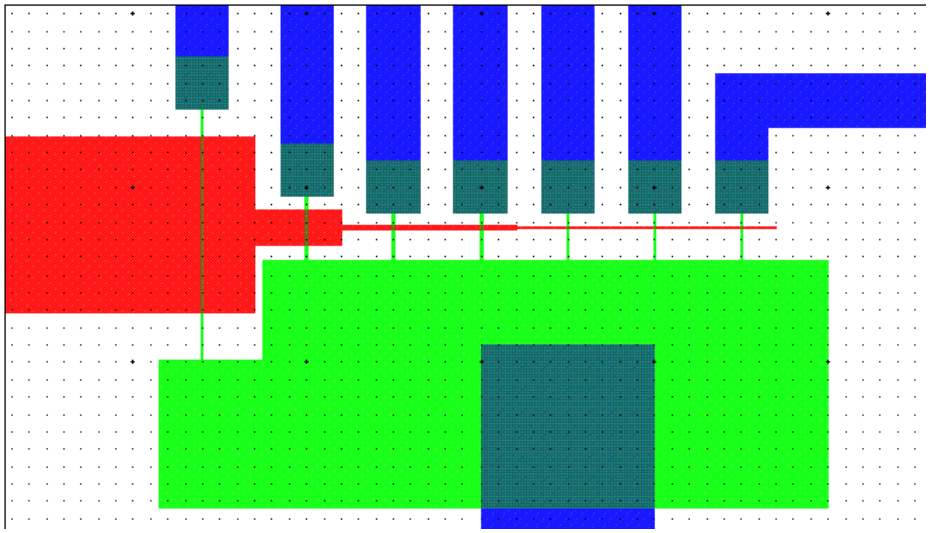


Figure 32 BSIM model fit test structure. This has various gate length and active width transistors that are not multiplexed in any way. There is a chart of transistor widths and length that are needed to extract all the BSIM parameters. This 30-pad cell has transistors with these dimensions.

4.4.10. SRAM and Standard Cells

To add some randomized layout geometries that might prove sensitive to different process parameters a set of standard cells and SRAMs have been implemented in the parallel transistor format. The parallel transistor format was more amenable for this task

as a lot of standard cells have minimum pitches that do not leave enough space for a contact. The SRAM design was added as SRAM designs traditionally push the envelope and violate design rules to maximize density.

4.4.11. Dopant density test structures

To study long range dopant fluctuation in the channel and source/drain a set of MOSCAP and active resistivity structures were designed. The capacitance of large MOSCAPs can be measured directly and the channel doping can be extracted from the low frequency C-V curve. If dopant density varies across the wafer or from wafer to wafer this structure will be able to detect it. Van Der Paw and ELM structures were designed in the active layer to measure source/drain dopant concentration. If the source/drain dopant concentration varies it can affect the electrical channel length so measuring the resistivity is an independent way to identify a source of channel length variations. The main goal of these test structures was to have a less complicated and relatively litho-independent method of measuring dopant concentration variations.

4.4.12. Pass Transistor Logic

In addition to single transistor test structures, a non-standard design technique using pass transistor logic has been implemented to study random and systematic variation. In pass transistor logic each signal is routed through multiple transistors that are not hooked up to the power supply. The goal is to evaluate how pass transistor logic responds to random and systematic variation. Since signals in pass transistor logic are passed through multiple transistors, the hope is that random variation will be averaged out. This can be experimentally examined by inducing defocus in the scanner, which degrades image

quality and increases LWR for random variation or change the dose induce a systematic shift. Louis Alarcon designed different depth logic trees, some with pre-programmed inputs to maximize sensitivity to systematic variation.

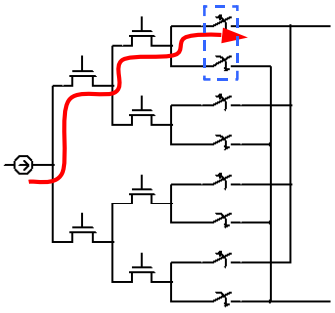


Figure 33 Example of a 2-level pass transistor logic tree. Any two input logic function can be implemented in this way. The testchip includes 3, 5, and 7 level pass transistor logic trees.

4.4.13. Reproducing Pattern Dependent Variations

Previous experiments executed by Liang-Teck Pang using a 90nm ST flow showed significant amounts of pattern dependent variation with suspected sources being mask errors, image quality related variation, and coma[15]. Each of these errors is suspected for various reasons, but definitive conclusions are impossible to draw as the entire process, including OPC, is proprietary foundry IP. The same test structures were reproduced on the current testchip with the hopes of reproducing similar pattern dependent variation. Since the current process is fully known, post OPC data is available, and all the characterization test structures will be used to identify process non-idealities, stronger conclusions will be formed when analyzing the same data on the current experiments. The different patterns can be seen in Figure 34.

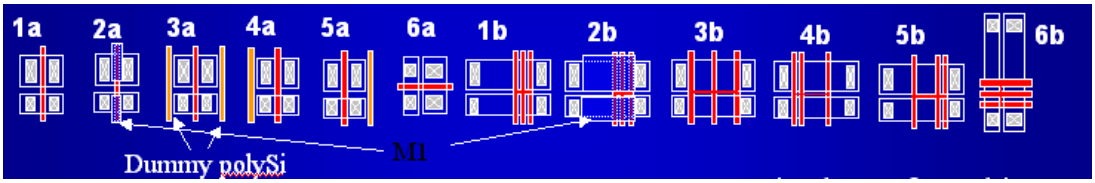


Figure 34 **Twelve different layouts that have previously shown significant pattern dependent levels of variation. Now the pattern dependent variation can be better analyzed with a much better understanding of the process from the other test structures**.

4.5. Conclusion

A collaborative testchip has been designed from a process/device/design perspective with the aim of characterizing the lithography process in a 65nm NMOS flow. To help facilitate combining designs and test structures from multiple designers a standard 30-pad cell was used as the basic building block. The primary test structure is based on Enhanced Transistor Electrical CD metrology that enables automated measurement of dense structures with arbitrary geometries. Most of the testchip is covered in probe pads, so test structure density was not a main concern, but scribe line or small footprint versions of most test structures should not be hard to design. Most transistors have an enhanced and standard version for comparison and potentially for comparing random dopant fluctuation (RDF) and LWR. Test structures have been implemented to specifically look at LWR, focus, misalignment, proximity effects, and systematic variation, although the hope is to have enough variety to capture all significant effects. With test structures and conclusions from six different students, this comprehensive study of process variations can generate sets of strong conclusions that reinforce each other and paint a clearer picture of the DFM challenges we will face in the future.

5 Collaborative Database

The final piece of the Collaborative Platform for DFM is the collaborative database that serves as the glue between experiment and simulation as well as the front end for the platform. With over 15,000 test structures the need for an infrastructure to deal with the massive amounts of data is obvious. The benefits of a database is that it enforces structure to all data, the data is centrally located and web accessible, and the barrier to exchanging information on similar projects is greatly reduced. In fact not only data is stored in the database, but queries that form the basis of data mining are also stored and accessible by others. This is a key enabler for multi-designer projects where project overlap may not be significant enough alone to instigate collaboration. The database can also drive the Parametric Yield Simulator and store simulation results in an annotated fashion so that trends can be identified and analyzed. A process extraction strategy is based on this capability as well as the shotgun approach of including a wide array of test structure. It consists of sets of queries that calculate test structure statistics and filter out good process monitors. As a platform for comparing simulation and experimental results and sharing data mining techniques, the database is a key component of the Collaborative Platform for DFM.

5.1. Database Design

The online accessible database has been designed and implemented in MySQL and Ruby on Rails to store layout design data, simulation data, processing data, and experiment measurement data. The database enforces structure on all uploaded data, which enables easy comparison between datasets, and is online accessible, which enables access from any type of computer anywhere in the world. Beyond serving as an interface with data, the database website serves as a platform for collaborative data analysis and a driver of the Parametric Yield Simulator. Queries of each user can be saved, loaded, searched, and ranked by others, so good queries can be reused and improved in a collaborative fashion. The database also stores past analysis results or extracted process conditions, so contributors can compare results or use previous results to filter out confounding effects. Several tables in the database are specifically allocated to store extracted process-non-idealities and their repercussions on the printed test structures. This section will describe how the database is structured and the tables it includes.

5.1.1. Database Structure

Figure 35 shows an Extended-Entity Relationship Model (EER) diagram of the database, which is essentially split into four sections; tables or rectangles in the upper left describe the layout, the bottom left describe simulation results, the bottom right describe experiment measurements, and upper right describe process conditions. Each rectangle represents a separate table and each diamond represents a relation. Splitting the data into multiple tables reduces the need for storing redundant data as a lot of test structures share similar attributes. The numbers in parenthesis (x,y) between each rectangle(table) and diamond(relation) represent the “x” number of rows in the given table can be related to

"y" rows in the related table. For example, a cell may have 1 to n (1,n) individual transistors, but each transistor can only be associated with one cell (1,1). Each table has a set of attributes or columns and tuples or rows that correspond to individual instances. Each attribute also has a specific datatype and some attributes in each table have to be defined for each row or cannot be NULL. This enforced structure is the key to running queries and slicing and dicing the data as needed. In essence each test structure is associated with hundreds of attributes and dozens if not hundreds of simulation and experimental measurement results that could be used to identify test structures of interest. Generally each test structure has at least one row associated with it in every table in the database.

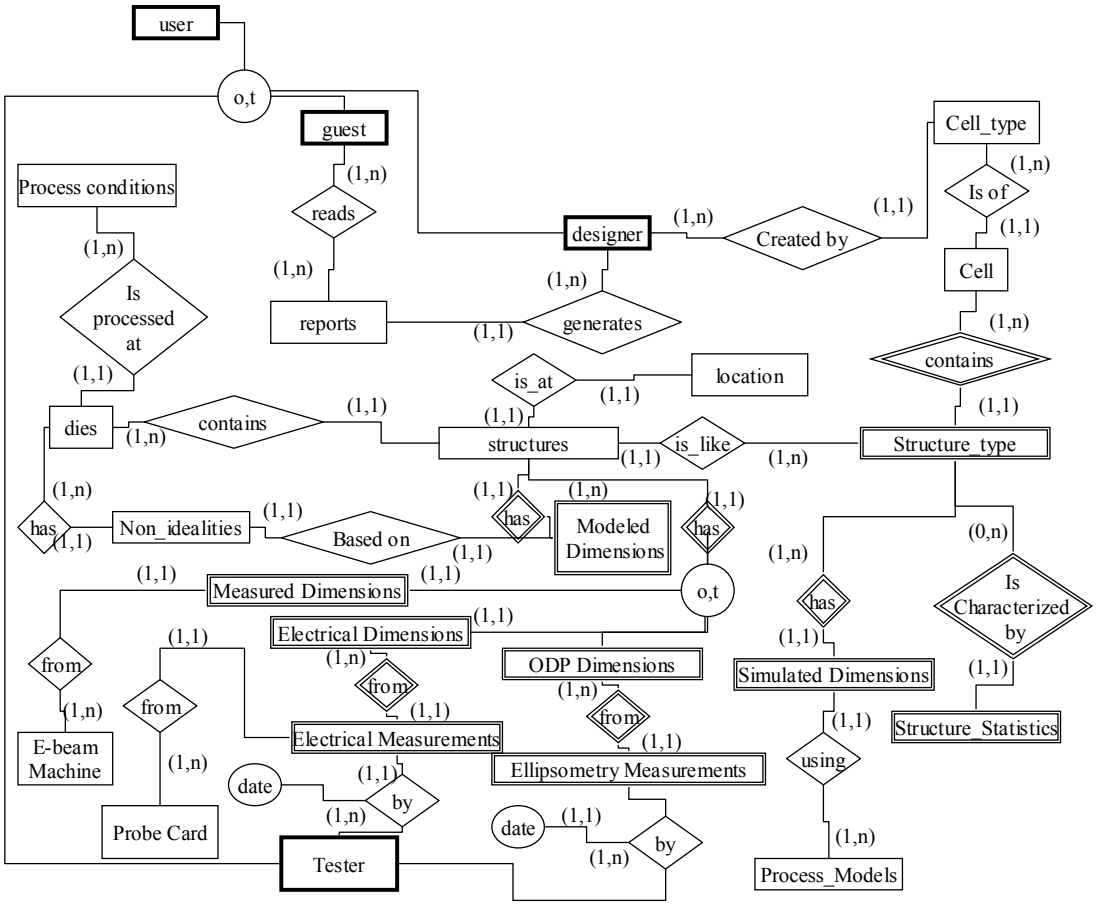


Figure 35 EER Diagram or schema of the database

The upper right section of the EER diagram holds layout and design intent information and can be seen independently in Figure 36. The designer table holds information about each designer such as contact info, background, and advisor. The current testchip is split into 201 30-pad cells that can be grouped into 74 different categories or cell_types. The cell_types table stores a description of each cell_type in terms of the designer of the cell and the intent. Each cell_type has a cell_name and description. This table could also be used to store separate layouts or chips as separate cell_types if the database where to be used for other testchips. The table would then have to be expanded to accommodate pointers to layout file-names, which will no be an issue as the database has been created to be flexible in such ways. Every table in the database has a comments attribute, which is a generic text field that can store more data with XML style meta-data. Additional columns can be added by tagging text in the comments field with new attribute names, this process is explained in the next section.

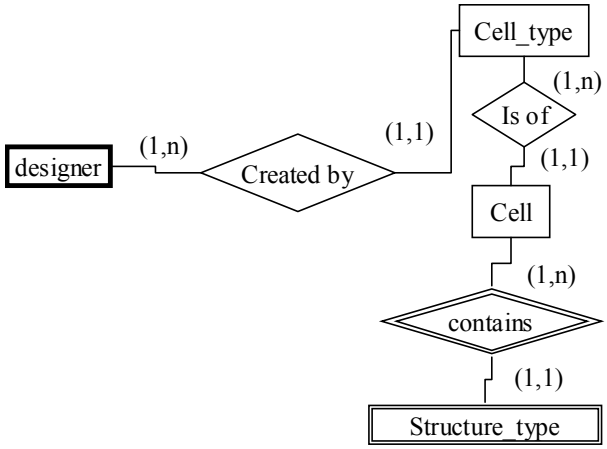


Figure 36 Section of the EER diagram associated with describing the layout and design of the test structures

The cells table holds information about specific cell instances in the layout, including cell location, primary CD, if the cell has OPC, and if it has Enhanced Transistors. The

structure_types table holds layout specific information for every transistor or structure in the chip. Each structure_type row is annotated with a CD, width, pitch_left, pitch_right, poly_overlap, poly_elbow_below_active (poly to active space), active_corner_left (distance to step on left), active_corner_right, and four miscellaneous columns that are cell specific. This is a particularly important table as trends in test structure response to process non-idealities will be identified based on attribute values in this table. For example plotting change in gate length through focus vs pitch would be done by creating a query based on pitch_left and pitch_right values in this table. This table can also be used to describe other types of test structures, but since this chip was primarily populated with transistor based test structures, some of the attribute names are transistor specific. ELM structures for example only use the CD, width(is length), pitch_left and pitch_right attributes. Also, since some cells have non-standard designs, such as the cells with triangular transistors, miscellaneous columns are used to accommodate different layout parameters. For example for triangular transistors it is important to note that the top CD may be different than the bottom CD, but in the case of transistors with a step or corner on active the active step height may be important. Having miscellaneous columns allows them to change from cell to cell and hence hold important information without creating a lot of extra columns for all transistors. The benefit of having miscellaneous columns over storing the data in the comments field is that queries can be run a lot faster and are simpler, so there is a tradeoff between size and speed here. In general the strategy for this section of the database is to differentiate different test structures, which enables looking at small subsets of data. The details of all the attributes in all the tables can be found in Appendix B.

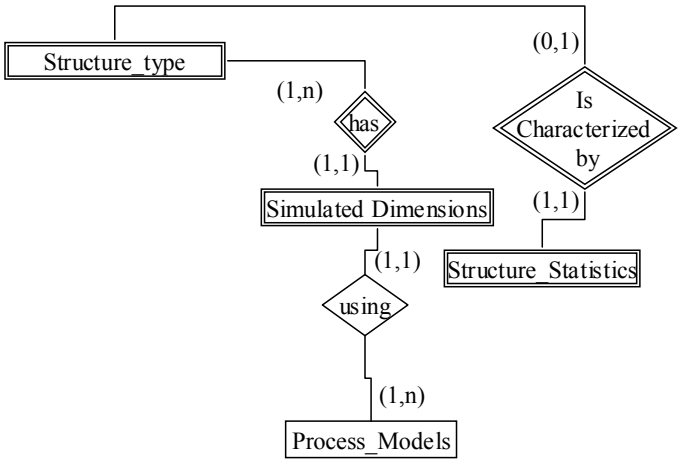


Figure 37 Section of the EER associated with simulated dimensions, currents, and effective lengths

The simulation section of the database stores simulated dimensions, currents, and effective lengths as well as the process conditions used in simulation. The process models table holds specific process conditions such as defocus, dose, slice lookup table filename, and misalignment. The slice lookup table filename is a hook for selecting slice current lookup tables that may correspond to different transistors or transistors with non-lithographic parameters changing. Specific process conditions can be uploaded into this table and then rows of this table can be used to drive the Parametric Yield Simulator(PYS). Each row in the simulated_dimensions table needs to have a unique process model and unique structure type referenced. The simulated_dimensions table holds most of the pertinent simulated data, such as average CD, average width, image slope, effective length for OFF current (L_OFF), Ioff, L_ON, Ion, etc. A finer granularity of simulated dimensions and image parameters is held in slice_dimensions, which holds simulated values, CD, Imax, Imin, Islope, for each slice in each transistor. As each simulated slice or dimension is associated with a structure type and process model, each simulation result is annotated with layout parameters as well as simulated process

conditions. This table can grow very rapidly as it can have up to every single combination of structure_type, of which there are about 15,000, and process model of which there are over 120. A structure statistics table was added on to help deal with this large amount of data by summarizing common statistics, such as standard deviation through focus for each structure type. This table has a one to one correspondents to the structure types table.

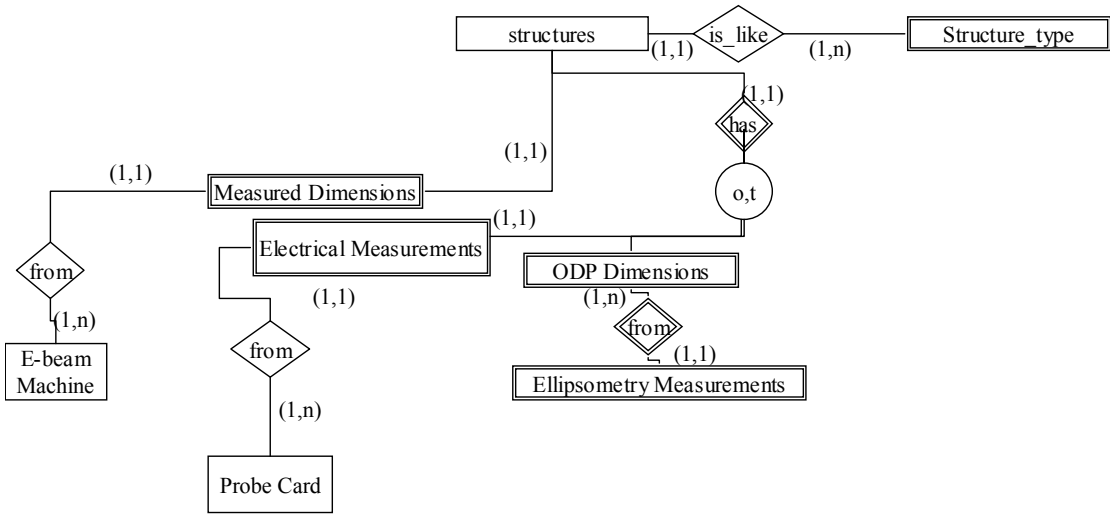


Figure 38 The measurement section of the database holds all data associated with the measurements from the measured currents to the temperature at the time the measurement was taken.

The experimental measurement section of the database holds all pertinent information about the actual measurement of the test structures. This is of particular importance as noise in the data can then be attributed to specific measurement noise sources such as temperature or probe card offsets. The structures table holds a row or entry for every measured transistor or structure on every wafer that can be uniquely identified by a structure_type and die. Each transistor or structure is then associated with measured_dimensions, electrical_dimensions, ODP_dimensions, and/or modeled

dimensions. Measured dimensions are physically measured dimensions in a CD-SEM. Only a portion of the structures are expected to have this form of measurement as it is very expensive. Each measurement is associated with a SEM table row, which includes the tool that was used, measurement conditions, and day of measurement. Electrical dimensions are dimensions that are extracted from measured currents, which are stored in the electrical_measurements table. Some test structures such as ELM have proven methods of CD extraction with high accuracy, but transistors are susceptible to a lot more noise due to random dopant fluctuation among other effects. High Vds values that increase Drain Induced Barrier Lowering as well as Enhanced Transistors can be used to increase the correlation between leakage current and gate length and hence reduce measurement error. Each current measurement is annotated with a row in the probe card table that holds the name of the probe card, the date of measurement, time and temperature. In general the strategy of this section of the database is to store all measurement data along with information such as temperature that could prove useful in the future.

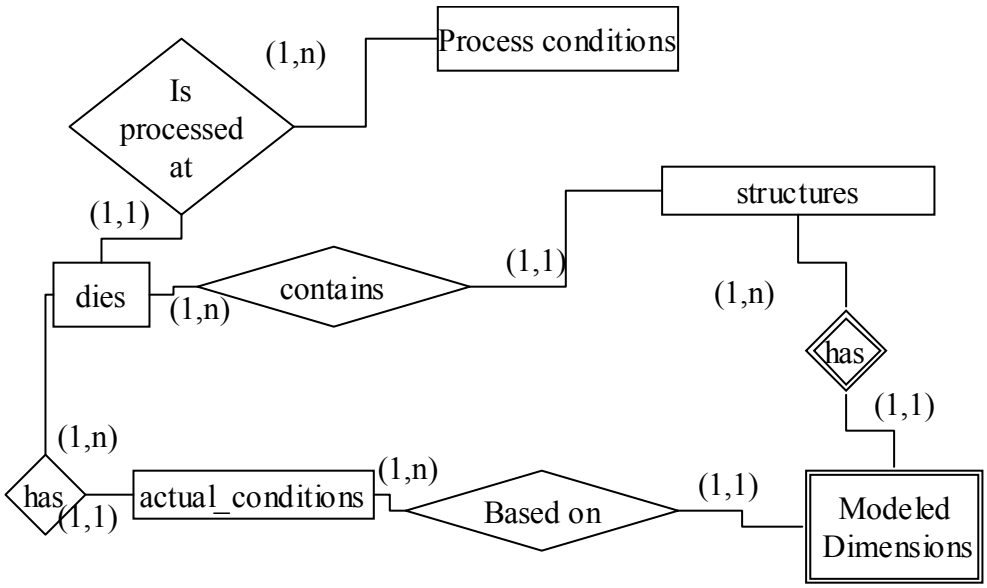


Figure 39 The process description section describes the programmed process conditions as well as deviations from ideal. This section also stores modeled dimensions, which are based on extracted process condition and test structure dimensions that were calculated from the programmed experiments.

The process description section of the database stores process condition data as well as deviations from set process conditions (Figure 39). The process conditions table holds process parameters such as defocus, dose, and misalignment for each die as they were programmed in the scanner. This table can be expanded if other process steps can be programmed or set outside the preset values. The actual conditions table is a manually updated table by database users based on the measured electrical dimensions of each die and an iterative process extraction strategy described in this chapter. The modeled_dimensions table holds expected gate lengths given conditions specified in the actual_conditions table. The primary strategy for converging on a set of actual_conditions or non-idealities is iterating until the error between measured_dimensions and

modeled_dimensions data is minimized. This process will be explained in more detail in section 5.4.

5.2. Implementing Structure with Flexibility

A key benefit of the database is that it forces structure into all data, independent of the source, which enables queries and comparison of data, but flexibility also needs to be incorporated to allow description of attributes that have not been predicted during the design of the database. If database users want to start keeping track of attributes that are not in the database design, they need to be able to do so without modifying the database schema. This is important as adding extra columns to a table after the database has been populated can lead to loss of data. For this reason every table in the database has a generic "comments" attribute that can be used for adding extra data with XML style tagging (or metadata). This means if you want to add an attribute to a table, you add a line to the comments field that looks like this "<attribute_name>value</attribute_name>. The XML style metadata can easily be parsed in a query with a REGEXP command or post processing can with any of the free XML parsers that are readily available. If queries need to be optimized, then data can be extracted from the comments field into helper tables that have specific columns for the data.

Another benefit of this design is that the comments field can be used for adding attributes to only a small portion of the rows. This has a performance benefit as adding an attribute or column that is only utilized by a few rows and is NULL for the rest is inefficient. An example of this flexibility is in the structure_types and cells tables. Most

transistors or structure can be fully described by the set of attributes in the structure_types table, but some specialized cells require some further detail. For example one type of cell has triangular transistors that are wider at the bottom than at the top. The structure_types table has four miscellaneous columns of type float, two of which can be used to describe the CD at the top and the bottom. To store a description of what the miscellaneous columns are used for, the comments field is utilized in the cells table and would look something like

<misc1>CD at top of gate</misc1>
<misc2>CD at bottom of gate</misc2>

Since the top and bottom CD are values that may be used in a query, separate miscellaneous attribute columns are important to have, but since the miscellaneous column descriptions are mainly data that just needs to be read and not evaluate in a query conditional statement, the miscellaneous column descriptions will be served well in the "comment" field. Hence built in flexibility can be used to account for unpredictable attributes as well as make the database more efficient. This flexibility will make the database adaptable even if the only access point is the database website.

5.3. Database Website as a Collaborative Platform

A key component and facilitator of the database serving as a collaborative platform is the web interface, which can be used for uploading, downloading, and querying data. The database website has been implemented in Ruby on Rails and can be accessed by any type of computer with Internet access. This means one student can upload measurement

data in the lab and another can run queries on the new data without the need for any direct exchange of files. In addition to storing experiment related data, the website serves as a platform for collaboratively mining the data. Queries can be saved, loaded, and even rated by different users. Queries can be reused and refined by multiple people with the hope that the most effective queries bubble up to the top.

Queries can be formed either in advanced mode where SQL commands are typed in directly or in basic mode where sets of conditionals are created to formulate a query. Basic mode is useful for beginners to learn how to build queries, but advanced mode is required for creating queries that perform more complicated data analysis such as finding variance or correlation. Once formed, each query can be saved and annotated with a name, keywords, and a description. All saved queries are saved in a separate database along with usage and rating data. This is the main collaborative aspect of the website and database as each query can then be loaded, reused and modified by all users. Since queries can prove to be long and tedious and/or complicated, this feature of the website can be very useful. The query page seen in Figure 40 lists queries along with a star rating and hit counter. Users can rate each query, which will help the best queries bubble up to the top. When several users upload dozens of queries each, it is important to have a quantitative metrics that can be used to identify the good queries. With access to old queries along with their descriptions it will be easier for novice users to utilize more complicated queries. As both queries in basic and advanced mode are self-explanatory it is not to difficult to customize them to ones particular needs, even with little SQL experience.

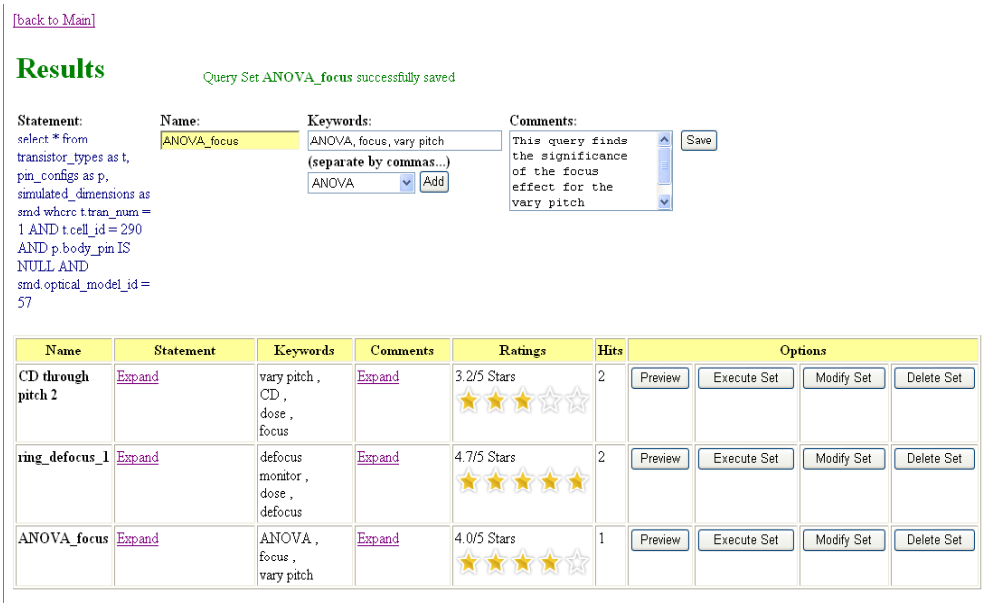


Figure 40 Screenshot of the load query web page. Queries can be rated, previewed, executed, modified, and deleted.

Currently the web page takes as input tab delimitated files, but limited effort can expand the capability to take in outputs from various tools directly. An application specific parser can easily be written in Perl and added as an extension to the website. This is an important aspect of the collaborative platform as developing tools that can parse data automatically will motivate students to use the database if their alternative approach is to use excel. As collaborative results can take some time to bear fruit, it is important to provide some instant gratification when asking students to collaborate out of their own free will.

5.4. Process Characterization/Data Analysis Strategy

The Process Characterization Strategy is split into two parts, identification of good process monitors and extraction of process conditions based on process monitor dimensions. A basic set of process monitors is chosen based on test structure response to

programmed process variations either experimentally or in simulation. Once a set of process monitors is identified, the response of those process monitors through focus, dose, and misalignment is characterized by either a lookup table or a compact model. Once a set of relations exist, it is possible to iterate to a solution (a set of process conditions) given a set of measured dimension. Although this project concentrated on dose, focus, and misalignment, this technique is expandable to looking at more sources.

5.4.1. Identification of Process Monitors

The identification of process monitors can be done either by analyzing measured experimental or simulation results of test structures over a range of process conditions. There are three basic types of process monitors: single variable monitors that are essentially only sensitive to one process variable, linear response process monitors that are easily modeled and can be characterized with a relatively simple compact model, and threshold-spike test structures that have a huge jump in signal (ex. Short or open) if a process crosses some threshold. Single variable monitors are the best type as they are orthogonal to other process non-idealities and can be used independently to lock in corresponding process parameters. Single variable monitors can be identified by looking at the variance of test structure dimension while varying one process parameter at a time. If a test structure has a low variance for all process parameters except one, it will probably be a good single variable process monitor. A more sophisticated method, that may need to be employed if somewhat noisy experimental measurements are made, is running an ANOVA analysis on each test structure with a full factorial experiment on all the process parameters. The measured dimension is treated as the response and the process parameters, such as focus, dose, and misalignment, as the factors. Process

monitors can be identified based on the p-values that are calculated for each test structure. This analysis can also be used for finding good linear-response test structures. Good linear response test structures should have strong main effects, but weak interactions (Figure 42). Low interactions p-values are potentially bad as they lead to more complicated equations that will be hard or impossible to invert (which is necessary for process characterization). Specifically if one of the terms in the equation CD=f(dose, focus, etc) is a product of two parameters, those parameters may be difficult to separate. Another approach to finding linear-response test structures is by looking at the stdev in response to a step in one parameter when swept across another parameter. For example, how much does the gate length change with a 40nm defocus change at five different dose levels. If the change is similar (variance is low) across all levels, then the structure is a good candidate for having a linear response. Finally threshold-spike test structures can easily be found by looking for a relatively large response with a small change in process parameters. For example in the defocus monitor, the leakage current may spike dramatically with a small change in focus that bleeds in just enough light to cause a dramatic change in the resist image(Figure 41). As there are over 15,000 test structures in the FLCC testchip, data mining the database should lead to sets of test structures that fall under each type of process monitor and can then be used to extract process parameters.

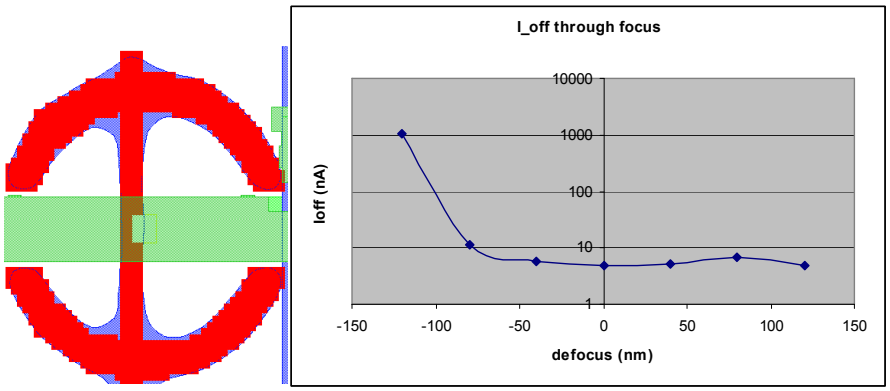
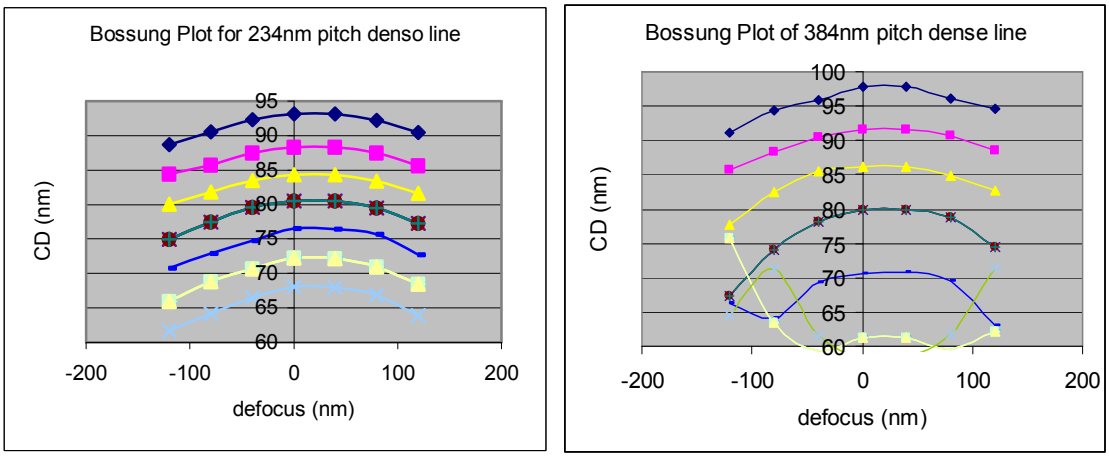


Figure 41 An example of a threshold spike monitor. There is a huge jump in leakage current between –80nm and –120nm.



(a) (b)

Figure 42 Example of a good linear response monitor on the left and a poor linear response monitor on the right. These are Bossung plots of a 80nm line. The left is a denso line at a 234nm pitch and the right is a dense line at a pitch of 384nm.

Since all the data is in a structured database, process monitors can be found by executing queries that can find characteristics such as those mentioned above. As process monitors or test structure sensitivity can change from technology node to technology node, a shotgun approach paired with the right data mining technique can mitigate this

issue. For example test structures with slightly varying layout parameters, such as pitch, can be data-mined for best fit or process sensitive designs such as the defocus monitor can be evaluated to see what group of process monitors they fit in. Since MySQL comes standard with basic statistical tools like finding averages and standard deviations, looking at variance will be the preferred method of identification. ANOVA analysis is also possible, but is relatively slow and unnecessary if there is a small amount of random noise or redundant data. Using the "group by" command in a query one can find standard deviations of groups of measurements based on various test structure attributes. Based on a specific "group by" condition, such as identical defocus values, the standard deviation and average can be automatically calculated within each group. This information can also be plugged into algebraic equations, such as those used to find correlation. ANOVA could also be implemented in SQL if a lookup table for the Fischer distribution was created. If intelligent queries are used to select small subsets of data, relatively complex calculations on each tuple or set of data can be done with little computational complexity. As the database is populated with data and used by different users, new queries and analysis strategies will be formed and refined.

5.4.2. Process Condition Extraction

Once good process monitors have been identified, they can be used to estimate a process condition and then iterate until dimensions of all process monitors fit the extracted process conditions. The basic strategy is to start with single parameter monitors that are sensitive to one process parameter and use linear response and threshold spike monitors to dial in the remaining process parameters one at a time. To estimate an initial guess of process conditions an upper and lower bound can be established by looking at

threshold spike monitors. Based on if a process monitor is at its high value or low value the threshold (plus an extra budget for error) can be used as a lower or upper bound respectively. If threshold spike monitors exist that can reduce the range of possible parameter values, new single variable test structures can be identified by testing structures within the limited range. The single variable monitors can then be used to dial in their respective values. For example Figure 43 shows a Bossung plot of a good candidate for a single variable dose monitor that can be described by CD = a*dose + b or dose = (CD-b)/a. The line has about a 1nm/1% dose change sensitivity to dose, but does not change by more than 1.7nm across all focus conditions for a given dose. Since the goal of an initial guess is to be close and not necessarily accurate, this is a very good margin of error. It may be useful to see if new single parameter process monitors exist in the now smaller range of possible parameter values. Once all the single parameter monitors have been used, the linear response monitors can be used to lock in parameters that do not have single parameter models. For example a linear response monitor could follow the following equation CD = f(dose) + g(defocus) or more specifically CD = a*dose + b*focus^2 +c. If an initial estimate of dose is known then defocus can be extracted as sqrt(CD – a*dose – c)/b. Once dose and focus are figured out using test structures that are insensitive to misalignment, misalignment sensitive structures can be used to figure out overlay errors (this time treating dose and focus as known values). Using this approach and a few select structures, it is possible to get initial estimates for all process parameters.

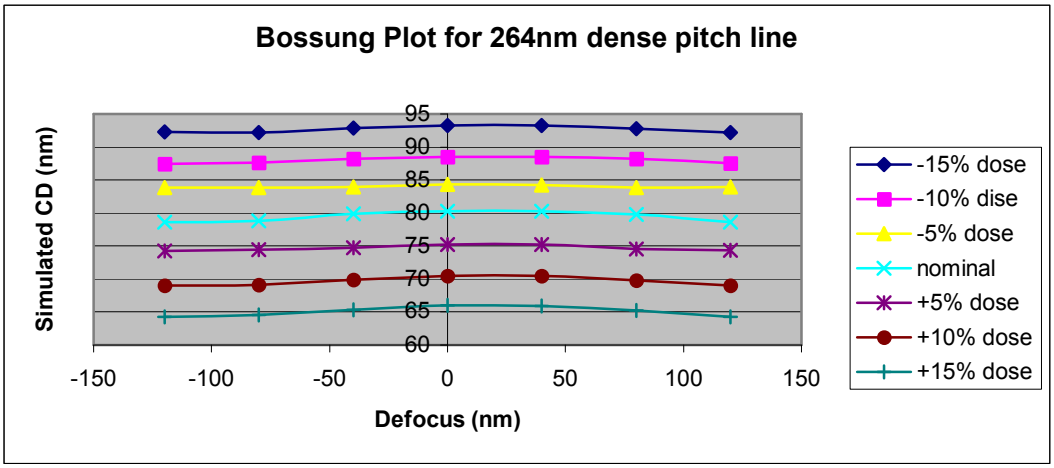


Figure 43 Example of a good single variable dose monitor. Bossung Plot of 80nm dense pitch line at a pitch of 264nm. The CD does not change more than 1.7nm through focus for all dose settings.

Once an initial estimate of process parameters is known, one can iterate to a solution that minimizes the error between "expected dimensions" given extracted process conditions and measured dimensions. "Expected dimensions" are held in the modeled dimensions table and can either be calculated through compact models or lookup tables. One slightly simplified approach using lookup tables assumes that between any adjacent process conditions, for example within one defocus step, the CD varies linearly. With this assumption it is possible to interpolate critical dimensions at any process condition within the range of those simulated. Using lookup tables has the benefit of being able to create a model without finding a potentially complex equation to fit the data. The iterative process can either be done manually as a database user can guide the direction in which process conditions change or automatically where a range of process conditions around the initial estimate is evaluated (NOTE: the automatic process can become very long if a lot of process parameters are accounted for). Using the automatic method, the process condition resulting in the minimum error is used as the next estimate and again a range of process

conditions, centered on the new estimate, is evaluated. If the current estimate is the lowest error, then the iterative process is done. This process can be susceptible to finding local minimums and ignoring global minimums, but this can be mitigated by using a larger number of process monitors for the initial guess and by increasing the range of process conditions around each estimate that is simulated. As multiple users will be looking at data from the same testchips that were printed under the same conditions, comparing results will lead to further reinforcement if they agree.

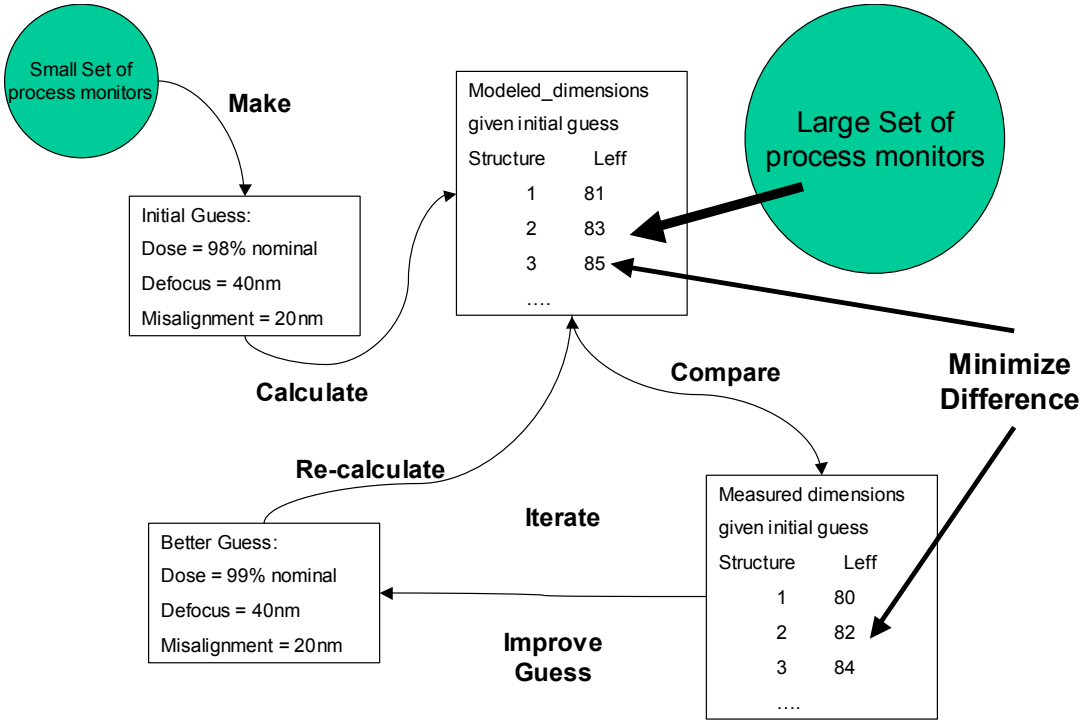


Figure 44 Iterative process of extracting process conditions. First make an initial guess using a small set of high quality single parameter, linear response, and threshold spike test structures. Then iterate using a much larger set of process monitors until error is minimized.

5.5. Database Optimization

One final detail that can make a huge difference in scalability and usability is database performance, which is a very strong function of design. Two of the most effective

techniques for improving query runtime is using indexes and helper tables. Using indexes is a standard strategy in database design, but using helper tables is a strategy more specific to using databases for making calculations. Traditionally databases have been utilized for storing data in an easily accessible fashion, such as finding contact information for customers or getting specs on a specific product. Those databases were optimized for minimizing the amount of redundant data or data stored and the time it takes to execute queries[50,51]. The collaborative database on the other hand not only requires easily accessible data, but numerical data mining as well. This can become an issue if the results of calculations are what end up being evaluated in a query. Hence a helper table, which stores results of various calculation such as standard deviations and averages can prove to be useful. In a sense this is using the database to reduce the number of redundant calculations. Instead of making a calculation the helper table can serve as a lookup table for the results. The helper table can be built as a view, which is a result of a query or as a separate table. Views offer the benefit of being automatically updated, but using separate tables is a lot faster and more efficient. The big benefit of a separate table is an index can be used that will help filter out results even faster. The way an index works is it sorts all the data in the table based on the values in one or more columns and stores the sorted list in an index table. This way when an entry with a specific value in the indexed column is needed, it is not necessary to read through the entire table to find it, which takes O(n) time. Instead the index table, which is generally stored in a binary tree data structure, is used to find the value in O(log(n)) time. For large tables using indexes saves orders of magnitude time. Using a helper table the performance savings is even

greater as storing the results of calculations trims down the amount of data from all the variables to just the results, which are then indexed for quick retrieval.

This strategy has been used in the Collaborative database to make queries that find process monitor run faster. A structure_statistics table was created that stores averages and standard deviations of test structure responses across dose, focus and misalignment. An example of a more complicated statistic is the standard deviation of CD change due to one step in focus across different dose values. In this query one needs to use nested queries where the first query finds the standard deviation across one specific step in defocus, say 40nm to 80nm, and then an outside query that averages across all focus steps. This query can be seen below

```
update (select id, avg(stddevs) as ans
from (select p1.poly_defocus, st.id as id,
stddev(abs(s1.CD_ave-s2.CD_ave)) as stddevs
from structure_types st, simulated_dimensions s1,
simulated_dimensions s2, process_models p1, process_models p2
where s1.structure_type_id = st.id and s2.structure_type_id =
st.id and s1.process_model_id = p1.id and s2.process_model_id
= p2.id and p1.poly_dose = p2.poly_dose AND p1.misalignment_y
= 0 and p2.misalignment_y = 0 and p1.slice_OFF_table_fname
REGEXP "10X" and p2.slice_OFF_table_fname REGEXP "10X" and
p1.poly_defocus=p2.poly_defocus-40 group by st.id,
p1.poly_defocus) as t1
group by id) as tt1, structure_statistics ss
set stdev_step_through_focus = tt1.ans
where tt1.id=ss.structure_type_id;
```

This data is used for finding linear response structures that have a small stdev_step_through_focus value. The only caveat is a good process monitor will vary significantly so one wants a relatively high standard deviation across focus. To create such a query strictly on the data one must first sweep dose while keeping focus constant and then sweep focus while keeping dose constant. The alternative is to run two queries, one to find the stdev of step sizes for given focus conditions and one to find stdev of CD

through focus and store the results in the structure_statistics table. Then query the helper table to identify the best linear response monitors. For a table with 2154 structure entries, the single query approach takes 129 secs, while the helper table method takes 34 secs, a quarter of the time. The real benefit is that subsequent queries using the values in the helper table will only be 0.04secs or five orders of magnitude faster. Helper tables will not only make queries run faster, but will make writing queries easier to understand. An example of this is the original query written to find threshold spike monitors, which took 9 hours and 26 minutes. Compared to 36 seconds when the problem was divided into parts and conquered by storing intermediate results in the structure_statistics table. The divide and conquer method could then be rebuilt into a single query that took a little over a minute, but the operations where not obvious from the start. Since most data mining will be done on simulation and experimental data, two helper tables were added to the database with preset statistic columns, 2 miscellaneous columns, and a comments column for needs that exceed the original column count.

5.6. Conclusion

A database and data analysis strategy has been presented that can be leveraged in multi-student testchips and experiments. The database has been designed to enforce enough structure as to enable queries or comparisons between sets of data, but also with flexibility to adapt to requirements that may be present in the future. The database has 25 tables that consist of 390 different attributes that store layout specific information, simulation and experimental results, process conditions, as well as results from data mining that may be done by various users of the database. Hence each test structure is annotated with 390 attributes and potentially hundreds if not thousands of simulation and

measured data points. The relational database model allows for efficient storage and access to each bit of information in a logical and organized fashion. Comparing sets of data becomes a matter of running a set of queries, which enables very fast hypothesis testing. In addition a complementary website allows access to the database from any computer with internet access. Granting access not only to data, but also queries or tools for doing data analysis. Queries can be saved, annotated with keywords and descriptions, rated by other, loaded and modified so that the best queries will be reused and refined with time. A set of queries already forms a basis for a process characterization strategy that is split into a process monitor identification stage and a process parameter extraction stage.

Adding indexes to most columns that will be used in queries as well as adding two additional helper tables has greatly reduce query runtime, simplified query complexity, and helped reduce redundant calculation of the same statistics. A big challenge in building the database has been creating enough flexibility for the database to adapt to the data, which is made possible through ambiguous comments fields that can be populated with XML style data. Similar strategies may be used in the future where additional helper tables are used to store data extracted from the comments field in a way that could be more efficiently queried. The hope is that this database will be easily adaptable so that it can be utilized by different groups for various projects where data from multiple sources can be combined and analyzed. This database can hence be used to facilitate collaboration and maximize utilization of multi-designer testchips.

6 Enhanced NMOS Testchip Dry-Lab Simulation

The tools developed in this thesis, consisting of the Parametric Yield Simulator for process simulation and the Collaborative Database for data aggregation and analysis, have been systematically applied to the test structures on the FLCC Enhanced NMOS testchip to demonstrate the efficacy of the high volume electrical test structure process characterization strategy. The three goals are to evaluate the efficacy of designed test structures, the Parametric Yield Simulator under heavy simulation loads, and the process extraction strategy in the presence of random variations. The strategy here is based on simulated electrical currents, which would corresponds to measured leakage currents on experimental wafers. As an initial proof of concept only dose, defocus, and misalignment were varied on the poly layer, but this strategy should be expandable to looking at more process dimensions. Process test structures were evaluated in the simulation portion of the database based on sensitivity of leakage currents and effective gate lengths to programmed process parameters. The experimental portion of the database was populated with simulated currents under randomized process conditions with LER noise added inside the PYS. The goal of this dry-lab simulation experiment is to mimic a planned Enhanced NMOS testchip experiment at SVTC for process extraction and characterization. This chapter will first describe the dry-lab experimental setup, then the

performance of various test structures, and finally the accuracy of the process extraction strategy.

6.1. Experiment Setup

This dry-lab simulation experiment is designed to mimic a set of planned experiments at SVTC where measured leakage currents from Enhanced Transistors would be used to characterize the main sources of transistor performance variation. Model based OPC and aerial image simulation was done using Mentor Graphics Calibre Work Bench, TSUPREM4 and MEDICI were used to model the process. Model based OPC was relatively easy to implement as it does not require a complex set of rules in rule based OPC and produced good results, at least under nominal conditions. Rule based sub resolution aspect feature (SRAF) insertion is used in most OPC algorithms, but ignored in this test chip as increased sensitivity to focus and dose is preferred to help discover parameter sensitive layouts. The big benefit of model based OPC is that it biases all features so that under nominal conditions isolated and dense lines should print with minimal Edge Placement Error (EPE).

An optical model was built with lamda of 248nm, NA of 0.78, and quasar illumination with a sigma_in/sigma_out of 0.72/0.92 and a 40 degree illumination angle (Figure 45). This is a deviation from the lithography system at SVTC where a 193nm scanner is used with an annular illumination with an NA of 0.75. A wavelength of 248nm was used instead of 193nm in an effort to reduce the k1 from 0.44 to 0.35 for the minimum pitch of 224nm. This is an aggressive illumination scheme that yields good resolution, but more layout dependent sensitivity. This change does have some significant ramifications on the results in this study. First of the Rayleigh Unit is increased from 158nm to 203nm, which

increases the depth of focus and hence makes a 1nm change in defocus 33% harder to detect for most test structures. Secondly the through pitch performance will be skewed as the dead zone is bigger and pushed from ~280nm-490 to ~350nm–600nm. The iso-focal pitches that have little or no sensitivity to focus are also larger than if 193nm light was used. Finally the optical radius of influence is increased by the ratio in k1 factors, so 2D patterns such as in the corner rounding test structures may be more sensitive at this wavelength and as the programmed defocus monitors have not been designed for 193nm, they are not going to be as sensitive. Since this process characterization strategy is aimed at cutting edge processes and not specifically the SVTC flow, these changes were intentional as most aggressive processes are at k1 factors well below 0.44. The general strategy of this testchip was to have broad coverage and to be prepared for the unexpected, so this in a sense is also a test of how comprehensive this generic set of test structures is.

The film stack for both the poly layer and active or FOM layers have been specified according to resists and BARC thicknesses used in the poly process. Optical models were generated for a matrix of dose and defocus values with dose varying in 2% steps from 94% of nominal to 106% of nominal and defocus varying in 40nm steps from –120nm to 120nm . The focus window range of –120nm to +120nm was chosen based on how much focus is expected to actually vary in a production setting[56]. This range should capture defocus in the scanner as well as that due to CMP non-uniformity. Misalignment was swept in 10nm increments from –60nm to 0nm and 20nm steps from 0nm to 60nm in the y-direction for nominal dose and defocus conditions. Misalignment in the x-direction

was ignored as it is analogous to misalignment in the y direction if the overlay test structures were flipped by 90 degrees (which they were in some instances).

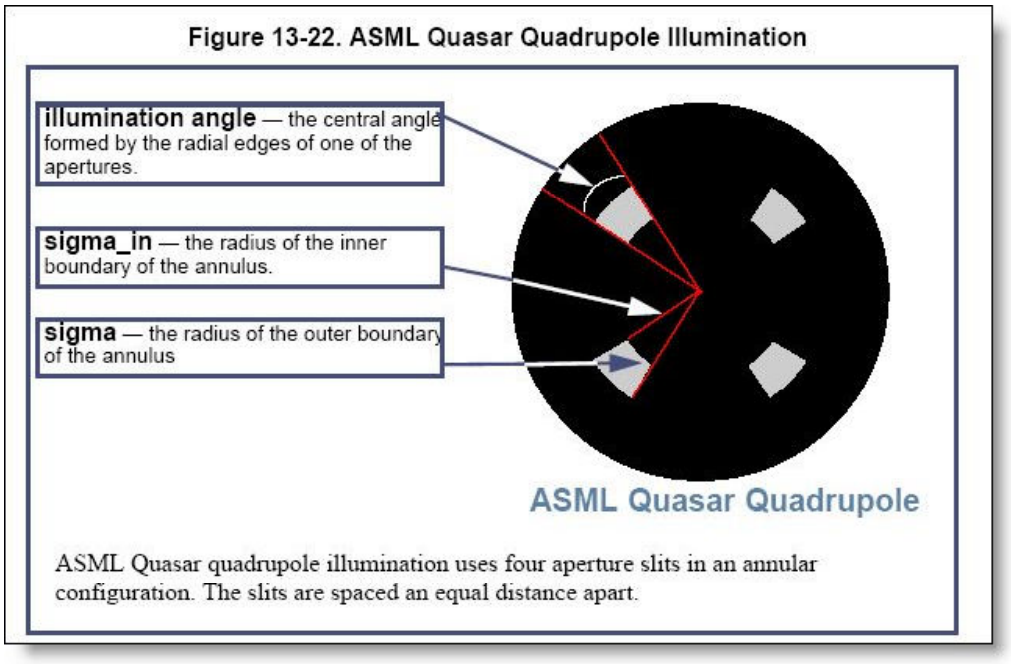


Figure 45 Quasar Illumination was used for simulation with a sigma of 0.92, sigma in of 0.72, illumination angle of 40 degrees, lamda 193nm and NA of 0.78.

The slice lookup table library of 10nm slices was built for the non-rectangular transistor model from TSUPREM4 (Version X-2005.10-0) process simulation and MEDICI (Version X-2005.10-0) device simulation. The TSUPREM4 process deck was built from an approximation of the process runcard that would have been used to manufacture the Enhanced NMOS wafers. Looking at the plot in Figure 46 of leakage current vs gate length, the resulting flow results in stable transistors down to about an 80nm gate length. As some of the process parameters had to be guessed, more fine tuning would lead to a stable 65nm transistor. Since the testchip was designed with 80nm, 100nm, and 120nm transistors, this ended up being a good match. The three different transistor sizes were intended to be a hedge in case the 80nm devices did not work.

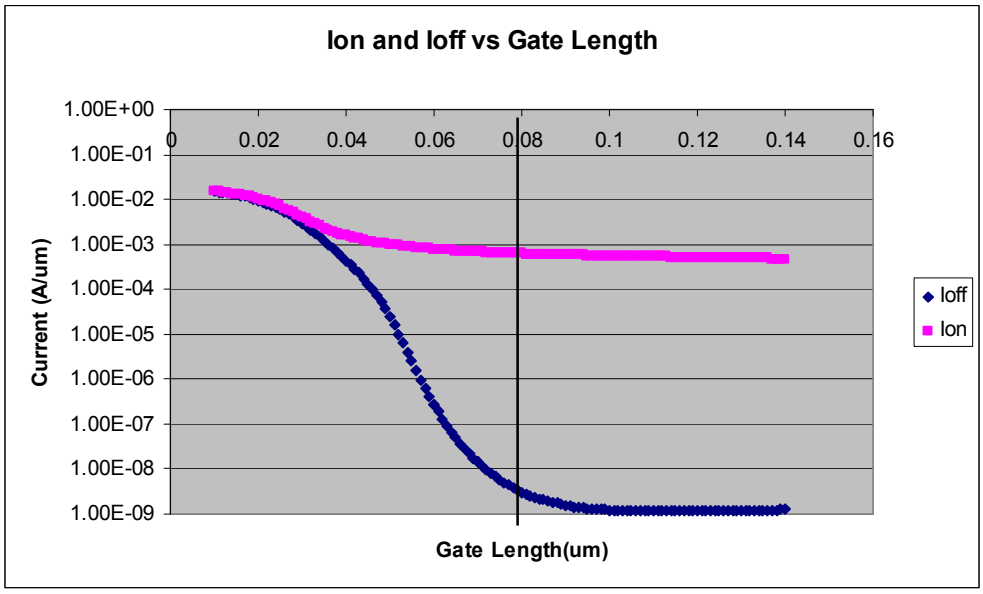


Figure 46 Ioff and Ion vs gate length for the standard transistor. Notice the leakage current starts climbing around 80nm meaning the 80nm transistor is still pretty stable.

In order to create enhanced transistors that has high leakage sensitivity to gate length at 80nm, a second slice lookup table was built from a similar process flow, but one with a 10X higher LDD implant. The resulting 3X increase in sensitivity of leakage current to gate length can be seen in Figure 47. The entire Ioff vs gate length curve has been shifted about 20nm to the right so the 80nm enhanced transistor will behave closer to a 60nm standard transistor, but the gate length will change as for an 80 nm transistor.

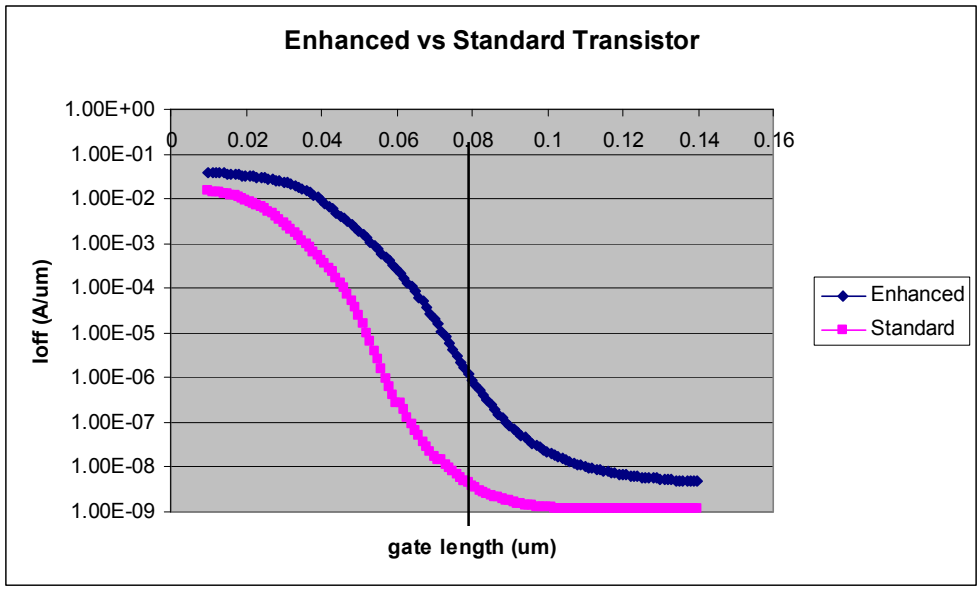


Figure 47 Ioff vs gate length for Enhanced vs Standard Transistors. The 80nm enhanced transistor has a 233X higher leakage current and 3X higher sensitivity to gate length

Since MEDICI is a 2D device simulator, transistor width effects had to be supplemented from a 90nm BSIM model. Specifically, the fitting constants for the exponential explained in chapter 3 were fit to the BSIM model and applied to estimate edge slice currents. Center slice currents were taken from simulated current densities in MEDICI to capture short channel effects for the enhanced NMOS and standard transistors. The use of MEDICI vs the BSIM model for building the slice lookup table facilitated the use of enhance transistors for Enhanced Transistor Electrical CD Metrology.

Simulation results with programmed process parameters were used to populate the simulation portion of the database and simulation results with randomized process conditions and added noise were used to populate the experimental part of the database. The simulations with programmed process parameters correspond to using a Focus Exposure Matrix that would be used to characterize the signature of dose and focus

changes for each test structure. To mimic experimental noise, Line Edge Roughness (LER) and random measurement error were added to each simulation that was used to populate the experimental portion of the database. LER was added in the device module as two sine waves, one with a correlation length of 1um and one with a correlation length of 800nm, both with a randomized uncorrelated phase and an amplitude proportional to the Image Log Slope (ILS). ILS has been shown to be correlated with the amount of line edge roughness[52]. The amplitude was calculate as (6-ILS), where ILS ranged from 1 to 3.5, so at worst case with an ILS of 1 a single slice could be changed by +/-10nm.

Once the current through the non-rectangular transistor was calculated, it was the only value uploaded to the experiment portion of the database as that would be the only form of measurement data in silicon. Each extracted gate length was rounded to the nearest nanometer, which added an extra random component of noise. Figure 48 shows that the added LER and rounding error leads to about 4.8nm 3-sigma CD variation for an identical isolated feature in a single die, which is pretty bad as across chip linewidth variation at the 90nm node is supposed to be 4.7nm, but this is for all patterns, which includes pattern dependent variation[53]. If different pattern proximities are included, the 3-sigma variation for 80nm lines increases to 17.2nm.

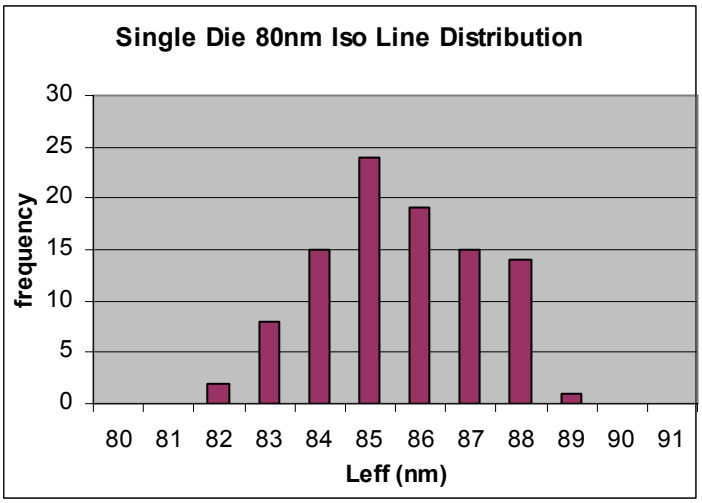


Figure 48 CD distribution of isolated 80nm gate on one of the "experiment" dies. The random LER noise lead to a 4.8 m 3-sigma CD distribution.

6.2. Process Monitor Identification and Evaluation

With simulation results from each test structure over a range of process conditions uploaded in the database, the process monitor identification strategy described in chapter 5 was executed. The goal is to show that the features of the database system facilitate rapid interaction with the data to identify systematic effects, residual noise, parameter specific test structures and characterization strategies. Process monitor identification is a key part of the "shotgun" approach used in this testchip. The "shotgun" strategy is supposed to leverage the database and wide array of test structures to capture all types of variations. When looking at a specific process parameter, it is important to identify the most pertinent structures and then use them individually to gain a more physical understanding of how things vary.

This section will describe the results of the three different types of queries used to identify threshold spike, single parameter and linear response test structures as well as

some results of specifically designed overlay process monitors. The threshold spike monitor is a monitor that has a spike in signal across one step in one process parameter. Single parameter monitors are only sensitive to one parameter and linear response monitors vary across multiple process parameters, but do so in an independent manner. For example variation due to focus is independent from variation in dose, but dose does have an impact. Sections 6.2.1-6.2.3 will address each of the three groups and section 6.2.4 will address overlay test structures that consist of many transistors and were described in Section 4.4.4.

6.2.1. Threshold Spike Monitors

Threshold spike monitors were identified by identifying structures that had at least one change in signal that was at least 10X larger than the average change while varying one parameter. To reduce query time the structure statistics table was used that had pre-computed statistics for each structure. Both the OFF current and Leff were looked at while using the enhanced transistor slice lookup table. Query results showed the expected suspects such as the defocus ring (Figure 41) as well as some unexpected structures that actually had the strongest response. One such structure was originally designed to help validate the non-rectangular transistor model by showing that slice position matters (Figure 49).

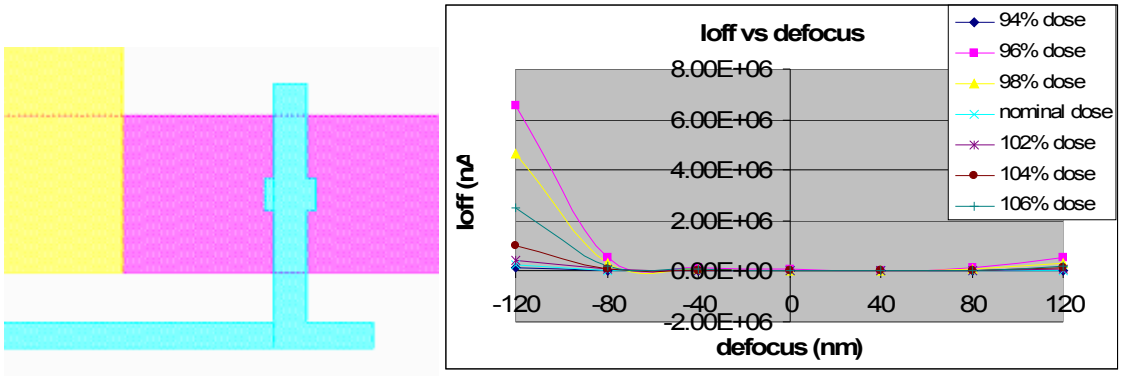
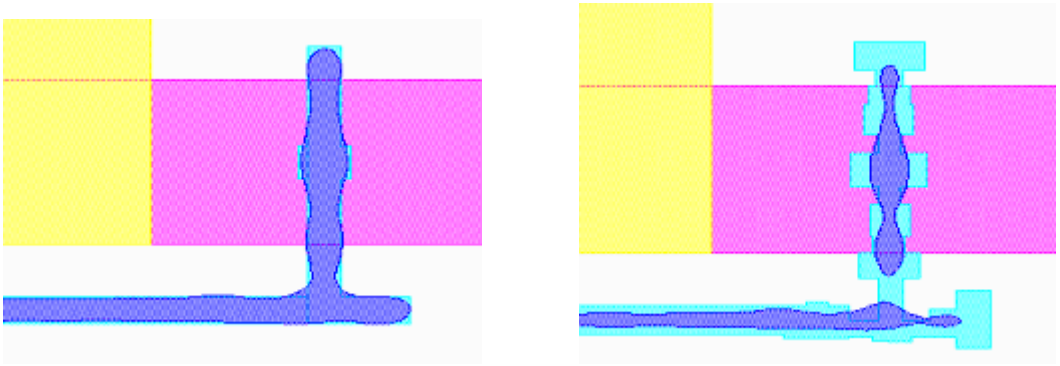


Figure 49 Non-rectangular transistor model test structure. Although not its designed intention, this is one of the best threshold spike structures in the testchip.

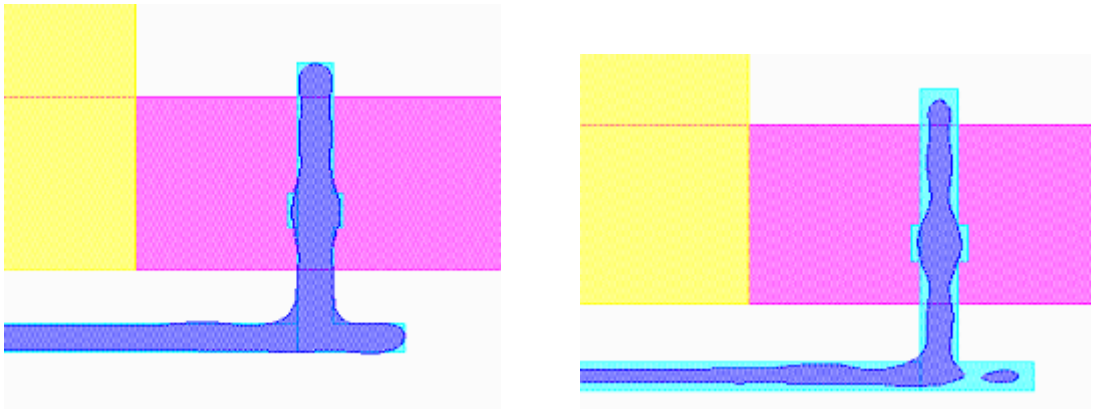
Simulation results showed that this is one of the best candidates for a threshold spike monitor as this structure has a synergy between the line end, the bump, and the poly elbow that causes two significant pinch points through focus. Figure 50 (a) shows how the figure simulates at nominal conditions, where model based OPC clearly did a good job. Figure 50 (b) shows the same transistor simulated with 6% overdose and out of focus showing drastic changes in CD. Particularly in two pinch points where leakage current ends up dominating.



(a) (b)

Figure 50 The non-rect structure simulated at nominal conditions (a) and at 6% overdose and -120nm defocus (b). Figure (b) also shows the post OPC layout in the background. The hammerhead, bump, and poly elbow lead to additive necking on both sides of the bump.

This phenomenon is a product of illumination conditions and additive spillover effects that demonstrate the potential benefits of a shotgun approach. A similar structure in Figure 51 shows how a small change in the location of the bump looses this effect. The only difference between these two structures is the location of the 120nm bump, which in the second case is 80nm lower.



(a) (b)

Figure 51 A similar non-rectangular structure with the 120nm bump 80 nm lower does not show the drastic change through focus and dose. (a) shows the structure at nominal conditions and (b) shows the structure at 6% overdose and -120nm defocus.

The addition of the 120nm bump in the middle made the transistor or gate 2X more sensitive to defocus than an isolated line (Figure 52,Figure 53).

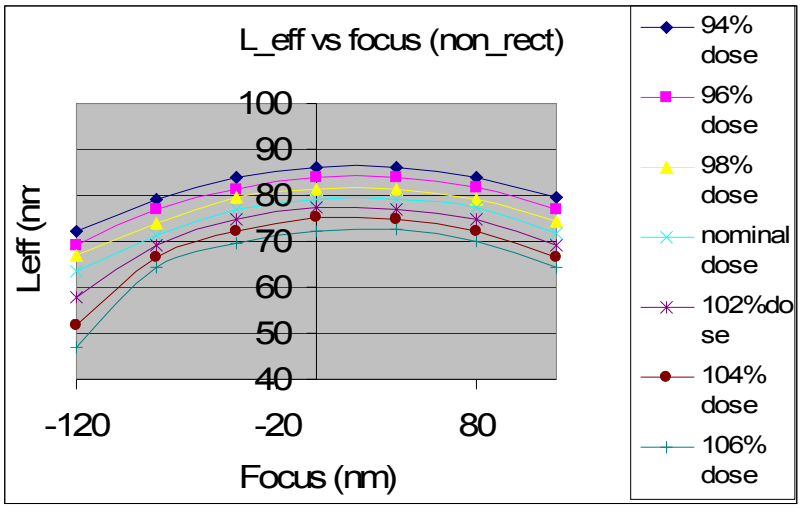


Figure 52 Bossung plot of non-rectangular structure. Notice the last focus step at higher doses has an extra large step in effective gate length. The gate is pinching at these conditions.

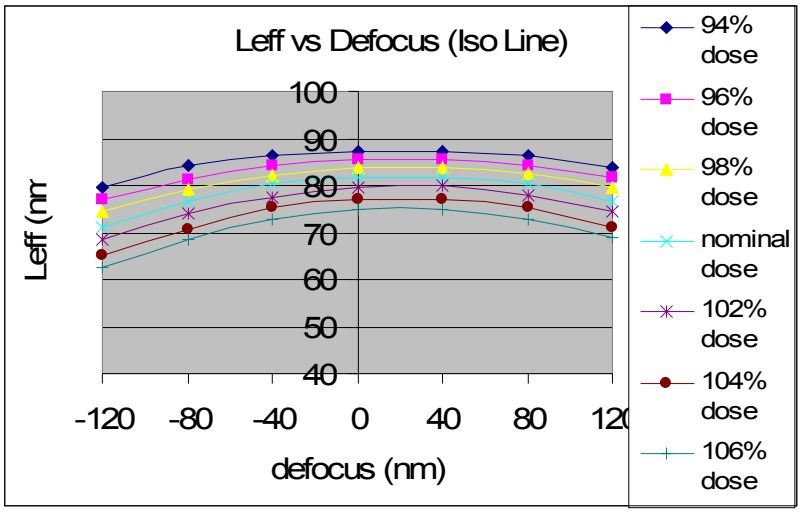


Figure 53 Bossung plot of an 80nmisolated line. Notice the range of CDs on this plot is about half of the non-rectangular structure shown above.

One other aspect of this test structure that made it stand out is that it is relatively insensitive to small amounts of misalignment. The only structure that beat this one was an overlay test structure that had a gate that did not completely overlap the active region.

As line end shortening is very sensitive to dose and defocus it acts well as a defocus threshold spike monitor, but is very susceptible to misalignment so is not as good an initial structure to look at. The programmed ring defocus monitors actually did not perform as well mainly due to not being tuned to this dose and illumination. Chapter 7 describes a second-generation electrical monitor that is sensitive to defocus over a broader range of doses. As far as threshold spike monitors for dose, none existed as the CD change due to dose is generally pretty constant from dose setting to dose setting. Misalignment on the other hand has good threshold spike test structures, but these are very sensitive to dose, focus, and other sources of line end pullback. Misalignment is hence best addressed looking at a group of transistors and will be discussed in section 6.2.4.

6.2.2. Single Parameter Process Monitors

Single parameter process monitors have high sensitivity to only one process parameter and little response to other parameters. They can be found by looking at the standard deviation of effective gate length across a given process window. These monitors serve as a foundation or initial step to guessing process values as good estimates can be attained from a simple equation.

Unfortunately all structures have a strong response to dose, so only dose specific single parameter test structures exist. It is thus essential to first identify test structures for determining dose with little confounding effects from other parameters. Primarily these are structures with a high poly overlap and a dense pitch that is stable through focus. The graph in Figure 54 shows the STDEV of Leff through focus. One can see dense pitches performing the best, followed by a second minimum somewhere between 400nm and

500nm. There is a hump at ~ 350nm due to the 3rd diffracted order not completely being caught by the lens.

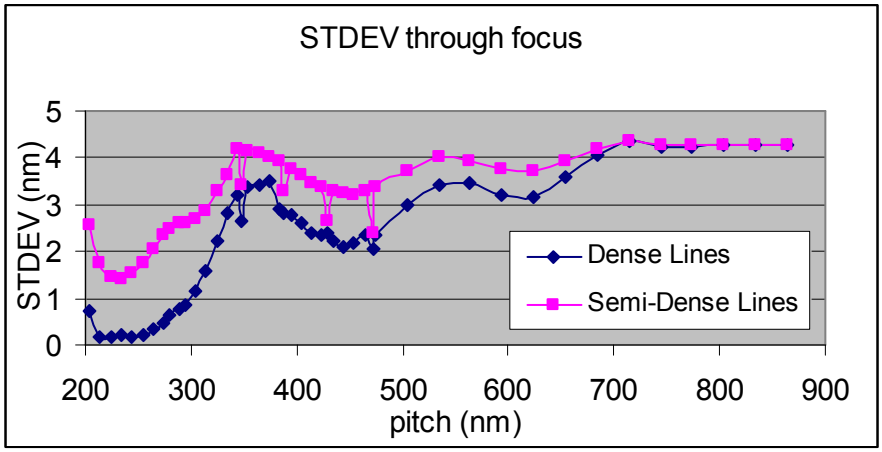


Figure 54 Standard deviation of effective gate length through focus for various pitches. Dense pitches have the smallest stdev.

The chart below shows that both sensitivity to defocus and misalignment can almost be negligible. In order to maximize sensitivity, the standard deviation to dose is also looked at and the 214nm pitch is identified as the best candidate. Not only is the Bossung plot flat, this pitch is unusually sensitive to dose. All the points in Figure 56 can be fit with the equation CD = -1.6*dose (% of nominal) + 272.24nm to an accuracy of 0.5nm. Hence dose can be extracted from effective length as dose = (272.24-CD)/1.6.

pitch_left	pitch_right	stdev_focus	stdev_misalignment	stdev_dose
274	274	0.11	0.31	3.37
214	214	0.11	0.95	6.51
280	280	0.12	0.35	3.73
264	264	0.13	0.05	3.43
284	284	0.14	0.32	3.69
244	244	0.15	0.4	3.41
224	224	0.15	0.42	4.32
254	254	0.17	0.07	3.71
304	304	0.19	0.06	3.73
234	234	0.21	0.46	3.69
288	288	0.22	0.29	3.66
248	248	0.29	0.24	3.91

Figure 55 Chart of sensitivities (or stdev) to defocus, misalignment and dose. A pitch of 214nm has a low sensitivity to defocus, but very high sensitivity to dose.

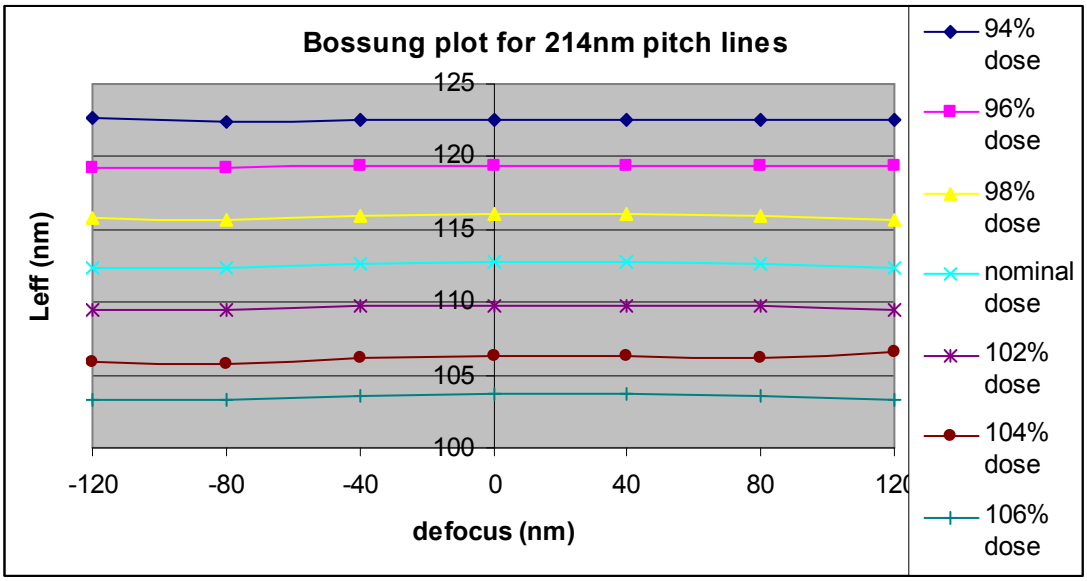


Figure 56 This Bossung plot shows a flat response to focus and a very significant ~ 1.5nm/% sensitivity of Leff to dose.

6.2.3. Linear Response Process Monitors

Linear response process monitors can be sensitive to more than one process parameter, but the contributions from each process parameter are additive so can be easily separated.

Linearity is important as it makes it possible to invert the relationship and extract a process parameter given a linewidth and the rest of the process parameters. Linearity of course is commonly the case as a lot of the process parameter affects are independent and their linewidth changes are additive. Even in the case of dose and focus, two parameters that have a significant impact on image quality, the two can be separated so that CD = f(focus) + g(dose) + h(everything else). To search for these structures a query is created that compares the change in gate length from one focus condition to the next across all doses in the process window. Since single parameter monitors are a subset of linear response monitors, a condition has to be added that checks for a minimum standard deviation across focus. Misalignment sensitivity can also be filtered out by looking at the stdev through misalignment. The structure statistics table is used again to simplify and speed up the queries.

The results of the most obvious query had a pretty intuitive result, a lot of isolated lines, which have high sensitivity through focus and at least for bigger CDs negligible change in step size at different doses. The 130nm isolated transistor showed a high standard deviation through focus of 4.3nm with very little sensitivity to misalignment and variation across dose settings. This can be seen in Figure 58 where each curve of CD through focus at different dose levels is in a sense parallel to the next. In this case it is fairly simple to create an equation in the form of CD = f(dose) + g(focus). Which can then be inverted to focus = f^{-1}(CD-g(dose). In fact the isolated 130nm line can be described by the fitted quadratic equation for focus and linear for dose (Figure 57). This equation is accurate to within 0.5nm for all simulated points and to within 0.2nm for most simulated points. If this test structure is analyzed with the single parameter dose

structure, which gives the dose value, the focus value can be calculated with pretty good accuracy. Multiple isolated lines can be looked at and averaged to reduce random sources of noise such as LER.

$$CD = f(focus) + g(dose)$$
$$CD = -0.00065 * focus^2 + 0.266 * focus + 124.28 - 0.85 * dose + 89.92$$
$$CD = -0.00065 * (focus - 20.46)\text{^}2 + 124 - 0.85 * dose + 89.92$$
$$focus = \pm \frac{\sqrt{(CD - 213 + 0.85 * dose)}}{0.00065} + 20.46$$

Figure 57 This equation describes the behavior of a 130nm line seen in Figure 58

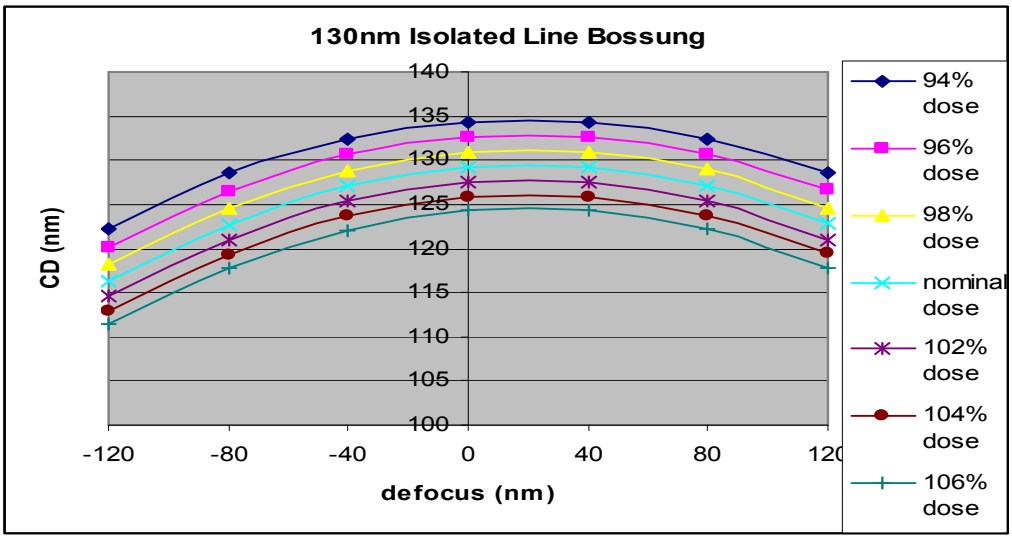


Figure 58 Bossung plot of isolated 130nm line/transistor. Notice how each curve parallels the next, showing that an isolated line will work well as a linear response monitor. Once the dose is know the focus can be read off the plot if the level of defocus is greater than +/-40nm.

Isolated lines offer high focus sensitivity and with enough poly overlap are insensitive to misalignment, but they have little sensitivity at low defocus values, where a lot of the action may happen. For this reason the search results had to be expanded to look for targets that may not be as insensitive to misalignment and have slightly varying

performance through dose, but at least a couple nanometer change from 0nm to 40nm defocus. The strategically designed defocus targets show to do the job well as they do not have a symmetric profile around focus and change significantly between –40nm and +80nm defocus (Figure 59). If the isolated lines end up having a measured CD similar to in focus the single ring defocus targets will have to be used. The relationship between focus and CD is established using a lookup table.

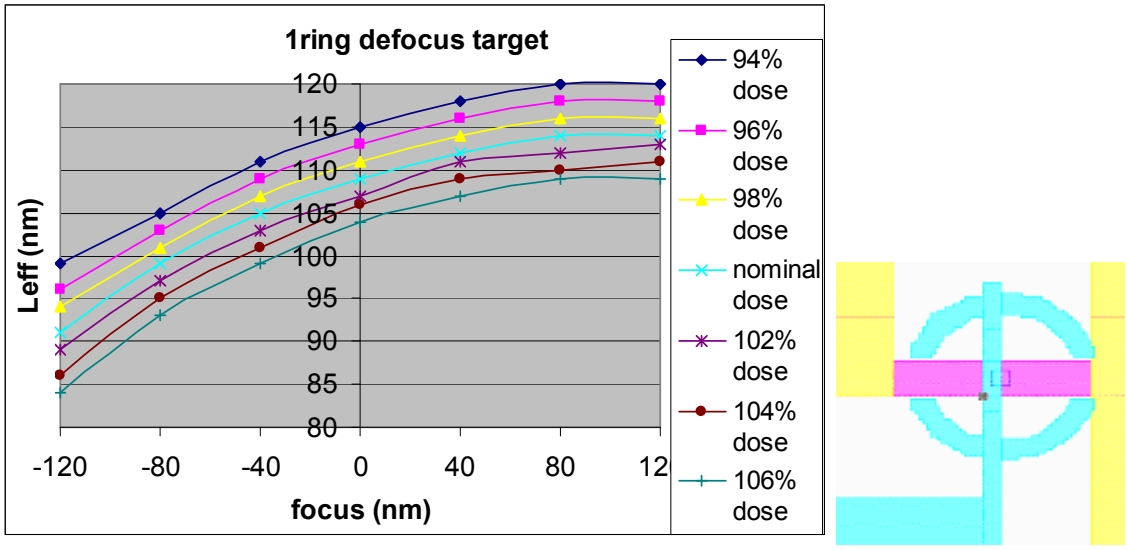


Figure 59 Single ring defocus target and corresponding Bossung curves. Notice the asymmetric behavior through focus, which leads to high sensitivity at low defocus values. If the defocus is between –20nm and +60nm, this target is necessary to extract defocus.

6.2.4. Overlay Test Structures

Since overlay test structures have a complicated non-linear response to misalignment as well as defocus and dose due to line end shortening, these structures are better analyzed in groups rather than individually. The primary test structure or cell was described in section 4.4.4 and consists of 196 transistors with a varying amount of programmed poly overlap. Since the leakage current of each transistor is a strong

function of line end pullback, which is a strong function of dose and defocus, it is helpful to look at this structure as a whole one dose and defocus values have been extracted. Figure 60 shows the leakage current plotted vs poly overlap. One will notice the curve is flat for larger amounts of overlap and then starts shooting up as the rounded portion of the top of the transistor starts to overlap active. This current increases exponentially as the top slice dominates leakage current until the transistor gate no longer fully overlaps active, at which point the leakage current increases linearly as a wider and wider section of active is uncovered. Looking at the steep portion of the curve, the blue curve is shifted 20nm from the pink curve, which corresponds to the 20nm difference in misalignment. Hence looking at this shift will indicate the amount of misalignment. This shift calculation is particularly useful as the only curves necessary to calculate it is the curve with no misalignment under the extracted dose/defocus settings and the measured curve. So there is no need to do an exhaustive simulation study for all possible dose, defocus, and misalignment combinations. Test structures are only simulated with misalignment at nominal conditions to filter out test structures that are insensitive to misalignment and hence good dose and defocus monitors.

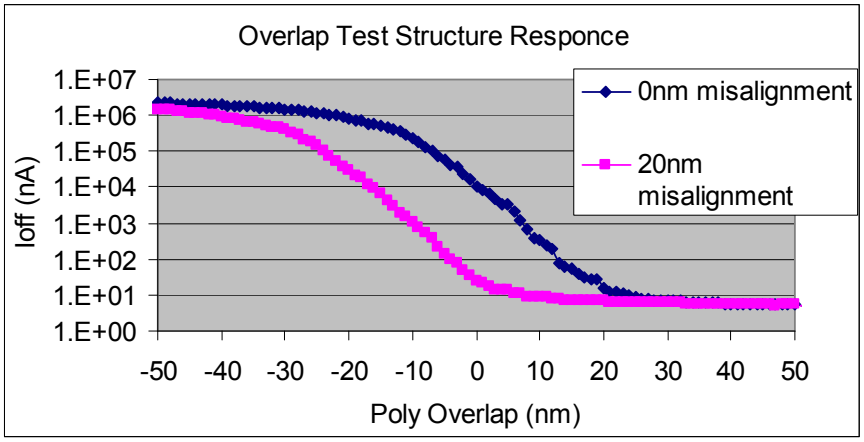


Figure 60 Overlay cell at two different misalignment values. Nominal dose and focus. There are a 196 transistor with varying amount of poly overlap, shown on the x axis. Hence each point corresponds to the leakage current in an individual transistor in the cell.

6.3. Simulation Experiment Analysis and Results

Simulated currents of all the test structures in the testchip at fourteen different process conditions that corresponded to fourteen dies were uploaded into the electrical measurements table. Effective lengths were extracted from currents using a perl script similar to the one that found equivalent rectangular transistors. Half the dies had randomized defocus and dose conditions and half had randomized, dose, defocus, and misalignment conditions. All dies had randomly generated LER and an extra rounding error as all extracted dimensions were rounded to the nearest nanometer. As a sanity check the LER noise strategy was tested by looking at a distribution of 196 identical transistor on one die. Figure 61 shows the distribution with a 3-sigma CD value of 4.8nm, which is pretty bad for 90nm process. The 2003 ITRS specifies a 4.7nm 3-sigma CD control spec at 90nm, but this is for all patterns in an ASIC[53]. Looking at all 80nm transistors in the testchip the 3-sigma value is 17.2nm, but this is due to poor OPC as well

as the added noise. This section will describe the accuracy and ability of the described process extraction technique to extract process conditions from this relatively noisy leakage current data.

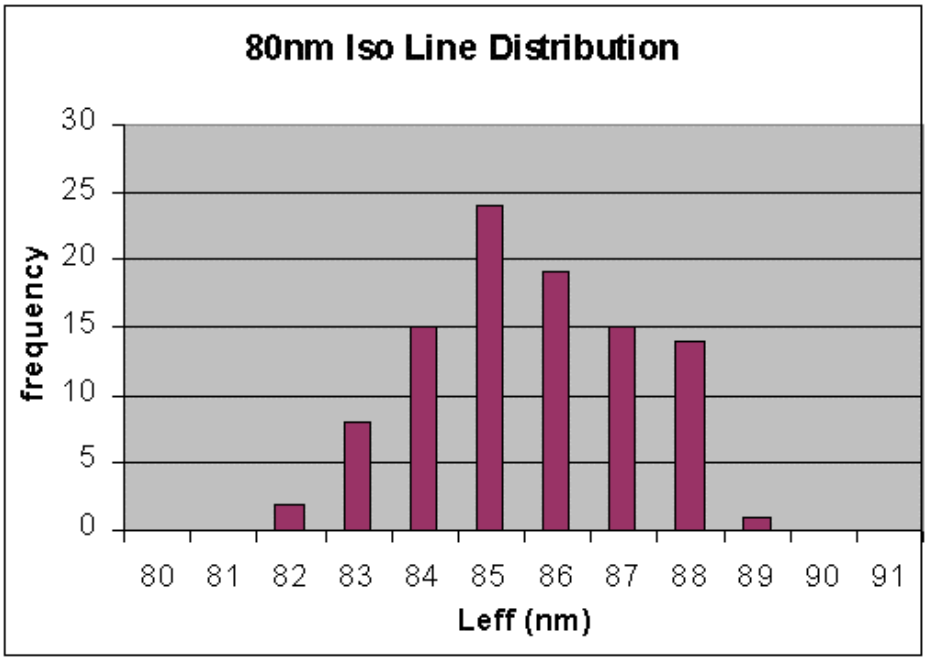


Figure 61 CD distribution of isolated 80nm gate on one of the dies simulated. The random and LER noise lead to a 4.8 m 3-sigma CD distribution.

6.3.1. First Guess of Dose and Defocus

Initially single parameter dose monitors were used to estimate the dose. To increase accuracy the average extracted gate length from a group of identical structures was used to lookup the dose from the modeled dose curves in the simulation part of the database. As the error due to random variations and measurement error decreases by the number of test structures measured, any level of random noise can essentially be eliminated by using enough redundancy. As an added measure to resolve model fitting errors, five different pitches of 214, 224nm, 234nm, 244nm, and 254nm were used to

extract dose. Each pitch had a linear equation fit in excel and dose was calculated from the average electrically measured CD. All test structures where insensitive to misalignment as they had long poly overlaps and did not have poly corner rounding anywhere close to the active region.

The initial dose estimate was then used to extract the amount of defocus from isolated lines and the ring defocus monitors. The primary reason for using isolated lines vs the non-rectangular transistors is that there was only one instance of the hyper-sensitive non-rectangular transistor per die, so no averaging could be done to reduce the random noise. The ring defocus targets were required as they were the only structures that did not have a symmetric through focus behavior and hence had a significant CD change at small amounts of defocus. This proved to be necessary in more than half the dies as the amount of defocus was small. A lookup table approach was used as the defocus targets did not have an easily fit quadratic. Defocus values were estimated by selecting a curve corresponding to the first guess dose and finding the defocus value that corresponded to the average electrical CD.

die #	dose 1st guess	iterated dose	actual dose	1st guess error	iterated error	defocus 1st guess	iterated defocus	actual defocus	1st guess error	iterated error
1	99	99	99	0	0	50	42	42	8	0
2	104	103	103	1	0	0	-9	-7	7	-2
3	103	102	102	1	0	28	15	17	11	-2
4	98	98	98	0	0	-22	-24	-24	2	0
5	96	97	97	-1	0	-3	7	6	-9	1
6	102	105	101	1	4	-49	-28	-58	9	30
7	103	103	103	0	0	25	28	29	-4	-1
8	98	99	99	-1	0	-80	-76	-79	-1	3
9	95	96	96	-1	0	40	34	34	6	0
10	101	102	102	-1	0	25	29	25	0	4
11	98	98	98	0	0	-73	-75	-76	3	1
12	98	98	98	0	0	15	18	15	0	3
13	95	96	96	-1	0	40	43	37	3	6
14	98	99	102	-4	-3	110	89	3	107	86

Figure 62 Extracted dose and defocus values from leakage currents. The 1st guess value is an initial guess from a small set of structures. The iterated value uses the database to iterate to a solution that minimizes error between modeled and measured dimensions.

6.3.2. Iterated Process Extraction

Once a first guess estimate is known, it is inserted into the actual_conditions table and the execution of one query updates the modeled dimensions table. The modeled dimensions table holds modeled CD values given the dose and defocus values in the actual conditions table. These values are interpolated from the simulated dimensions table between adjacent simulation points. The simulated dimensions table has a simulated Leff value for a focus exposure matrix with a 40nm focus step from –120nm to 120nm and a 1% dose step from 96% nominal to 100%nominal dose. Leff can be calculated for any defocus and dose value by looking at the closest four simulated data points and linearly interpolating between them. As Leff does not change by more than several nm across a single step in dose and focus, this method is accurate enough as both the electrical dimensions are rounded to the nearest nanometer.

Extracting process conditions becomes an iterated process that now uses hundreds of transistors to evaluate the accuracy of the dose and defocus estimates. Again dose is estimated first where the average, minimum, maximum, and standard deviation of the error between the modeled_CD and measured CD is looked at. Since misalignment is not yet known and the defocus estimate could be wrong, a set of 127 single parameter dose structures is used for this calculation. The goal is to make the average error zero, which is generally achieved within a couple iterations. Since each step in dose leads to a 1 to 2nm change in average dose error, the dose is easily fit to under 1% accuracy. Once a better estimate of dose is known, 731 defocus sensitive test structures are used to estimate defocus. Again the goal is to minimize the error between modeled dimensions and measured dimensions.

When looking at defocus it is important to also look at the stdev of and range of errors. Sometimes it is necessary to filter out outliers if certain test structures have a particularly large error. This is generally the case when the linear approximation breaks down, which may happen, especially in threshold spike monitors. Outliers need to be eliminated so that they don't skew the average. Once the error is minimized the defocus can be estimated with nanometer accuracy. As seen in Figure 62 the defocus can be extracted to within a few nanometers in most cases. Sometimes it is necessary to look at different subsets of test structures such as isolated lines, defocus monitors, or threshold spike monitors to get a more confident estimate. It was found that at low defocus values looking at all the test structures resulted in a very accurate estimate, but at high defocus values looking at just isolated lines changed the answer by several nanometers and resulted in a more accurate estimate. At high defocus values the isolate line has a very

predictable behavior so is more accurate, but at low defocus values the ring defocus monitors as well as other focus sensitive structures is necessary.

Dose can also be estimated or checked by looking at very dose sensitive structures that are no longer insensitive to defocus. This is ok as the defocus value is known. One may also notice that the accuracy decreased with misalignment. This is due to the fact that a lot of the defocus test structures have some sensitivity to misalignment (especially the ring defocus targets). This can be fixed by designing the defocus targets to be insensitive to misalignment and using that subset if the range of errors becomes large.

Once dose and defocus estimates are known and improved, the overlay test structure is used to evaluate misalignment. The iterative approach is not used for finding misalignment as simulating all test structures across a full FEM at different misalignment values would explode the 1.5 million entries in the simulated dimensions table into at least 15 millions. Hence the strategy uses the best defocus and dose estimate to create a plot leakage current vs poly overlap for both the modeled dimensions and measured dimensions. The misalignment is then extracted by looking at the shift in both curves. This can be done in excel by shifting the values in the measured currents column until the curves overlap. Figure 63 shows the accuracy of this method to be several nanometers. It is important to note that only one overlay test structure was used and hence random CD variation played a big role in the accuracy. If multiple overlay test structures were designed into the testchip it would be possible to average the electrical CD at each poly overlap value and this misalignment strategy would be a lot more accurate. Accuracy should also be improved when defocus estimates can be made more accurately.

extracted misalignment	actual misalignment	error
-16	-19	3
0	-8	8
-13	-15	2
15	21	-6
2	3	-1
-38	-43	5
15	16	-1

Figure 63 Extracted misalignment vs actual misalignment. Misalignment was extracted from one structure so CD errors could not be averaged out.

6.3.3. Quality of Fit

The range of errors and stdev proved to be a good indicator for goodness or quality of fit. All dies had a stdev of ~1.6nm except dies 6 and 14. Dies 6 and 14 are obvious outliers with very large errors in extracted process values and a high standard deviation. Different sets of test structure like defocus monitors vs isolated line showed significantly different results. The large error in process condition estimate was in a sense anticipated as the range in errors for die 6 and 14 were twice as large as the others. The expected reason for this was a bug in the code that resulted in overwriting of results when running the Parametric Yield Simulator in parallel. The Parametric Yield Simulator has several dozen files that need to be separated into separate folders between multiple parallel runs. Unfortunately one such file overlapped between two simultaneous simulations and given the right timing the wrong process conditions could be simulated for a portion of the gates. This is expected to have happened for these two dies. Mainly because using the actual conditions in the actual_conditions table still lead to significant errors between modeled and measured dimensions. The data for the dies was included to indicate that

unexpected rare non-idealities can potentially be filtered out. This can be the case if there is a big particle in the resist or defect density is high on one die. The range and stdev of error can be used as a litmus test for such gross non-idealities and as an estimate of goodness of fit.

6.3.4. Lessons Learned

Using the large set of test structures and the database for process extraction has shown to be very accurate, but the ultimate accuracy has proven to be somewhat a function of art and science. This is not too surprising given the shotgun approach and the intricacies of sub-wavelength lithography.

Firstly some hyper-sensitive non-rectangular transistors were found by simulating the entire layout and pushing the statistical capabilities of the MySQL database. These patterns are probably only sensitive with this particular version of OPC and illumination, but just happen to be twice as sensitive as isolated lines. This structure underlines the complex optimization problem of making a sub-wavelength feature printable as well as stable through dose and focus. In fact this could be an orthogonal optimization problem as adding edge biases in OPC that minimize edge placement error at nominal conditions may degrade through focus behavior. There are many ways to apply OPC, but given infinite layout possibilities it is very hard to resolve this problem for all instances. Simulating the entire layout at various process conditions may be computationally prohibitive given an iterative process such as OPC. Hence a shotgun approach with a wide variety of test structures can be used to evaluate the robustness of an OPC algorithm. Potential canaries can be identified and flagged in production chips and more sensitive test structures can be designed.

Another result of this study is that process characterization is a mix of art and science where experience can make a big difference in accuracy and convergence time. It is important to analyze the test structures themselves and make sure the sets of test structures used for process characterization are well balanced. This means they have the right combination of test structures so that their linewidth can be used to uniquely identify the signature of defocus or dose, ie. defocus could not be mistaken for dose or another process parameter. Figure 64 shows this where one set of test structures leads to a less accurate result. The original set of 440 structures was used to characterize die #4, but an expanded set was used later that lead to the correct answer. A small 2nm discrepancy was due to the fact that some defocus sensitive structures did not have enough redundancy and the random error caused a slight bias. Adding more structures resolved this issue.

It is important to look at the standard deviation of the errors or the residuals to evaluate goodness of fit. Notice in Figure 64 the stdev decreases as the average converges to zero, but in Figure 65 the stdev does not decrease to levels that are expected from the generated level of random noise. The stdev in Figure 64 corresponds to a 3-sigma value of 4.2nm, a little lower than the isolated line in Figure 61, but that is expected as some of the structures are dense pitches that have better ILS and hence less random noise. Figure 65 is from die #6 where the actual conditions where guessed wrong. This was due to the fact that a portion of the transistors in die #6 were simulated with different and wrong process conditions. A lesson learned in this dry-lab experiment is that the accuracy and confidence of extracted conditions will improve with experience. Future experiments will

hopefully refine this strategy and create a more scientific approach to maximizing accuracy.

	with 440 structures					with 732 structures	
guess #	defocus (nm)	avg	stdev		defocus (nm)	avg	stdev
0	28	-1.01	2.26		28	-0.48	2.05
1	24	-0.71	1.84		24	-0.3	1.74
2	20	-0.42	1.5		20	-0.12	1.5
3	16	-0.09	1.38		18	0.06	1.38
4	14	0.05	1.39		17	0.01	1.4
5	15	-0.02	1.38				

(a) (b)

Figure 64 Iterative guesses for defocus for die #4. (a) uses 440 test structures to estimate error and (b) uses an expanded set of 732 structures. Set (b) arrives at the correct defocus value of 17nm.

guess #	defocus	avg(error)	stdev(error)
0	-50	2.2	2.65
1	-35	0.5	2.44
2	-30	0.14	2.4
3	-28	-0.018	2.44

Figure 65 Iteration from die#6. Notice how the stdev(error) does not decrease as the average is minimized. This indicates something is wrong and a portion of the gates have some type of bias on them.

6.4. Discussion

Using automated electrical measurements from a large set of structures and an iterative approach inside the database has demonstrated sub 10-nanometer accuracy. It has been shown that 2-dimensional patterns can have higher defocus sensitivity than isolated lines and OPC can be even more counter-productive in stabilizing images through focus. Using transistors as measurement structures enables very high packing density, which can be used to reduce noise by redundancy. Electrical testing is the only inexpensive testing solution for a wide variety of 2D test structures, which is required for

high accuracy and comprehensive process extraction. This process requires structures that have a strong, moderate, and weak responses to each process parameter as then the signature of each process parameter can be uniquely encoded in the linewidth responses of all the structures.

The beauty of this approach is that lithography is the only process that has a die specific signature, so this approach should have similar accuracy in silicon. If enough test structures are used, defocus is modeled correctly, and the defocus condition is accurately extracted, the distribution of (measurements – modeled_dimensions) should simply be a Gaussian with a bias or systematic shift. The systematic shift can be subtracted out in the database and the sigma can be reduced by redundant data. Since focus, exposure, and misalignment are the only variables systematically changed from die to die, their impact on transistor performance will be captured in programmed exposure experiments. If other processes have pattern dependent effects, such as strain, then they will be captured in the programmed experiments model as the pattern remains essentially the same from die to die even if defocus or dose is changed systematically. If there happens to be a bias from another effect that only affects one die, the standard deviation of the errors can be used to evaluate the goodness of fit. If all structures are modeled correctly and the process conditions are extracted accurately, then the residual plot should essentially be a measure of random noise and should have a similar stdev to a histogram of many instances of the same feature (ie. The isolated line in Figure 61). So even though this simulation study did not capture all sources of process variations, this procedure should work in an inherently noisy silicon experiment.

As far as modeling lithography, dose, defocus, and misalignment are the only parameters that need to be characterized to represent the process window in lithography DFM tools. This is because those are the only parameters that will vary independently as they depend on wafer uniformity, laser consistency, and stage accuracy. Even though mask edge effects, lens aberrations, and polarization can play a significant role in pattern dependent effects and will change the response to defocus, there is no need to characterize them in a DFM tool itself. These affects need to be captured in the optical model, which is empirically fit to capture all optical non-idealities. It is the responsibility of the fab to characterize their optical systems and release optical models that generate contours similar to what is seen in silicon. Once these optical parameters are captured, they themselves will not change significantly from die to die. They will effect how defocus will change an isolated or a dense line, but that should be captured in the defocus optical model. So given a DFM tool and optical model, the last component is confidently establishing a range of expected defocus, misalignment, and dose values. If a DFM tool does not model the lithography step accurately, this will be evident in the distribution of errors between simulation and experiment results in programmed experiments. In which case the guard bands on each transistor will need to be adjusted appropriately or the optical model will need to be fixed. So in a sense, this type of characterization study can also be utilized for evaluating the DFM process model fit. Either way, the ultimate the goal of this characterization study is to figure out how much variation can be confidently captured with the DFM tool and how much can transistor guard bands be decreased.

Finally, when extending this strategy to look at other process effects, a strategy similar to extracting misalignment is recommended. Once misalignment, dose, and defocus are

extracted, etch and other sources of systematic variation in the pattern transfer process can be addressed. The key assumption, which should be accurate to at least first order, is that the subsequent pattern transfer process steps are independent of litho. If this is the case, it is possible to subtract the signature of litho in the database and then analyze the data for other sources of variation. It has been has shown that spatial correlation can be eliminated with better models of the systematic variation[54]. So the data in the database can slowly be massaged with models that eventually account for all variation and the remaining residuals across all patterns are normal and a metric of pure random noise. Note that the residual distribution in the defocus extraction is the difference between the modeled dimension and electrical dimension, the residuals do not include the systematic bias that may exist between the two models of two features with different proximities. Those differences will be the non-litho pattern dependent effects.

6.5. Conclusion

This chapter described the implementation of a simulation based dry-lab experiment using the Parametric Yield Simulator and the Collaborative Database. The goal was to extract process conditions based on simulated leakage currents in test structures on the Enhanced NMOS FLCC testchip. The process extraction strategy leveraged a high volume iterative approach inside the database and lead to sub 10nm resolution in misalignment and defocus and sub 1% resolution in dose extraction under significant levels of noise. This extraction strategy should still work with comparable accuracy, even with higher levels of noise in silicon experiments. Large numbers of test structures and redundancy have shown to significantly reduce noise and the signature of dose and defocus can be captured in programmed lithography experiments or simulation runs. Also

the distribution of residuals between modeled dimensions and measured dimensions is a good sign of goodness of fit, so each extracted defocus, dose, and misalignment value can be quantifiably evaluated for accuracy.

The key differentiator for this work is the large number of process specific test structures used in process extraction. Most other approaches use either one or two types of structures with limited averaging as measurement cost is high or electrical testchips that do generate lots of data, but do not have enough process specific structures to generate a broad range of responses to defocus and dose. More specifically, structures such as those using a 90 degree phase shifting probe for an asymmetric response through focus, which is necessary for high accuracy at low defocus values, have not been implemented in a transistor based testchip to the best of our knowledge. In process control, which has slightly different goals than process characterization for DFM tools, phase shifting masks and inline monitors are commonly used along with scatterometry, which is a lot cheaper then CD-SEM metrology. These techniques are optimized for inline process control and not process window characterization, hence their accuracy is in the 30nm range[55,56]. An electrical "shotgun" approach can not only result in high accuracy, but also the ability to measure many different process parameters.

Looking towards the future, this process could be improved by strategically using a smaller set of test structures with more redundancy that are specifically aimed at the processes that have corresponding DFM tools. It should be possible to create a set of scribe line test structures that can effectively be used to extract dose, defocus, and misalignment from electrical measurements. The first guess and iterative procedure used to extract dose could also be automated. With automation it would be possible to process

hundreds of dies with little manual labor. As the testing is electrical, data aggregation costs can also be minimized by using intelligent self-testing circuits. As more experiments are executed with these strategies the accuracy, silicon area, and testing efficiency can be optimized. Chapter 8 will describe improvements in test structure design that can help make this happen.

7 Programmable Defocus Monitors

The transistor defocus monitor with ring and 90 degree phase shifted probe (section 4.4.5) has been modified to work as a single layer open/short electrical monitor and provide a vehicle for comparing simulation with experiment inside the database. This short loop strategy was created as an alternative when the Enhanced NMOS wafers could not be completed due to spin-off of SVTC from Cypress. This chapter first covers the simulation and design of the defocus targets and then the experimental characterization through focus.

The goal of the new defocus monitors is to create a set of single layer circuits which will either be open (high resistance) or short (low resistance) depending on the amount of defocus. The benefits of this target are that it can be tested in a short loop single layer experiment and can be implemented on layers other than the poly layer, as was the limitation in the transistor version. The primary drawback of this approach is that each monitor essentially has a digital signal vs the transistor version that has a grayscale of measurable linewidth. As shown in Chapter 6, the phase shifting probe leads to an asymmetric response through focus and high sensitivity of linewidth to defocus, which proved essential for extracting defocus accurately. In order to reproduce this sensitivity with a digital monitor, different targets are programmed to pinch open at different levels of defocus. So when a large sample of different targets are looked at, defocus can be

determined with high accuracy. These monitors can be programmed by varying four layout parameters; line CD, number of rings, phase shifting probe size, and position of the target on the line (offset from center) (Figure 66). Since these are single layer structures, electrical testing can be done immediately after etch, so they could be used for process monitoring as well as process characterization. In either case, the main part in this strategy is the large numbers of different structures analyzed together, which is enabled by electrical testing and the database for data analysis.

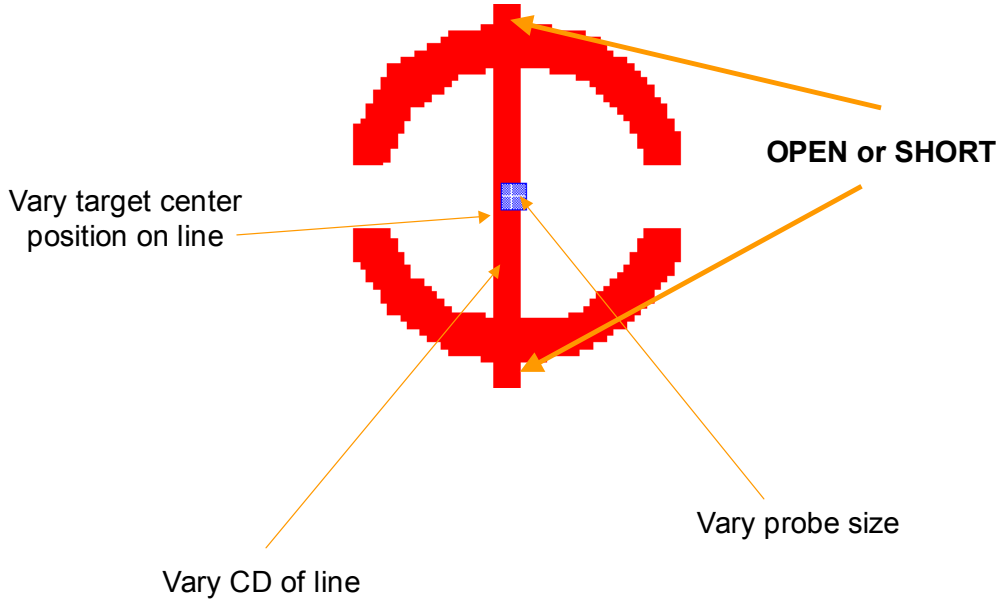
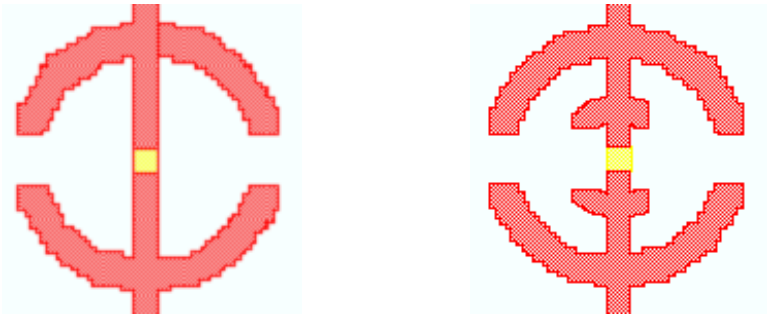


Figure 66 Programmed defocus monitor. By changing four parameters, line CD, (phase shifting) probe size, offset between ring center and line center and the number of rings, it is possible to program the defocus target to pinch open at different levels of defocus.

7.1. Test Structure Design

Two types of the programmable defocus test structure were designed inside standard 30-pad cells; a single layer version with 25 targets and a two layer version with 225 targets. Five versions of the two layer cell type contain 454 unique combinations of the

four main parameters and two versions of the one layer cell_type contains 50 unique combinations of the four main parameters; number of rings, CD, probe size, and offset. Half the targets have two rings and the other half have a single outside ring (Figure 67). The inside ring in the two ring target increases the amount of spill over through focus, but its proximity to the center makes it harder pinch open, so there is a tradeoff.



(a) (b)

Figure 67 (a) single ring defocus monitor. (b) two ring defocus monitor.

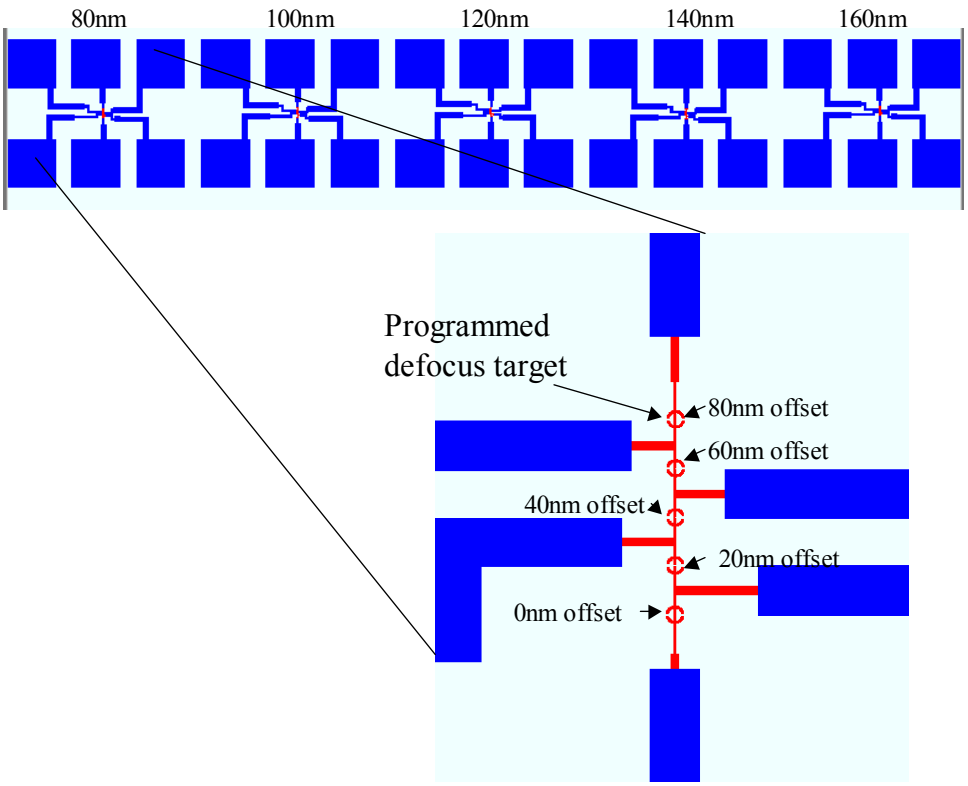


Figure 68 Single layer defocus cell. This 30-pad cell can accommodate 25 individually probable defocus monitors. There are five sets of monitors with five different CDs 80nm, 100nm, 120nm, 140nm, and 160nm. Each set of five targets has five offsets; 0nm, 20nm, 40nm, 60nm, and 80nm.

With routing signals being the main constraint in single layer electrical test structures, only 25 defocus monitors were fit into a 30 pad cell. Each test defocus monitor has one unique combination of two pads to address it (Figure 68). Each cell is designed with five main CDs, 80nm, 100nm, 120nm, 140nm, and 160nm with five alignment off-sets between ring and line that are either 0nm, 20nm, 40nm, 60nm, or 80nm. Two versions of the cells use the two ring target and two versions have the single ring version. The strategy is that the line CD will have a large influence on defocus sensitivity and probe offset will serve as a finer tuning for higher accuracy.

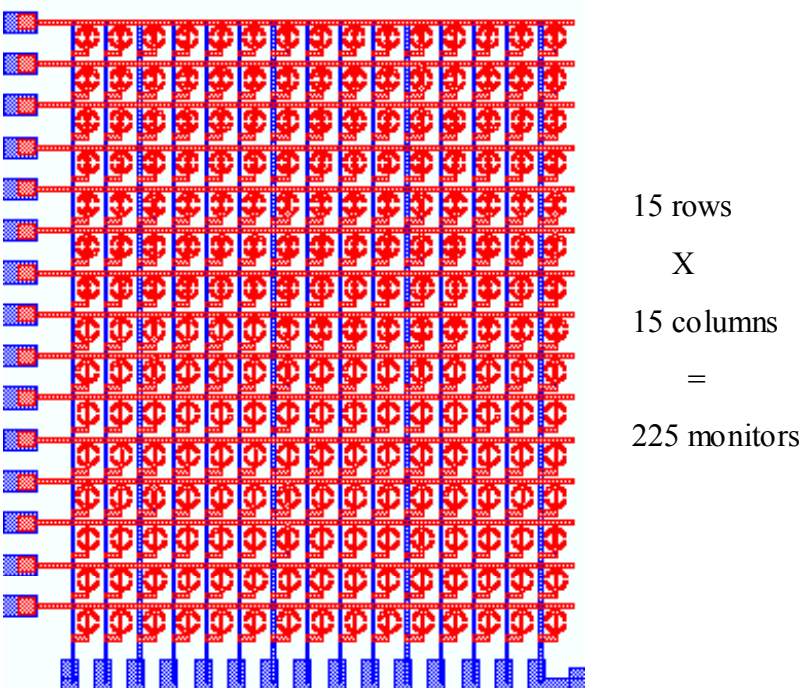


Figure 69 Two layer 225 monitor defocus cell. Each defocus target can be uniquely addressed by a combination of two probe pads. A short only exists if the defocus monitor is not pinched open.

The two layer version in Figure 68 can route a full 15x15 matrix of 225 individually probable defocus monitors. A script was written that automatically generates this matrix from an input file that specifies the line CD, probe size, offset, and number of rings for each target. The vertical spacing was strategically chosen so that the horizontal poly line has significant overlap with what would be the third ring of each target. The matrix seen in Figure 69 has vertical metal lines that potentially short out to horizontal poly lines only through individual defocus monitors. So depending on if the defocus monitor is pinched open, the corresponding poly line and metal line will be open. This Matrix measure 25umx30um, so is small enough to fit as a scribe line structure. The benefit of having a large number of monitors is an increased accuracy from either having finer differences between defocus monitors or from increased redundancy. The two layer version has been

simulated in the PYS with 454 unique combinations of the four layout parameters and varying amounts of redundancy.

7.2. Simulation Results

In order to evaluate the different defocus targets and sensitivity to the four different parameters (number of rings, CD, offset, and probe size) the new layouts were simulated in the Parametric Yield Simulator and results were uploaded to the database. The minimum simulated CD in the center of the target was used to evaluate if there was an open or a short. Calibre has a feature that outputs negative simulated CD values that quantify the severity of a pinch. The more negative the number the bigger the pinch. An example of this can be seen in Figure 70 where the center portion of a 2-ring defocus target is simulated through focus at different doses. This is the plot of minimum CD, which is positive at more positive defocus values and negative, or a pinch exists, at more negative defocus values. An easy way to quantify the breaking point is finding the point of intersection where the min_CD is zero. The pinch-point can be interpolated from simulated points on a 40nm defocus grid, so extensive simulation is not needed. Since the original PYS and database accounted for the minimum CD, no extensions needed to be written to analyze pinches and opens. The threshold value is in some ways arbitrary as resist models may not be accurate at small dimensions[57]. What is most important is that the threshold is consistent across the different targets.

In total 454 unique combinations of the four parameters were simulated at different dose and defocus values. Two illumination systems were used. Quasar with sigma in/out 0.72/0.92 and a 40 degree illumination angle and tophat with a sigma of 0.34. Both

optical models had a 193nm lamda and 0.85NA (Rayleigh Unit = 134nm). The same poly resist model was used from Chapter 6.

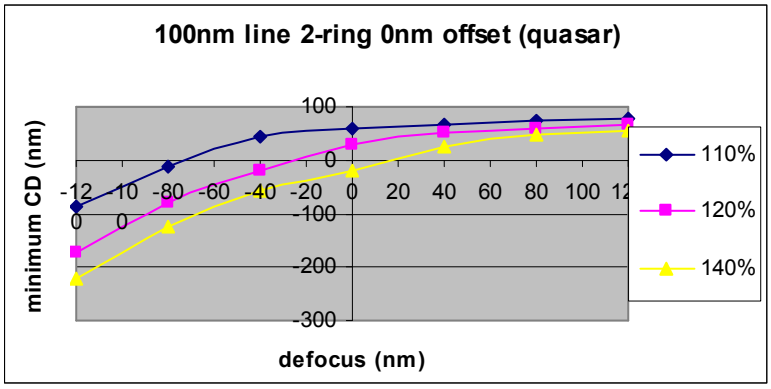


Figure 70 2-ring defocus target with 100nm CD, 0nm offset and a 100nm probe. The plot shows the minimum simulated CD vs focus for three different dose levels. 100% is the default level of 1 inside CalibreWB, which does not necessarily correspond to actual nominal dose.

7.2.1. Sensitivity

The primary strategy for electrically extracting defocus is to determine which defocus targets are open and then figure out what defocus value(s) could lead to such results. As different defocus targets have different layout parameters, and hence pinch open at different defocus values, different sets of defocus monitors will be open at each level of defocus. One way to evaluate sensitivity and range is to plot a number of defocus targets pinching open vs focus. Figure 71 and Figure 72 show this plot for quasar and tophat illuminations respectively. There are 454 unique defocus targets, of which a portion opens up at various defocus conditions. Note that increasing the dose in effect translates the range over which the targets pinch open. Tophat illumination also has an opposite trend with focus than quasar. Negative defocus pinches targets out with quasar

illumination and positive defocus pinches out targets with positive defocus. The reason for this is the off-axis illumination in quasar leads to a phase carpet across the mask and essentially flips the sign of the light spilled over from the surrounding features. There is one or two defocus targets pinching open for each nanometer of defocus. By looking at which targets are pinched, it should be possible to pinpoint defocus within a nanometer, this is assuming there is no noise.

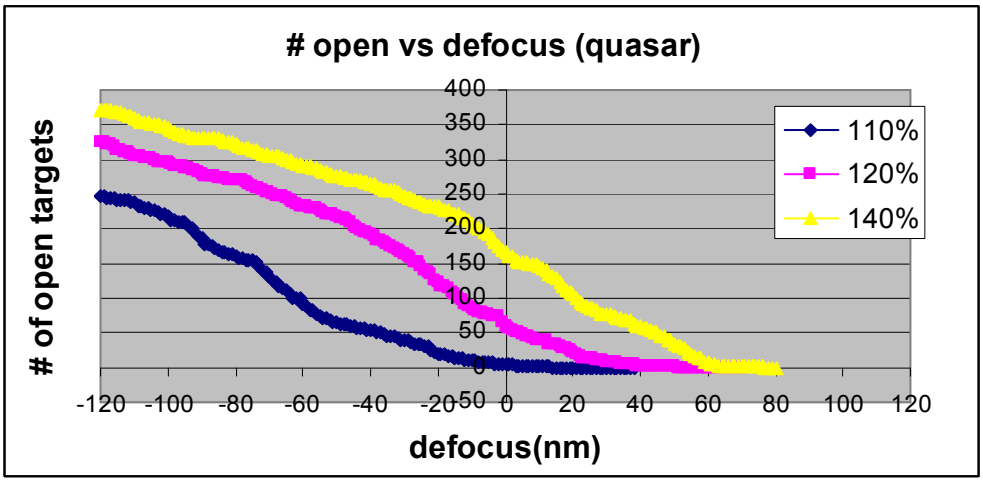


Figure 71 Number of defocus target open (out of 454) vs defocus for quasar illumination. Each point represents a defocus target, so there is almost 1 new defocus target pinching at each nanometer of defocus. The different lines represent 110%, 120%, and 140% dose.

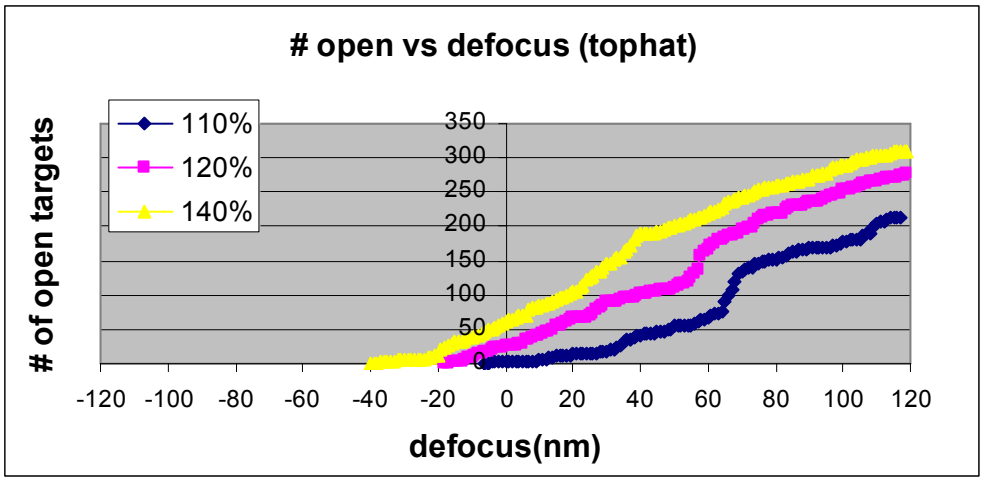


Figure 72 Number of defocus target open (our of 454) vs defocus for tophat illumination. Each point represents a defocus target, so there is almost 1 new defocus target pinching at each nanometer of defocus. The different lines represent 110%, 120%, and 140% dose.

Unfortunately there is a lot of sources of noise that could decrease accuracy. Dose non-uniformity can have a significant shift on the targets, but this can be accounted for by adding ELM structures that can characterize dose variation. Random mask errors, LER, and random development phenomena will be harder to account for. The main repercussion of random variations is that the target will no longer serve as a sharp threshold monitor. The only way to mitigate this is with redundancy. Then there will be a range of defocus values, most likely centered around the true threshold for that monitor, where the redundant monitors will pinch open. In fact, looking at the range over which the first and last monitor pinches open might be an indication of the level of noise. The slope of that curve can also be used to determine confidence intervals for defocus. Fortunately looking at a large number of targets is easy as all measurements are electrical.

Figure 71 and Figure 72 showed how defocus monitors can be programmed to have sub-nm differences in sensitivity. To better understand how this sensitivity changes, it is useful to look at each parameter separately to understand its role and also have more physical insight when analyzing the data. The next four sections will detail how the sensitivity of the defocus monitor will change if only one parameter is varied.

7.2.2. Sensitivity to CD

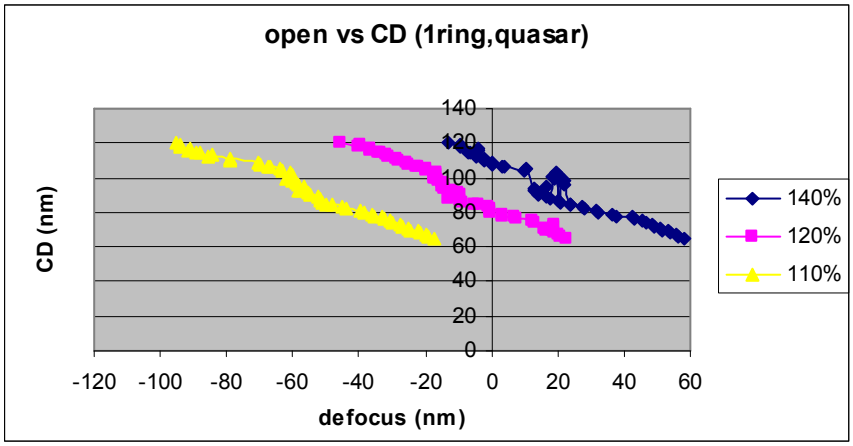


Figure 73 Sensitivity to CD or linewidth of the center line to defocus. The thicker the line the more defocus is required to pinch it open. The offset is held constant at 0nm and probe size is held constant at 100nm.

Center line CD or linewidth plays a significant role in when a defocus target pinches open. Intuitively a thicker line is going to require more defocus or more light spillover to generate a pinch. Figure 73 demonstrates this relationship with a set of targets where probe size, offset, and the number of rings is held constant and CD is varied from 60nm-120nm. The three curves, corresponding to the three doses, have a similar slope where a 1nm change in linewidth translates to a 1.1 nm change in defocus pinch-point. Using a

higher dose effectively biases the CD of the line and essentially pushes each curve towards lower defocus values. As a note on the dose percentages, nominal conditions in Calibre may be completely different from nominal conditions in reality. A dose of 1.00 in the input file is considered nominal dose for simplicity and consistency, but in reality nominal dose is chosen based on a target size of a target feature. It will be important to either program targets to operate at the doses specified in production or to span a broad range of doses.

7.2.3. Sensitivity to Probe Size

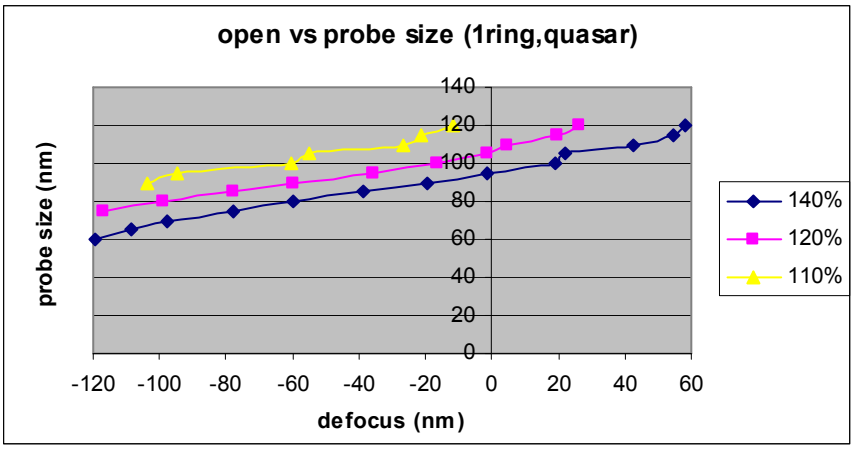


Figure 74 Sensitivity to probe size. The smaller the probe the more defocus is necessary to cause a pinch. This is for a 100nm line with 0nm offset.

The phase shifting probe size plays an important role in the defocus threshold as larger probes let through more light and create a larger breaks in the line. Thus a larger probe size requires less defocus to pinch the target open. Figure 74 shows how sensitive the defocus threshold is to probe size for a 100nm line with a 0nm offset. For the high dose of 140%, changing the probe size from 60nm to 120nm spans the entire range from –120nm to 60nm. A 1nm change in probe size translates to about a 3.3nm change in

defocus threshold. Hence the probe size is a very effective lever in spanning a broad range of focus conditions.

7.2.4. Sensitivity to Offset

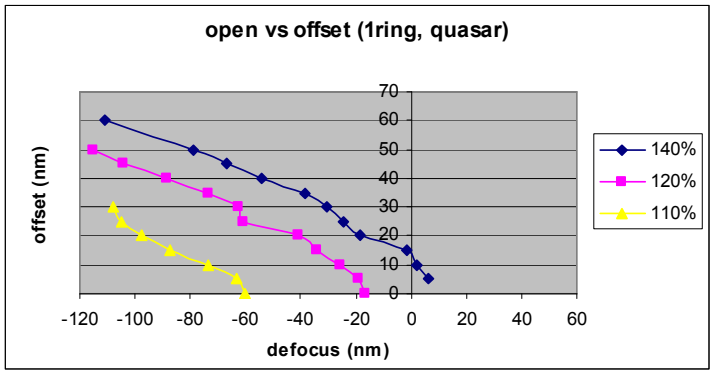


Figure 75 Sensitivity to probe offset. The further the probe is from the center of the line the more defocus is required to pinch it open. On average for every 1nm of offset it takes 2nm defocus. This is for a 100nm line with a 100nm probe size.

An additional layout parameter that can be monotonically changed to adjust the defocus threshold is the offset between the center of the defocus monitor and the center of the line. Naturally the closer the probe is to the center, the less defocus is needed to pinch open the line. This is a similar effect to probe size as higher offsets leave more and more chrome in the line to help it print. As shown in Figure 75 there is a 2nm change in defocus threshold for every 1nm change in offset.

7.2.5. 1-ring vs 2-ring

The number of rings can also have an impact on defocus target sensitivity and pinch threshold. The 2-ring version is expected to have a higher sensitivity as the inner ring helps spill over more light with defocus. The drawback of this target is that the inner ring is so close to the center that the proximity effect of the inner ring generally makes the

printed line thicker. It may also affect etch performance as it is harder for radicals to get into the center as there is less space. Looking at Figure 76 it is clear that the 2-ring target has a sharper drop-off for negative defocus levels, which indicates the pinch is growing faster. Changing the number of rings also slightly influences the pinching defocus threshold and can also be used for fine tuning.

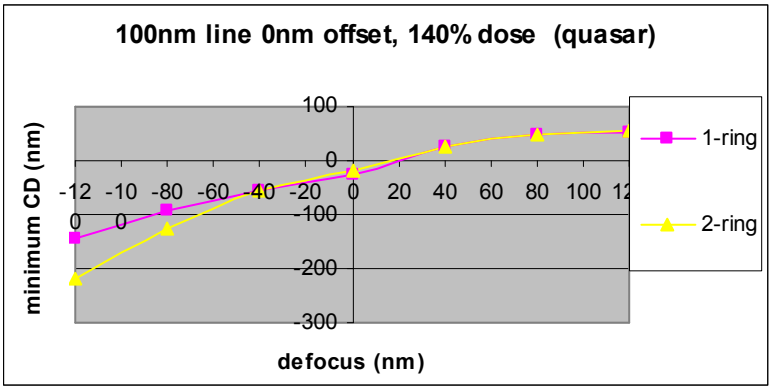


Figure 76 1-ring and a 2-ring defocus target with 100nm CD, 0nm offset and a 100nm probe. The 2-ring target has a stronger signal as it goes further below 0nm than the 1-ring version.

7.3. Experiment Setup

A set of single layer wafers have been manufactured at Silicon Valley Technology Center under ASML sponsorship to experimentally test the single layer defocus targets. A five step process was designed with a single litho and single etch step that lead to 500A thick W/Wti/TiN layer on top of oxide. Photoresist was used as the etch mask to avoid requiring a nitride strip on top of the 100um probe pads. An ASML twinscan1250 193nm, 0.85NA, scanner was used for litho with tophat illumination with sigma 0.34. The poly layer resist was used with a 95nm BARC and 220nm resist thickness. CD-SEM results were taken on the Vera SEM at SVTC and electrical measurements were performed on the electroglass autoprober in the UC Berkeley microlab.

7.4. Experiment Results

CD-SEM measurements have been made for a FEM wafer exposed with tophat illumination. Measurements show promising results with various defocus targets pinching open at different levels of defocus. Figure 77 shows a CDSEM of 30 targets that are in one of the defocus matrix cells. One can see how some targets are fully open, some are partially open, and some print just fine. The strategy is to extract defocus by examining the statistics of open targets. The targets in Figure 77 are split into 10 sets of three identical targets. The left three and right three targets in each row are the same, but targets change from row to row. This snapshot shows differences between different sets as well as reproducibility. The resist images are fairly consistent, but do show some LER effects, especially for the partially open sets. Since these are electrical targets, measuring the full matrix of 225 targets should be fast, inexpensive, and digital, so measurement noise should not be a major issue.

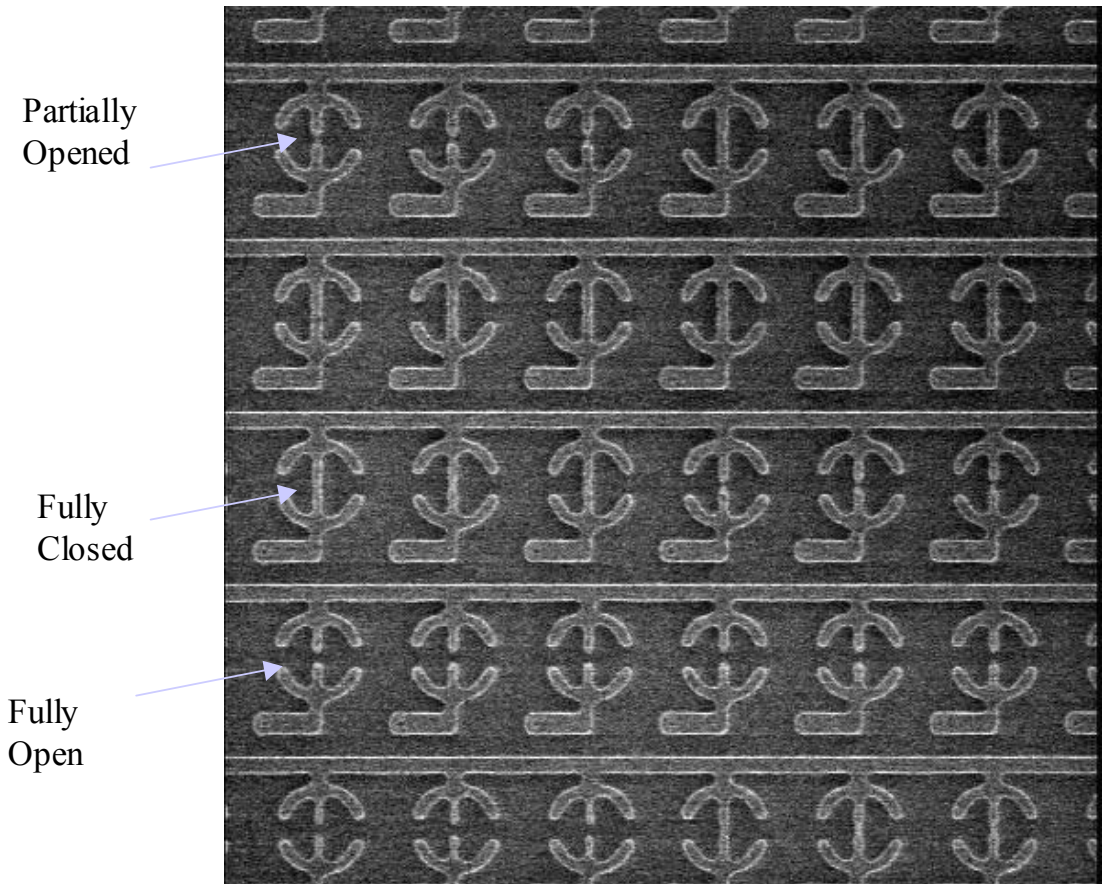


Figure 77 CDSEM of defocus target matrix at 40nm defocus with tophat illumination. Notice how different targets, which have different layout parameters, are pinched open.

When looking at a single target through focus, the center probe area clearly pinches open. Figure 78 shows a single ring defocus target at four different focus conditions. The four pictures spanning positive and negative defocus show an asymmetric response through focus, which is critical for attaining high sensitivity at low defocus values. Even though the center is not completely pinched in resist at 40nm defocus, it should pinch open during etching. Figure 79 shows a two ring target, this time the center pinches open with 80nm defocus. These two targets alone can identify defocus to fall either below 40nm, between 40nm and 80nm, or above 80nm. With a much larger set of structure it is possible to split up a full range of defocus values into much finer granularity.

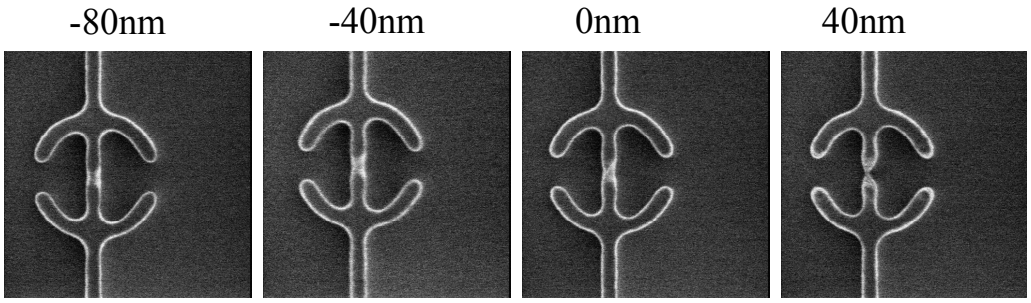


Figure 78 Defocus target at three different focus conditions. This target has a CD of 100nm, probe size of 100nm, and offset of 20nm. It was exposed at 31mJ/cm^2

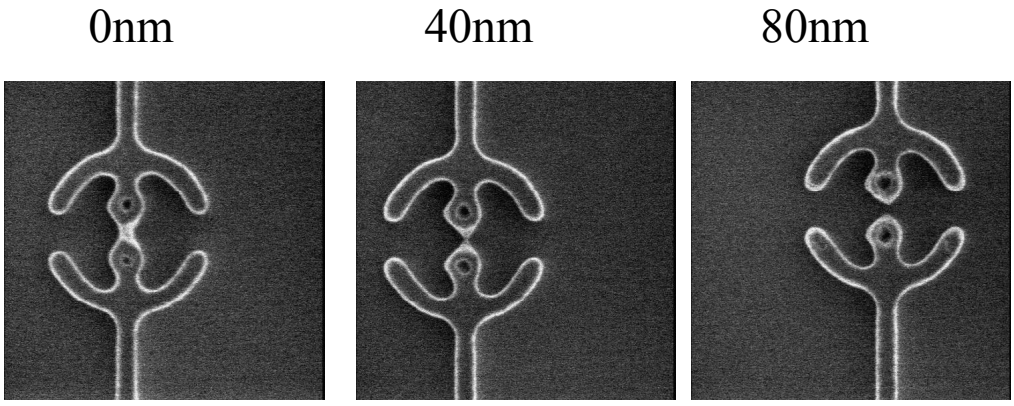


Figure 79 2-ring target at three different focus conditions. This target has a CD of 120nm, probe size of 100nm and a 0nm offset.

7.5. Electrical Results

A total of nine wafers were processed in the short-loop flow and three of the Focus Exposure Matrix wafers were probed. Original measurements applying 1V across a defocus monitor were unusable due to high and inconsistent contact resistance. High contact resistance could be attributed to a native oxide layer as well as the inability to scratch the hard tungsten probe pads. It was found that applying 2V helped break down the native oxide layer and resulted in more consistent results. There was also a significant

etch bias that essentially etched away all 80nm lines and biased all the results from images found in resist. The combination of using a higher measurement voltage and using 9 instances of the same cell in each die, it was possible to get a good signal to noise ratio. Looking at larger number of similar targets also helped paint a clear picture.

To help analyze the data a threshold model was used to translate resistances into 1s and 0s. Looking at the distribution of resistances for all targets, 5 megaohms was used as the threshold value for establishing an open or a short. Figure 80 shows a plot of average resistance vs defocus for 9 targets, where both extremes are far above or far below the 5 megaohm threshold. Figure 81 shows a plot of the same defocus target after the threshold model is applied. It demonstrates that looking at average resistance or even resistances of individual targets can be misleading. Looking at a single target there may be a 60nm uncertainty in extracting defocus, but looking at 9 the picture is much clearer.

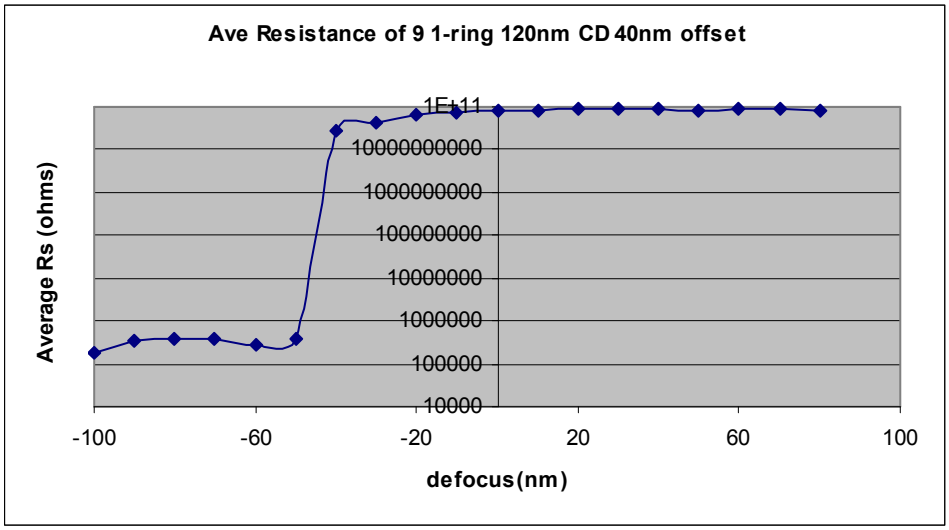


Figure 80 Average resistance of 9 defocus targets. Since an open resistance is much higher than when shorted, the first couple points in the graph actually only have a couple of the defocus targets opened up, which dominate the average.

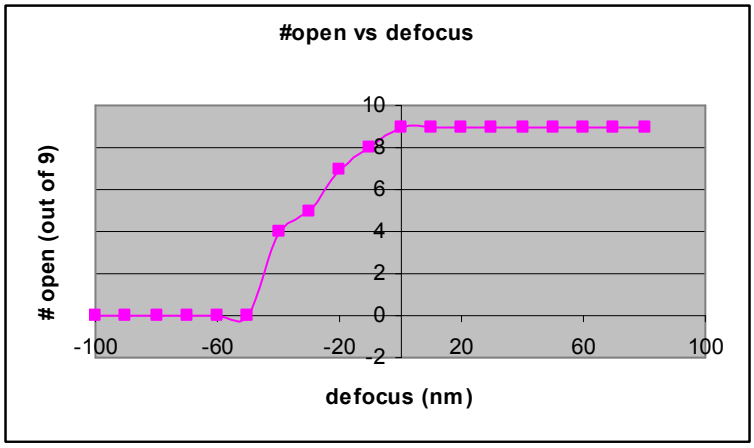


Figure 81 Same target as in Figure 80, but this time plotting number of targets pinched open vs defocus. This plot demonstrates a lot more noise where a set of identical targets takes 60nm of defocus for all of them to clear.

Looking at the number of targets pinching open vs defocus can give an idea of the range of sensitivity and accuracy. Figure 82 shows the number of open targets vs defocus for 600 targets, which consist of 12 instances of the 2-ring and 12 instances of the 1-ring 30 pad cells. The graph shows excellent range in sensitivity and an almost linear response with about 20 targets pinching open every 10nm of defocus. Note that this graph is smooth and not a step function. This is due to the programmed variety of defocus monitors that generate a continuous sequence of opens vs defocus and the random noise that acts as a smoothing function.

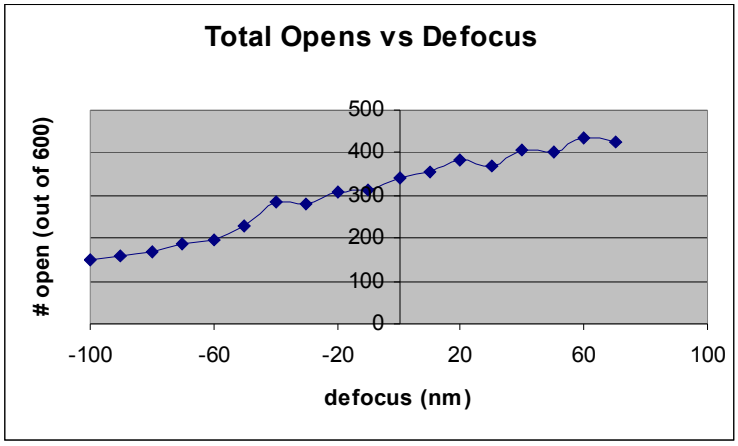


Figure 82 Total number of opens vs defocus for tophat illumination at a dose of 25mJ/cm^2 with a 10nm focus step. In the steep portion of the curve on the right, for every 10nm defocus 20 targets pinch open.

Looking at the data in a little more detail it is possible to evaluate the impact of linewidth and probe offset. Figure 83 shows a plot of all targets with a 40nm offset. One can see the curve is shifted to the right with a higher linewidth. In this case the 80nm line is always open and the 160nm line is almost always shorted. Eyeballing the graph the sensitivity to CD is not constant, but falls somewhere between 1-2nm defocus sensitivity for a 1nm linewidth change.

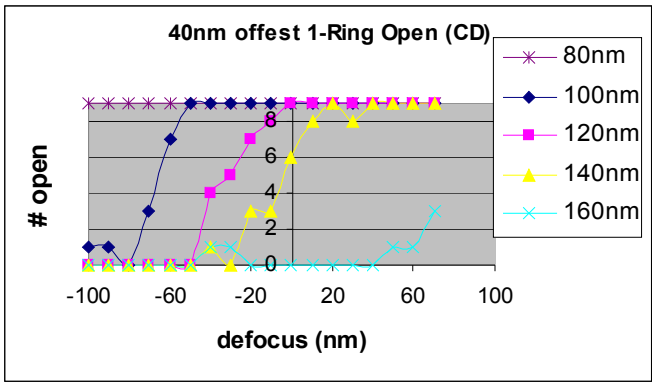


Figure 83 One type of target vs defocus for different CD values. The bigger the linewidth the more defocus required to pinch it open.

Expanding the linewidth comparison to sum targets across all offset values a similar trend is seen. One important detail is that the slope of the curve becomes more shallow with larger linewidths. This shallower slope indicates that probe offset has a bigger impact on larger CD targets. Looking at the 120nm line, the first target pinches open at –40nm and all pinch out at 0nm defocus. This means that all 45 targets pinch open in a relatively tight 40nm range indicating very little sensitivity to offset. The 140nm line on the other hand has the first targets pinching open at –20nm and never reaches all 45 targets opening at 70nm defocus.

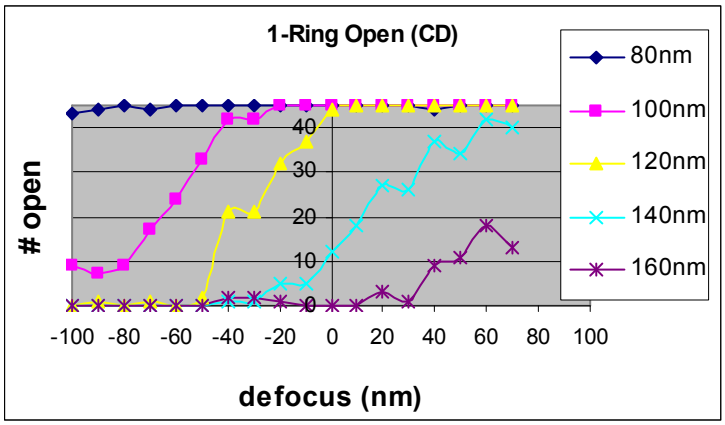


Figure 84 The number of opens vs defocus for different size lines. As there are nine instances of each cell and 5 targets per line, there is a total of 45 possible opens.

Looking at the 140nm line and 120nm line in more detail the range is indeed bigger for the 140nm line (Figure 85 and Figure 86). The 140nm line is significantly bigger then the probe and is much harder to burn through. In fact the most sensitive version for the 120nm line (no offset) is now the least sensitive version. This is most likely due to the fact that the 140nm line has too much signal in the middle of the line and attacking the edge is more effective. Clearly summing all four curves in both plots will lead to a steeper curve for 120nm than for 140nm.

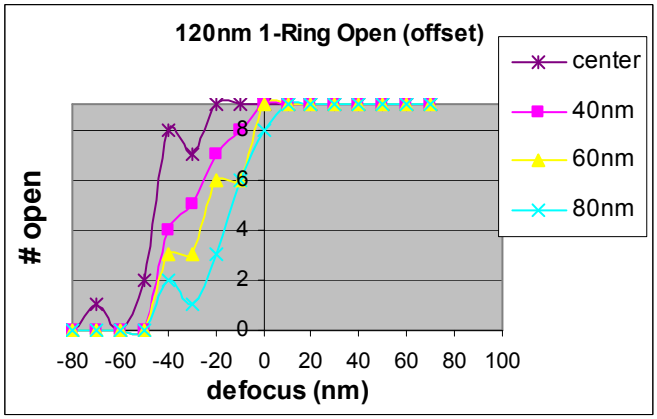


Figure 85 120nm 1-ring targets with different offsets. The smaller the offset or more centered the target the less defocus is needed to cause an open.

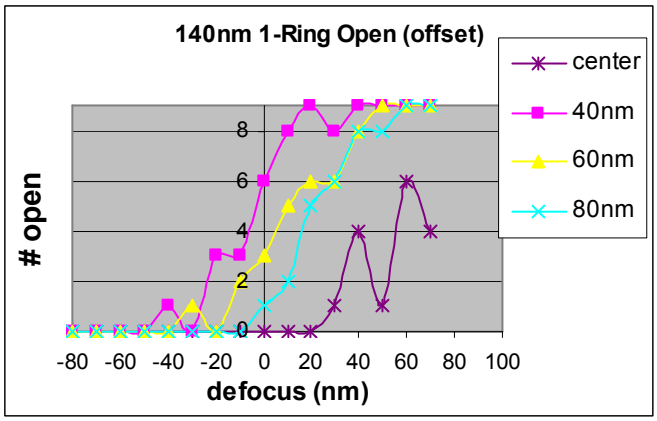


Figure 86 140nm 1-ring targets with different offsets. Notice how the range is bigger than the 120nm target. Notice how the centered target now requires the most amount of defocus to pinch open.

One issue with these targets is that the relationships between defocus and pinching changes with dose. In Figure 87 the 22mJ/cm^2 data is a lot more noisy as these measurements where made using 1V source. The slope of the curve is less steep, partially due to the higher level of measurement noise, but also due to the fact that the 120nm line is printing wider. The wider the 120nm line prints, the more it will act like the 140nm line in Figure 86.

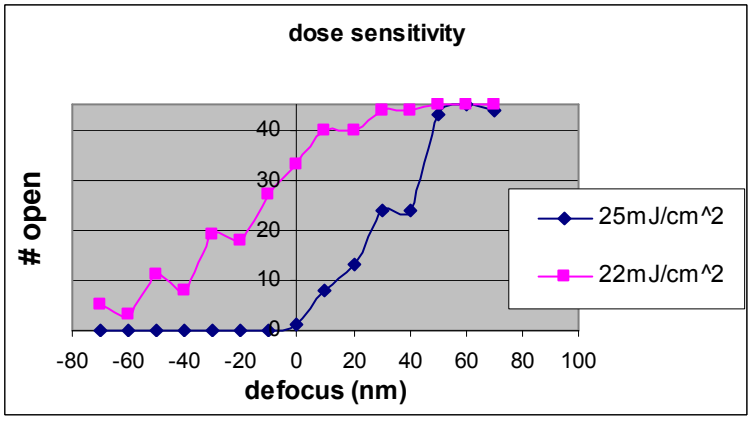


Figure 87 120nm 1-ring line at two different dose conditions. These values are summed across all offsets. (NOTE: 22mJ data is using 1V and is a lot more noisy)

7.5.1. Simulation vs Experiment

Both simulation and electrical results have demonstrated that the defocus monitors can be programmed to pinch open at various levels of defocus. The sensitivity or accuracy is related to the number of targets pinching open per nanometer of defocus. These plots are compared for both experiment and simulation in Figure 88. As there were different numbers of targets in simulation and experiment, due to the redundancy in experiment, the plots were normalized by dividing by the total number that pinched open. Due to the large etch bias that etched through the 80nm lines, the experiment data only had 100nm-140nm lines and simulation had 80nm-120nm lines. Other than the shift in defocus, the two plots line up fairly well. What is important to note is the slope of each plot. A steep slope indicates a large number of targets pinching open in a small range of defocus.

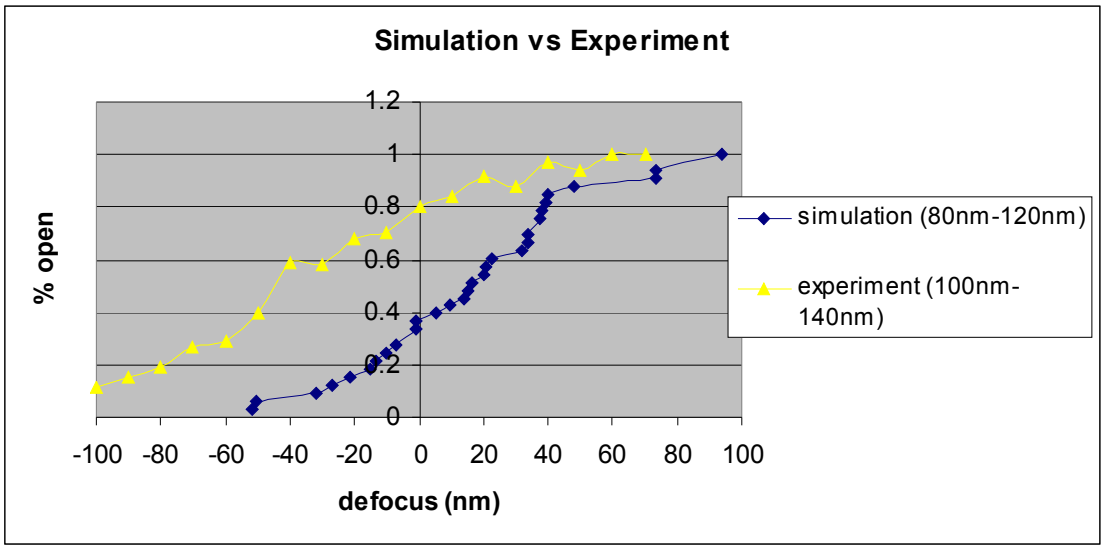


Figure 88 Plot of normalized data from Figure 82 with simulation data of the same targets under the same illumination conditions overlaid. The shift between the curves is most likely due to a significant etch bias during experiment as well as a difference in nominal defocus values in simulation and experiment.

After some more data manipulation to make the simulation resemble experiment, the curves line up even better (Figure 89). Indicating sensitivities or accuracy in simulation is well correlated with sensitivity in experiment. The new blue line in Figure 89 is a shifted version of the simulation curve in Figure 88 with some simulated noise. Each point in Figure 88 was reproduced 9 times with a random defocus bias, that corresponds to the 40nm or so uncertainty range found in Figure 81. The 50 nm shift between curves is most likely due to the offset between best focus in experiment and simulation. In simulation it is actually +20nm, so in experiment it is probably around –30nm. Adding the 50nm shift to the simulation data makes the plots overlap. Again the most important aspect is the slope of the curves, which matches well between experiment and simulation. Hence the sensitivity or fine granularity found in the simulation study with 454 unique targets should be reproducible in experiment if the two layer experiment is executed.

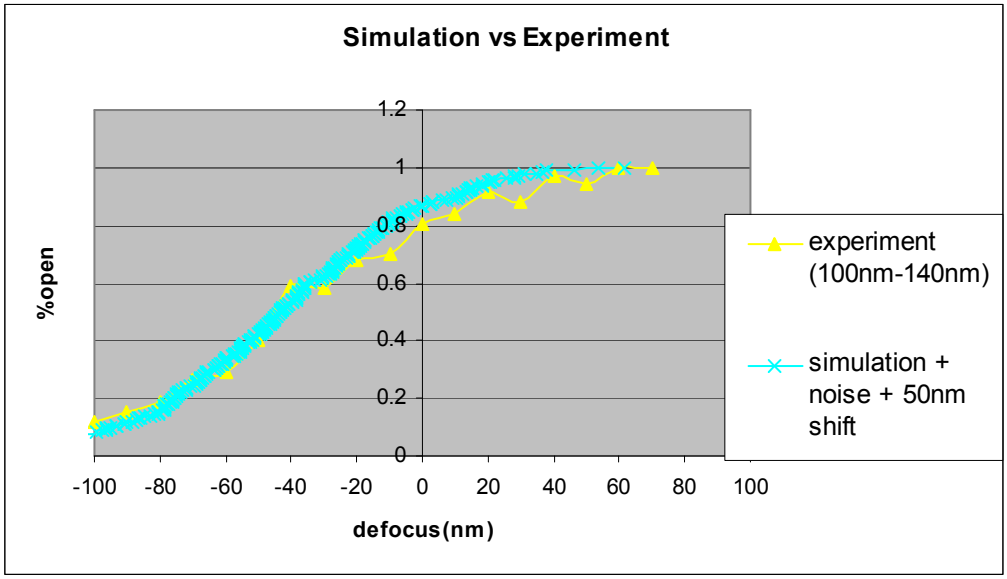


Figure 89 Similar plot to Figure 88, but with simulation results made to look like experiment. Each simulation point got reproduced into 9 points with about 40nm extra random noise in defocus pinch point. Each point was also shifted by -50nm in defocus.

7.6. Discussion

Overall the electrical targets showed high sensitivity or granularity over the full range of defocus values tested. Looking at individual targets it is possible to estimate defocus levels with a 40nm accuracy. With the strategy of redundant structures and multiple targets with different layout parameters it is possible to approach a few nanometers of defocus extraction accuracy. The accuracy is limited by random noise such as LER and dose sensitivity. Changes due to dose variation can be accounted for with complementary dose characterization structures, such as dense pitch ELM structures. The best remedy for random noise is redundant data. Fortunately one of the big benefits of this method is that electrical testing can provide the large volumes of data necessary for high accuracy. The noise itself can be analyzed and used for estimating the amount of noise in the pattern transfer process. Pinching has been shown to be hard to predict, so analyzing sets of targets with programmed pinch points and comparing them to simulation results can help calibrate models and identify hotspots[57]. Finally this structure can be improved by using a multi-layer process with aluminum probe pads, as is standard in a CMOS flow. This would lead to lower and more consistent contact resistance, which is necessary for other electrical structures, such as ELM.

As for defocus extraction, the database makes it easy to slice and dice the data into sets of targets and can help facilitate data analysis. A basic approach to defocus extraction would be to first find the relationship between defocus level and number of open targets for each set and then use the information in reverse to guess defocus. The defocus guess can then be further refined with a similar method to the one explained in chapter six where the database is used to iterate to a solution. The error between the

predicted amount of open targets in each set and the measured amount of targets in each set can be the metric for accuracy.

In terms of monitoring production, the full set of test structures can be used to identify outliers quickly and then smaller subsets can be used for detailed analysis and to extract actual defocus values wherever necessary. Looking at Figure 82 through Figure 87 one can notice a deviation on the –30nm and +40nm dies. There seems to be a consistent dip at these values in all the graphs. These deviations can be easily identified by looking at plots of total targets clearing vs defocus (Figure 82). These plots can be made more sensitive to small focus changes in production by designing most of the targets to break at small defocus values. This may require knowing the production doses and etch biases. If the total number of open targets deviates from the expected number open by some threshold amount, a more detailed analysis of the individual targets can quantify the amount of deviation from nominal.

In the multi-layer experiment the defocus matrix has 225 targets in a 30umx30um area, which would provide more statistics to identify defocus specific variations. Having the targets in a small area reduces their susceptibility to across slit dose non-uniformity and other systematic effects. If CMOS logic is available a decoder and a MUX can be used to address 2^n targets with only n-pads. Since the difference between an open and short is very large, parasitic leakage from control logic is not a concern. Targets can be tuned to have more redundancy and high sensitivity over a small range of defocus values. If these targets are used in parallel with dose sensitive structure such as ELM it should be possible to have accurate litho monitoring on the metal layers, which may have more

focus sensitivity than poly. The higher litho sensitivity stems from a wider array of geometries as well as more underlying layers that can lead to CMP non-uniformity issues.

7.7. Conclusion

Electrical and simulation data from the programmable single layer defocus monitor showed excellent results with high accuracy and large range over which defocus targets pinched open. This was achieved by modifying four layout parameters, CD, offset, number of rings, and probe size. Experiment results showed 20 defocus targets pinching open for every 10nm of defocus. Even though only 50 unique targets were used, random noise in experiments blurred the defocus pinch point of each target and created a fairly continuous string of targets pinching open. Simulation was able to reproduce these results indicating that simulation can be used as a guide for programming these targets. A more elaborate design with 454 unique targets showed in simulation at least one target pinching open for every nanometer of defocus. Although not enough experimental data was measured to evaluate the accuracy of the single layer defocus experiments, it is expected to be below 10nm, which is considered high accuracy for defocus monitors[58]. This level of accuracy is achieved by looking at large numbers of test structures and using the database to iterate to an extracted defocus value that would predict the targets that pinched open accurately.

8 Future on Process Characterization Methodologies

This chapter describes some low hanging research fruit that leverage process sensitive electrical test structures, characterization with the PYS, and the data analysis strategies in this thesis. These methods can provide further evolutionary steps in the process characterization strategy with implementations that are more pragmatic in a "real world" setting. Chapter 6 demonstrated high accuracy lithography parameter extraction can be achieved with a high number of transistor based test structures. The best structures were identified through simulation and have been improved and redesigned into a much smaller footprint that could potentially be fit into an 80um scribe line. Misalignment sensitivity issues with the defocus monitors have been mitigated by moving to a one dimensional version and multiple instances of the overlay structure were created to enable more averaging and hence more accurate misalignment extraction. Also a new set of electronic process characterization test circuits has been proposed in the past and can be expanded with the discoveries made in this thesis.

Electronic testing can be critical enabler for "DFM quality" process characterization. "DFM quality" process characterization should be very accurate, detailed, and comprehensive as DFM tools are only as effective as the processes are understood. What better place to provide process window information than to designers themselves when

they test their chips. This notion has already had some traffic in the designer world, so some silicon data from a previous testchip that had electronic testing and looked at ring oscillator frequency variation has shown pattern dependent variation. This data is compared to PYS results in this chapter. The ultimate goal for the test structures and the high volume extraction strategy is to create a set of electronically testable process monitors that could potentially be inserted into the white space on production chips for process monitoring and process characterization purposes.

8.1. Scribe Line Lithography Monitors

A redesigned set of structures based on the FLCC Enhanced NMOS testchip has a 110umx80um area, which is small enough to fit in a scribe line and can be used for accurately characterizing dose/systematic variation, defocus, and misalignment. These structures are electrically probable transistors that could be incorporated into electronic process monitoring.

8.1.1. Test Structure Design

A zoomed out view of the new layout can be seen in Figure 90 where there three main types of structures testing for dose, defocus, and misalignment. Dense pitch structures that have little sensitivity to dose were designed in sets of fourteen lines strewn across seven 2000nm active regions. Each line is extended an extra 2000nm below the bottom active region and 2000nm above the top active region to minimize misalignment sensitivity. This means that at each pitch there are 12 x 7 or 84 individually probable dense pitch structures (this is excluding the two outside denso transistors). The overlay structure from chapter six was redesigned into a dense pitch format with poly overlap

ranging from –20nm to + 120nm in 1nm increments(Figure 91). This structure was repeated 10 times so that each structure can be averaged and noise can be reduced. Finally the defocus monitor was redesigned into a vertical one-dimensional structure that is completely insensitive to misalignment.

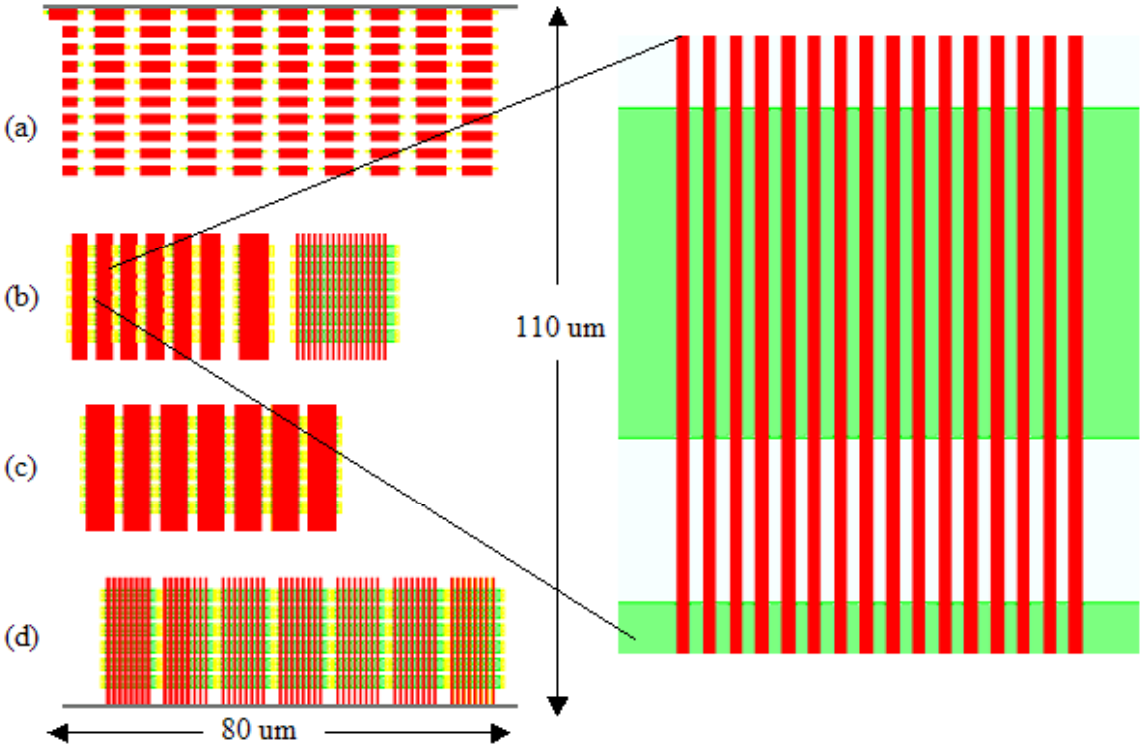


Figure 90 Four flavors new scribe line structures have a small footprint, less misalignment sensitivity, and more redundancy. (a) overlay (b)dose (dense pitch) (c) defocus dense (d) defocus iso.

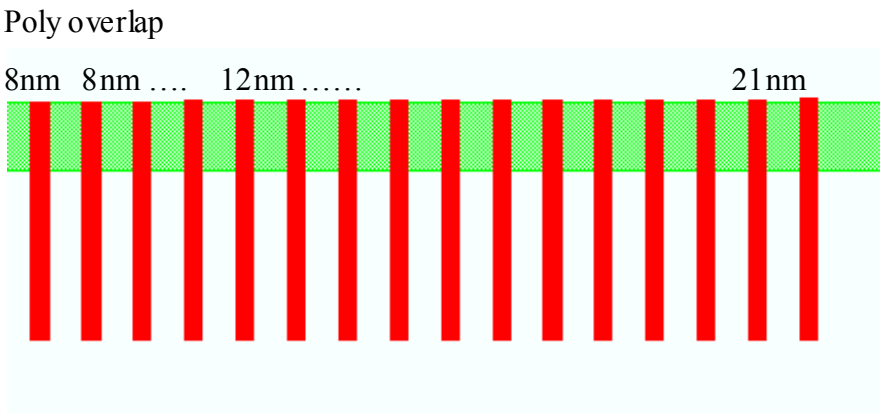


Figure 91 The new overlay structure is in a dense pitch for higher packing density. The outside two lines have the same poly overlap as they are considered denso lines and may have more line end pullback. In the inside 13 lines the poly overlap increments by 1nm from left to right.

The 1D defocus test structure again used the 90 degree phase etch to achieve asymmetric response through focus. The pitch was matched to the ring distance in the 2D structure and the 90 degree phase shift was implemented in seven different versions (Figure 92). The hope is to get different sensitivities through focus, a key factor in estimating focus with high accuracy. The structures are insensitive to misalignment in both the vertical and horizontal direction, so die to die misalignment variation will not decrease accuracy.

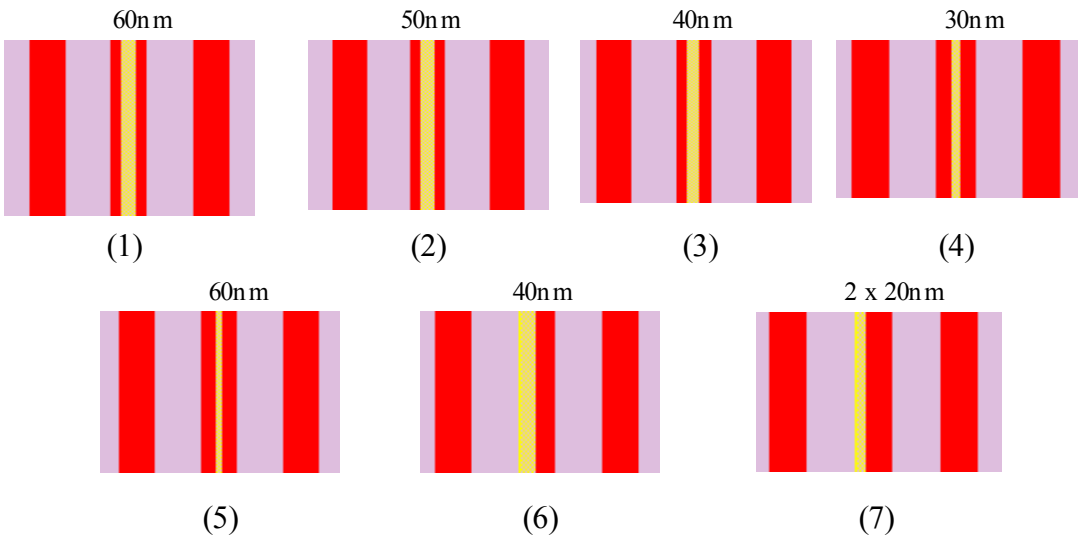


Figure 92 Seven different versions of the 1-dimensional defocus monitor. The red line is poly and the yellow line is glass with a 90 degree phase etch. The dimension above each version is the width of the 90 degree phase etch. The sum of the poly and 90 lines is always 120nm.

8.1.2. Simulation Results

The scribe line test structures were simulated in the Parametric Yield Simulator and results were uploaded to the database. An optical model was generated for Quasar illumination with 193nm lamda, 0.92/0.72 sigma out/in, and a 40nm illumination angle. The simulated CD did not change at all with misalignment in the dose and defocus structures. Pitches used for dose estimation needed to be scaled by lamda/NA or by 22.2%. The smallest pitch ended up being 160nm, which required an increased dose to print. The process window was hence centered around 110% nominal dose. No OPC was used in this design.

The key to these monitors is the 90 degree phase etch that enables asymmetric response through focus. One goal of this thesis is to underline the value in using such a mask. The 90 degree phase etched regions are written with a laser tool so extra mask cost is not

high. These lines are 4X more sensitive than an isolated line and the example in Figure 93 shows an 8nm shift in CD between 0nm and 40nm defocus. For an isolated line this change is essentially 0nm (primarily because the actual best focus is at +20nm and the plot is symmetric). This high sensitivity at low defocus values is a must for being able to accurately characterize lithographic variations in a production setting.

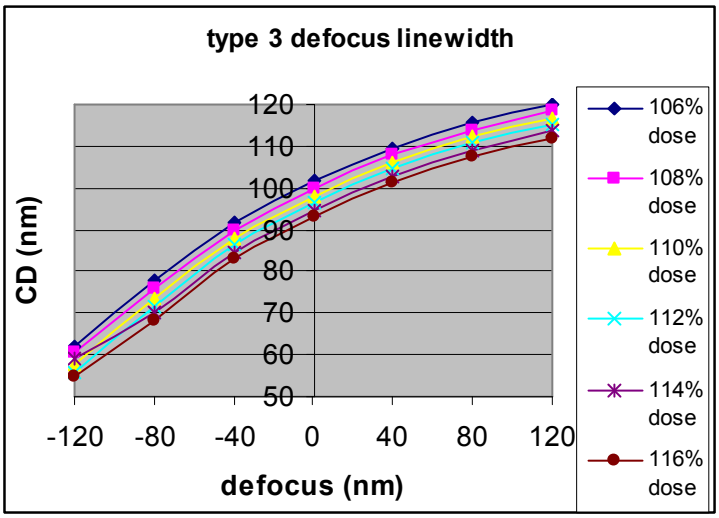


Figure 93 Bossung plots for the type 3 defocus monitors. Almost 4X higher sensitivity than an isolates line and a 8nm step in CD between 0nm and 40nm defocus.

The one dimensional defocus monitors performed better than expected. Although a fair comparison cannot be made as a different wavelength was used than in Chapter 6, the one dimensional structures had at least equivalent performance to the 2D structure. A big reason for this is that the entire transistor gate varied instead of just a center pinching area. If one were to look at the very center in a 2D defocus target, the line edge varies a lot more than the 1D version, but when looking at the transistor as a whole the 1D version has more signal. Most importantly it has no misalignment sensitivity, which was a source of accuracy loss in defocus extraction before.

8.2. Electronic Testing of Process Monitors and Interpretation by PYS

The ultimate goal of these test structures and the high volume process extraction strategy is an electronic implementation where data can be easily accessed in production chips and used for accurate process window extraction. Efficient electronic testing requires intelligent circuits for measuring these test structures with limited or no outside pins used for testing. This can be achieved by automatically testing test structures if a particular bit or flag or pin is enabled and by writing test results into memory or using a scan chain to read out results[16]. The benefit of writing results into memory is that no additional area is required for data storage and no outside devices are needed for measurement. Ring oscillator frequency and transistor leakage current can both be measured in an electronic fashion. Measuring ring oscillator frequencies gives information about drive current and delay variation and leakage current measurements can be used for accurate threshold voltage variation assessment.

8.2.1. Process Sensitive Ring Oscillators Study

Ring oscillators (RO) have been extensively used in testchips and offer a well understood platform for evaluating variations[16,59,60]. Ring oscillator frequency is closely tied with drive current and gate length and can also be designed to be interconnect loaded to measure resistances and capacitances in interconnect. A Berkeley testchip taped out by Liang Teck Pang utilized ROs to measure pattern dependent variations and found significant biases between dense and semi-dense (or denso) features. He had twelve individually addressable patterns for which he measured ring oscillator frequency and

leakage current on 36 chips[59]. He used a circuit approach that spits out a divided frequency that could then be measured by an off chip oscilloscope[60]. The results of the dense vs denso bias can be seen in Figure 94.

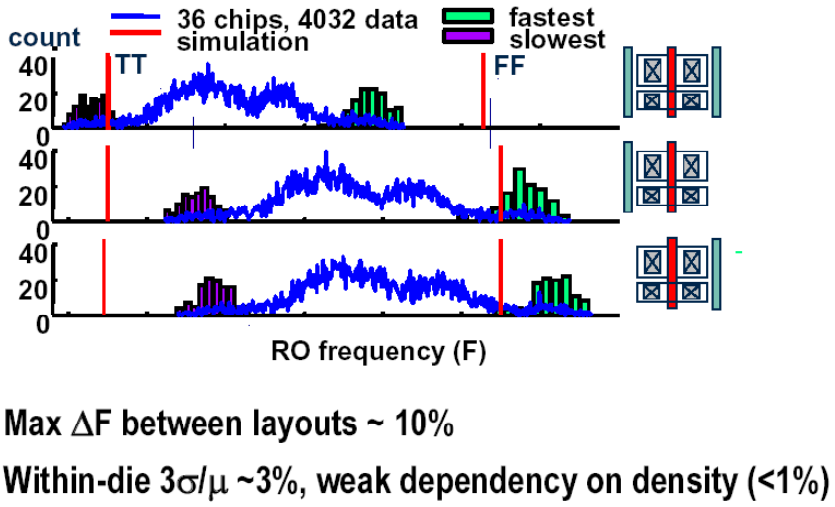


Figure 94 Ring oscillator frequencies for dense and denso transistors from Liang Teck's 90nm CMOS testchip.

Looking at the distributions in Figure 94 the denso transistors have a shift and a larger spread of frequencies so must be more susceptible to process variations. This shift or bias is most likely due to an etch bias on the edge that has no adjacent feature. If all the denso gates are smaller, the RO frequency is faster and hence there is a systematic bias in RO frequency where the denso gates are faster. It is expected that the isolated line might have twice as large an etch bias, but there is unfortunately no ring oscillator data from the isolated feature. The ring oscillator spread is bigger for the denso line, indicating a larger susceptibility to process variations. Naturally the dense and denso lines are going to have different image qualities that will respond differently to defocus, so defocus is high probability cause.

To better understand what the potential causes could be the two features were run through the PYS and their sensitivity to defocus was quantified. The dense lines simulated to have around 60% more frequency variation through focus, which could help explain the higher level of variation in denso line in Figure 94. Since the illumination settings and OPC strategy (which could include SRAF insertion) were not known it is not possible to make quantitative comparisons, but the trend has been reproduced and defocus variation is a highly likely culprit.

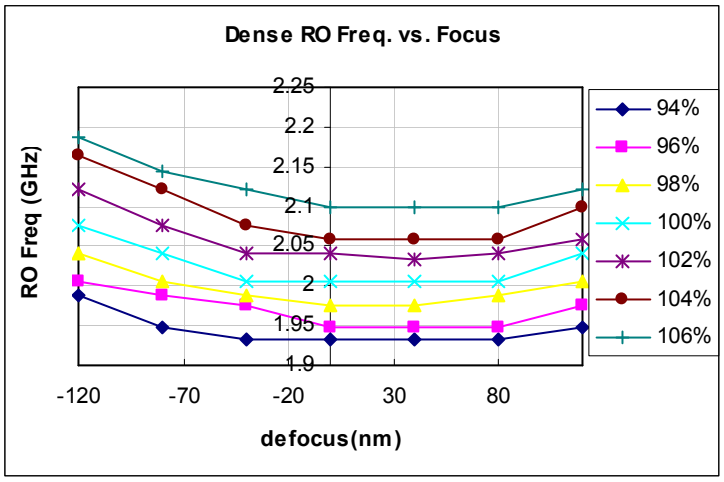


Figure 95 Simulated ring oscillator frequency vs defocus for the dense structure in Liang Teck's testchip.

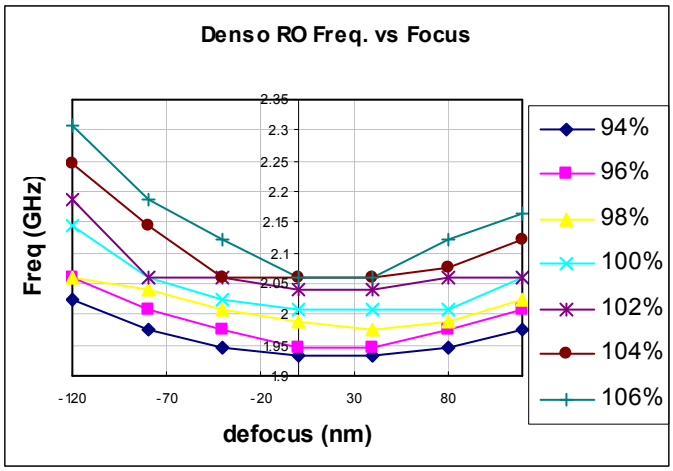


Figure 96 Simulated ring oscillator frequency vs defocus for the denso structure in Liang Teck's testchip. This structure shows 63% more variation than the dense line.

Looking at simulation and experiment more closely, the trend is predicted but the difference in stdev is a lot bigger in simulation. This cold be due to the mismatch in illumination conditions or from the fact that defocus is not the primary source of the discrepancy between dense and denso lines. One of the results of this study is that the Parametric Yield Simulator can be used to determine a quantifiable difference between layouts in terms of sensitivity to lithography process parameters. If the illumination settings were known and the post-OPC data was available, it is conceivable to use the PYS to diagnose the source of the problems. Being able quantify the difference in sensitivity to different process parameters can help conclude if specific sources of process variations dominate. If more process specific test structures were used, these conclusions could be made with more confidence. Ring oscillators offer an important insight into process variations as delay variability can play a very significant role in determining parametric yield.

	Experiment		Simulation
	mean freq	stdev freq	stdev freq
dense	1.1	1	1
denso	1	1.25	1.63

Figure 97 Table of RO frequency differences between dense and denso lines. Simulation predicts a 63% increase in sensitivity to defocus for the denso vs dense case and experiment shows a 25% increase.

8.2.2. Other Electronic Circuit Ideas

One draw back of the before mentioned test circuit is that it relies on an off chip frequency counter or oscilloscope. If ring oscillators were placed on a production chip, there would be a benefit to containing all the measurements on-chip. Care would have to be placed not to generate extra noise from varying measurement circuitry, but a solution should be possible if an outside clock is used. IBM used a counter matched with a decoder to quickly address and test 64 ROs[16]. This strategy could be expanded to using the original counter to address a memory space, use the RO output to drive a RO_FREQ counter, and then latch data into memory as the main counter is incremented (Figure 98). If the divider is larger enough to reduce the clock frequency and give the RO enough time to increment the second counter on average 10,000 times, RO frequency should be measured with 0.01% accuracy. A more in depth literature search is necessary to establish the novelty of this approach, but in either case it has an attractive value proposition.

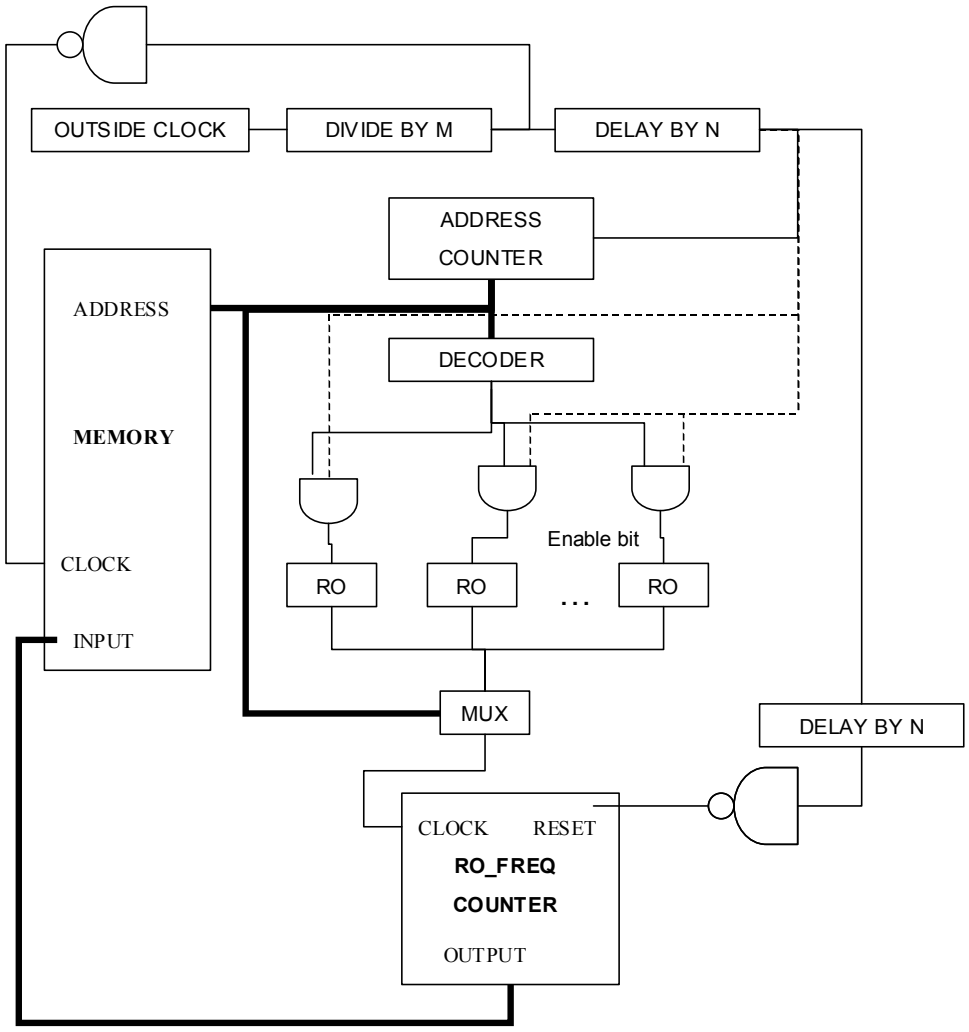


Figure 98 Schematic of an automatic RO variation testing circuit. The frequency out of the divided clock is four orders of magnitude greater than the frequency out of an RO test structure. Counter output at the bottom is going to be a linear function of RO frequency. The thicker wires represent more than one bit line.

The signals in Figure 98 propagate in the following way. The outside clock drives a divider, which is essentially a counter with one output that is a fraction of the input frequency. When this signal goes high it latches the fast counter into memory and after a delay of N it increments the address counter, which with the output from the decoder enables only one RO (NOTE the first memory address is going to have a dummy value of

0). The RO drives a RO_FREQ counter which is disabled with a delay until the RO frequency stabilizes. After a delay of 2*n the fast counter is enabled and starts counting. When the divider output finally goes low it latches the fast counter's value into memory and after a delay N disables the RO. After a delay of 2*n the RO_FREQ counter is reset. Now everything is off and the circuit waits until the divider output goes high at which pint the cycle repeats. If the RO frequency is around 1GHz, the divider can operate at 100KHz and measure 10,000 test structures per second. The data can then be read out of memory and all that is needed is an outside clock that makes sure the measurement time is constant between ROs. A test enable bit can be used to disable this circuitry during normal operation.

8.2.3. Electronic Leakage Current Circuit

A similar circuit can be created for measuring leakage current. The benefit of measuring leakage current is that it has a much stronger dependency on gate length, which enables the detailed process extraction strategy described in Chapters 5,6 and 7. The basic idea of this circuit is to count how many clock cycles it takes for the leakage current flowing through one device under test to charge up the input gate in a FLIP FLOP. This time can be programmed by the size of the input gate in the flip flop, so high accuracy can be achieved by using a lot of clock cycles on average.

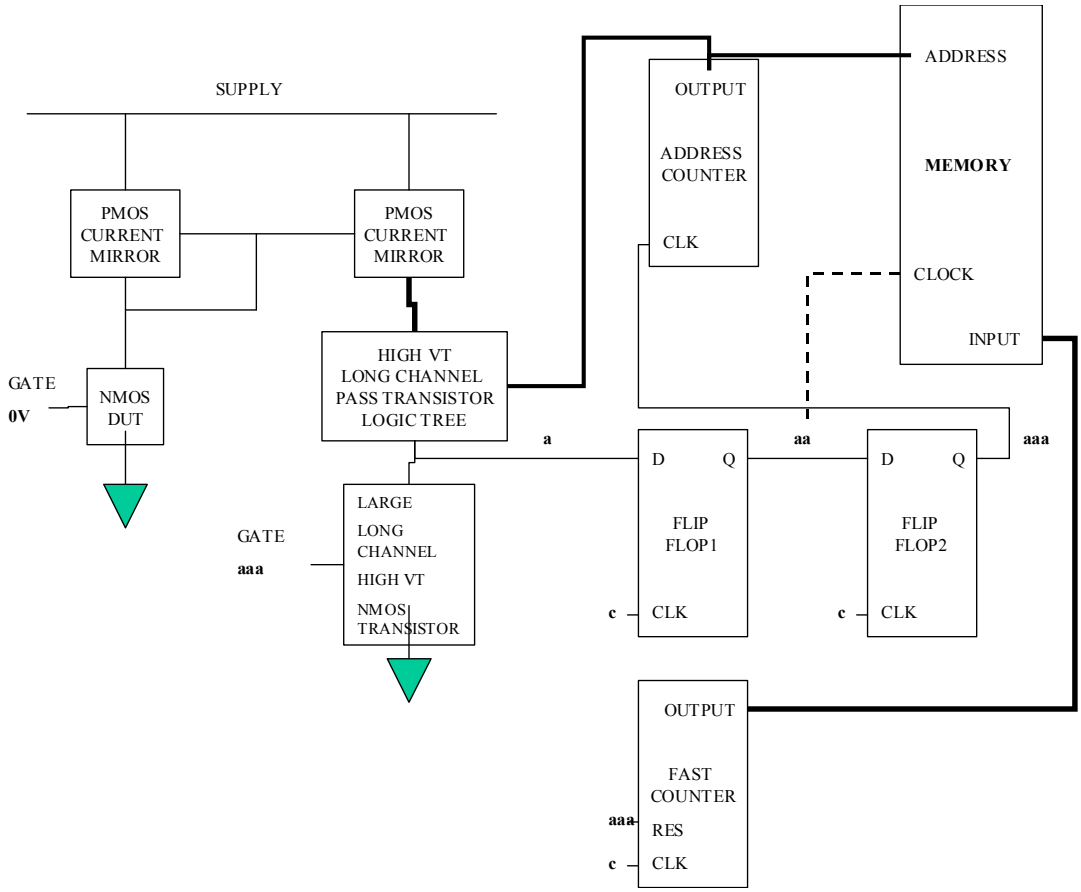


Figure 99 Electronic leakage current circuit. This circuit measures how long it take for leakage current to charge up the input transistors in FLIP FLOP1 and records the value in memory. The 'c' signal is the main clock, 'a' is the input into the first FF, 'aa' is output from the first FF, and 'aaa' is the output from the second FF. The thicker wires represent more than one bit line.

The circuit in Figure 99 is a mixture of discrete transistors, an analog current mirror, and some standard digital logic such as flip-flops and counters. To minimize parasitic leakage and variations in the sensing circuitry, all but the device under test (DUT) need to be high Vt long channel devices that have very little leakage current and are very stable through process variations. The signals are propagated in the following way. After setting up the circuit the 'a' signal should be at 0V as the large NMOS reset transistor would have grounded that node. The fast counter that is clocked on the main clock starts

counting as soon as the large reset NMOS transistor is turned OFF. Once the reset transistor is turned OFF, the leakage current through the device under test (DUT) is mirrored in a current mirror (one current mirror per DUT) and starts charging the input to Flip Flop1 (FF1). Since the leakage current is orders of magnitude smaller than the drive current, it will take many clock cycles to charge the input gate of FF1 to a point where it changes the output to high. On the clock edge after the input is high enough to register a VDD on the output, the fast counter value is written to memory. On the next clock edge, the address counter is incremented, changing the memory address and the input to the pass transistor logic tree that now routes the current from the next DUT. The main counter is also reset and the large reset NMOS transistor is turned ON, so the 'a' signal is now grounded or 0. On the next clock edge the '0' input gets propagated to 'aa' and on the next clock edge the '0' is propagated out to 'aaa'. When 'aaa' becomes zero again it turns off the large reset transistor and sets the fast counter reset to low and the process starts all over again for a new DUT. Now the previous memory address has the counter value for the amount of time it took for the FF1 input to charge to a state when the output was set to one. The key to consistency in this circuit is to make sure that input capacitance of FF1 is consistent from die to die and that a '1' is latched at the same input gate voltage. For this reason special care will be needed in designing FF1. All the devices under test need to be routed into the same FF1, so even though large gate length will lead to a bigger cell, it will not require a large total area.

8.3. Conclusion

This chapter describes additional steps that can deliver the process characterization strategies developed in this thesis into a small footprint and potentially electronic format,

which can then be utilized for accurately extracting the process window from production chips. A previous RO study from Berkeley showed pattern dependent RO frequency variation. It would be possible to us the PYS for hypothesis testing and analyzing the potential sources of the discrepancies. Trends can be identified, but to be able to quantify the significance of specific process parameters, an accurate optical model is going to be needed. Additionally, the accuracy and confidence can also be significantly increased by using more process specific test structures that have a broader range of sensitivities. The redesigned test structures could easily fit into a scribe line where similar RO frequency studies could be made. Other structures could be added such as minimum size transistors for monitoring random noise from LER and RDF as well as others from the original Enhanced NMOS experiment. The data provided from these process specific structures at volumes where sources of random noise can be averaged out and evaluated can prove very valuable. Electronic testing, which is already a standard practice in industry, can meet the high volume requirements for high accuracy process characterization.

9 Conclusions

This thesis addresses two of the biggest challenges in DFM, a methodology for injecting process information further upstream and a methodology for characterizing the process window and identifying what portion of transistor performance variations can be explained systematically. The challenge of pushing information into design has been met with the Parametric Yield Simulator, which uses the transistor model as the new bridge between design and manufacturing. The transistor model can translate any and all sources of process variations into transistor performance metrics the designer can learn to use consistently. The Parametric Yield Simulator has also been used to prototype and create a suite of test structures that could be used for comprehensive process characterization. To facilitate data analysis, a database has been created that can link experiment and simulation. The process specific test structures that have been designed were implemented in two testchips, both of which were thoroughly simulated in the PYS and one of which was successfully manufactures. In total a set of five main innovations/results make up the answer to the two biggest challenges in DFM. This chapter will summarize these five innovations and give indications as to how these projects can grow in the future.

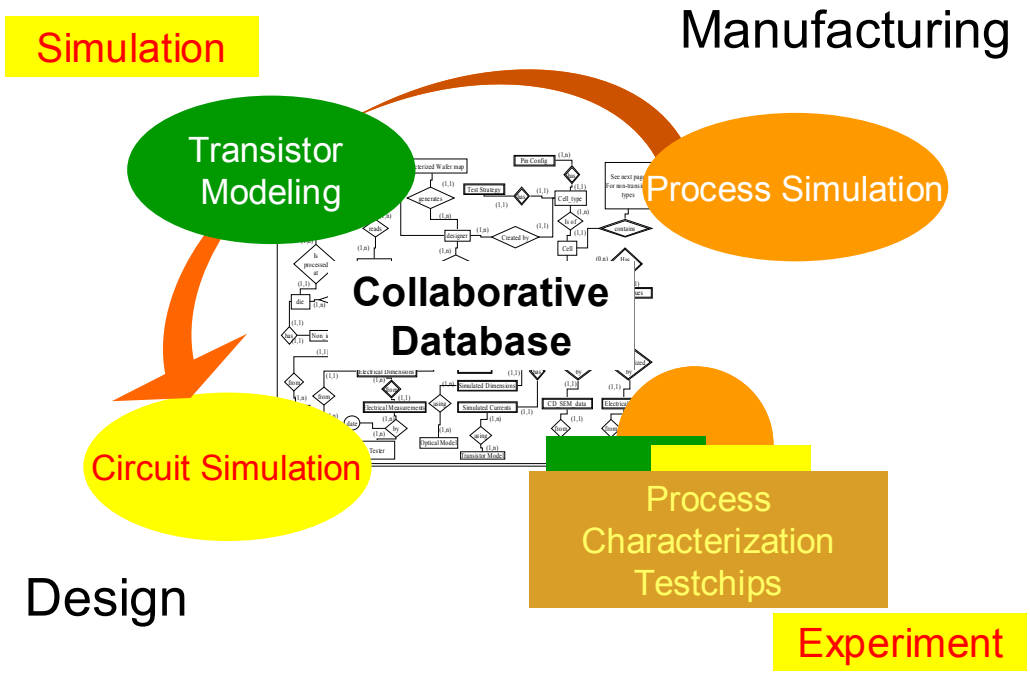


Figure 100 The collaborative platform for DFM consists of the Parametric Yield Simulator that links process, device, and circuit simulation and a set of process characterization testchips. The collaborative database serves as the glue between the two halves.

The first innovation is the Parametric Yield Simulator, which was built as a platform for injecting process information into circuit design. It is built around a non-rectangular transistor model that translates a two dimensional physical gate shape into an equivalent one dimensional compact transistor model. The transistor model forms a critical new link between design and manufacturing as it can encapsulate any and all sources of variation. As long as the standard interfaces between modules are maintained, improvements to the independent modules should integrate seamlessly. If the process module is expanded to include more pattern transfer steps such as etch or systematic across die biases, the device module will not need to be changed as long as the input is a list of slice lengths. If strain is modeled then both the device and process modules would need to be changed, but the output from the device module would remain a transistor model. On the circuit side,

process aware transistor models can be simulated in circuits and design centric tools can be developed. One low hanging fruit on this end is a performance based robustness metric that can be used in technology mapping or placement during synthesis. These tools can be developed assuming the input is always a set of process aware transistor models.

The second major innovation is the Enhanced NMOS testchip that has dozens of flavors of different types of test structures that have programmed sensitivities to different process steps. The testchips have a broad scope of structures with the hope to capture and characterize all the main sources of transistor performance variation. The Enhanced NMOS testchip leverages a passively multiplexed Enhanced Transistor Electrical CD Metrology testing strategy that enables very high packing density and minimizes parasitic leakage. Enhanced NMOS transistors are strategically destabilized transistors that are 3X more sensitive to gate length variations. This enhanced short channel effect can be achieved with a 10X higher dose extension implant. The resulting transistor has a higher correlation between gate length and leakage current so electrical leakage current measurements can be more accurately translated into critical dimension data, which can then be analyzed for sources of variation. This method could potentially be modified to make the transistors more sensitive to channel doping or oxide thickness if those variations became of interest. This method has none of the geometrical limitations of Electrical Linewidth Metrology and scatterometry so structures can be made sensitive to many different process effects. The Enhanced NMOS testchip has structures that look at defocus, dose, misalignment, LER, RDF, long range dopant fluctuation, active corner rounding, poly corner rounding, non-rectangular transistors, pass transistor logic, and

SRAM variability. With six contributing students the premise was six students, six sets of conclusions, all on one chip.

The third major innovation is the collaborative database which was built to facilitate data comparison and aggregation. The database has been customized to not only serve as centralized and organized data storage, but also as a platform for collaborative data analysis. This was enabled with three features; an organized data structure that enables queries to slice and dice the data as necessary, a flexible schema that allows the database to expand and account for attributes as they become important for data analysis and a website that serves as the front end to the platform. On the website students can reuse each others queries and make them better. The best and most frequently used queries are tracked so that they bubble up to the top. The database also has an interface with the Parametric Yield Simulator that enables running lots of simulations with automatic uploading of results to the database.

The fourth innovation or result is the high volume process extraction strategy that was tested in a dry-lab simulation experiment of the FLCC Enhanced NMOS testchip. TSUPERM 4 and Medici where used to create Enhanced transistors and the experiment portion of the database was populated with current values. A significant level of random noise was added inside the PYS in the form of LER and a random bias. Yet with 5nm 3-sigma CD variation for the same feature in one die and 17.2nm variation when different proximities were included, it was possible to extract defocus and misalignment with sub 10nm accuracy. This level of accuracy has not been demonstrated previously with electrical metrology, which would be necessary for high volume testing. The wide array of structures that had different responses to dose, focus and misalignment were the key to

attaining such a high accuracy. With a range of sensitivities, each process parameter has a unique signature on all the test structures. So looking at all the test structures together and using an iterative guessing process inside the database it is possible to address individual process parameters, even if confounding parameters are present. By looking at the residuals it is possible to evaluate goodness of fit and statistically establish if the extracted values are accurate. One of the most important structures was the defocus monitor, which had a 90 degree probe. This structure has asymmetric response through focus and a high sensitivity of linewidth to defocus at low defocus levels. This type of structure will be necessary in monitoring production lots as production wafers generally have very low levels of defocus.

The fifth major innovation or result is a programmable single layer defocus monitor that was characterized through simulation and experiment. By changing four layout parameters, CD, offset, probe size, and the number of rings, it is possible to program the sensitivity and range of focus values over which the target works. A second testchip includes 504 unique combinations of those four parameters in an electrical open/short test structure. Four hundred fifty four combinations in a two layer version and 50 combinations in a single layer version that was manufactured and electrically tested. The test structures showed great range and sensitivity in both experiment and simulation with various defocus targets breaking at different defocus points between –120nm and +120nm. On average 20 defocus targets pinched open per 10nm of defocus leading to sub-10nm defocus resolution if the database is used for looking at sets of test structures.

The Parametric Yield Simulator, process specific test structures, Enhanced NMOS transistors for electrical metrology, and database for high volume data analysis have been

developed and explored. Short loop experiments and high volume dry-lab simulation experiments have explored new frontiers of automatic process characterization at a breadth and scope that enables "DFM quality" process extraction. The key points of leverage proved to be, process specific test structures, programmable sensitivity, redundancy, characterization through simulation or programmed experiments, a structured yet flexible database for data analysis, and use of a novel attenuated phase shift mask with a 90 degree phase etch for defocus characterization. Perhaps the most enabling feature in the architecture and implementation of the database and software platform is adaptability and flexibility to grow and expand. The Parametric Yield Simulator, database, complementary helper scripts, layout generation scripts and process specific test structures will be made available through the TCAD lithography wiki page. With a flexible platform, DFM research can adapt to new challenges as processes become more complex and challenging to control.

Appendix A: Reticle Catalogue

This appendix contains a description of all the reticles created as part of this research

1. FLCC NMOS Mask – This mask has a four layer 4mmx4mm chip. It includes four different treatments of OPC and placement for the poly layer. This mask is labeled FLCCDUPONT1.
2. FLCC Enhanced NMOS Mask – There were two versions of this mask taped out. The first version had two masks, one CLEAR FIELD and one DARK FIELD. These masks are labeled FLCC06_020A and FLCC06_255A respectively.
   A second version of these masks has been redesigned with no Cypress structures. The clear field mask was taped out and is called FLCC06_020B. The dark field mask was never taped out as the contact size and strategy was never finalized. The clear field mask does include some new ODP structures, but all of these have been reproduced on the ODP_ELM mask.
3. FLCC ODP ELM Mask – This is the short loop mask used to expose the experiments in Chapter 7. This mask has ODP structures from Yu Ben, Marshal Miller, Jung Xue, and Wojtek Poppe. It also has the single layer and two layer defocus experiments as well as some single and two layer ELM structures. This mask is labeled FLCC07ODPELM

Bibliography

[1] Maynard DN, Runyon SL, Reuter BB. "Yield enhancement using recommended ground rules." IEEE. 2004, pp. 98-104. Piscataway, NJ, USA.

[2] Matthew I, Tabery CE, Lukanc T, Plat M, Takahashi M, Wilkison A. "Design restrictions for patterning with off-axis illumination" SPIE, vol.5754, no.1, 2004, pp. 1574-85. USA.

[3] Jie Yang, Capodieci L, Sylvester D. **"**Advanced timing analysis based on post-OPC patterning process simulations." *SPIE, vol.5756, no.1, 2005, pp.189-97. USA.*

[4] Pileggi L, Schmit H, Strojwas AJ, Gopalakrishnan P, Kheterpal V, Koorapaty A, Patel C, Rovner V, Tong KY. "Exploring regular fabrics to optimize the performance-cost trade-off." IEEE. 2003, pp. 782-7. Piscataway, NJ, USA.

[5] Ran Y, Marek-Sadowska M. "On designing via-configurable cell blocks for regular fabrics." Design Automation Conference. ACM. 2004, pp. 198-203. New York, NY, USA.

[6] http://www10.edacafe.com/nbc/articles/view_weekly.php?articleid=392097&page_no=4

[7] Dan Perry, Mark Nakamoto, Nishath Verghese, Philippe Hurat, and Rich Rouse, "Model-based approach for design verification and co-optimization of catastrophic and parametric-related defects due to systematic manufacturing variations", SPIE, vol. 6521, 2007

[8] Darsun Tsien, Chien Kuo Wang, Yajun Ran, Philippe Hurat, and Nishath Verghese, "Context-specific leakage and delay analysis of a 65nm standard cell library for lithography-induced variability", SPIE, vol 6521, 2007

[9] Sachiko Kobayashi, Suigen Kyoh, Toshiya Kotani, and Soichi Inoue, "Process window aware layout optimization using hot spot fixing system", SPIE, vol.6521, 2007

[10] Qiaolin Zhang, Cherry Tang, Tony Hsieh, Nick Maccrae, Bhanwar Singh, Kameshwar Poolla and Costas J. Spanos, "Comprehensive CD uniformity control across lithography and etch", SPIE, 5752, 692 (2005)

[11] Paul Friedberg, Yu Cao, Jason Cain, Ruth Wang, Jan Rabaey, and Costas Spanos, "Modeling Within-Die Spatial Correlation Effects for Process-Design Co-Optimization", ISQED March 2005

[12] Allgair J, Gong Chen, Marples S, Goodstein D, Miller J, Santos F. "Feature integrity monitoring for process control using a CD SEM." SPIE, vol.3998, 2000, pp. 227-31. USA.

[13] Tabery C, Page L. "Auto CD-SEM edge-placement error for OPC and process modeling." Solid State Technology, vol.49, no.7, July 2006, pp. 81-2, 84, 86, USA.

[14] Tabery C, Craig M, Burbach G, Wagner B, McGowan S, Etter P, Roling S, Haidinyak C, Ehrichs E. "Process window and device variations evaluation using array-based characterization circuits." ISQED, IEEE Computer Society. 2006, pp. 6. Los Alamitos, CA, USA.

[15] Liang-Teck Pang, Nikolic B. "Impact of layout on 90nm CMOS process parameter fluctuations." IEEE. 2006, pp. 2. Piscataway, NJ, USA.

[16] Bhushan M, Gattiker A, Ketchen MB, Das KK. "Ring oscillators for CMOS process tuning and variability control." IEEE, Feb. 2006, pp. 10-18, USA.

[17] Poppe W, Holwill J, Pang LT, Friedberg P, Liu Q, Alarcon L, and Neureuther A" Transistor-Based Electrical Test Structures for Lithography and Process Characterization." SPIE, vol. 6156, no.13, 2007

[18] Jie Yang, Capodieci L, Sylvester D. **"**Advanced timing analysis based on post-OPC patterning process simulations." *SPIE, vol.5756, no.1, 2005, pp.189-97. USA.*

[19] Cain JP, Spanos CJ. **"**Electrical linewidth metrology for systematic CD variation characterization and causal analysis." *vol.5038, 2003, pp.350-61. USA.*

[20] Dusa M, Moerman R, Singh B, Friedberg P, Hoobler R, Zavecz T. **"**Intra-wafer CDU characterization to determine process and focus contributions based on scatterometry metrology." SPIE*, vol.5378, no.1, 2004, pp.93-104. USA.*

[21] Friedberg P, Cao Y, Cain J, Wang R, Rabaey J, Spanos C. **"**Modeling within-die spatial correlation effects for process-design co-optimization." *IEEE Comput. Soc. 2005, pp.516-21. Los Alamitos, CA, USA.*

[22] Balasinski A. **"**A methodology to analyze circuit impact of process-related MOSFET geometry." *SPIE, vol.5379, no.1, 2004, pp.85-92. USA*

[23] Y. Trouiller, et al "65-nm gate OPC optimization with simple electrical model simulation" *SPIE, vol.5756, no.38, 2005*

[24] Ortolland, C. Orain, S. Rosa, J. Morin, P. Arnaud, F. Woo, M. Poncet, A. Stolk, P., "Electrical characterization and mechanical modeling of process induced strain in 65 nm CMOS technology", IEEE, Sept 2004, pp. 137-140

[25] A. H. Gabor, T. Brunner, S. Bukofsky, S. Butt, F. Clougherty, "Improving the Power-Performance of Multicore Processors Through Optimization of Lithography and Thermal Processing", SPIE, 2007

[26] Nardi A. Sangiovanni-Vincentelli AL, "Synthesis for manufacturability: a sanity check", IEEE, 2004

[27] Orshansky M, Keutzer K. "A general probabilistic framework for worst case timing analysis." *Proceedings 2002 Design Automation Conference. 2002, pp.556-61. New York, NY, USA.*

[28] Poppe W, Capodieci L, and Neureuther A. "Platform for Collaborative DFM." *SPIE, vol. 6156, no.13, 2006.*

[29] Balasinski A. "A methodology to analyze circuit impact of process-related MOSFET geometry." *SPIE, vol.5379, no.1, 2004, pp.85-92. USA*

[30] Y. Trouiller, et al "65-nm gate OPC optimization with simple electrical model simulation" *SPIE, vol.5756, no.38, 2005*

[31] United States Patent 6562638

[32] Axelrad V, Shibkov A, Hill G, Hung-Jen Lin, Tabery C, White D, Boksha V, Thilmany R. "A novel design-process optimization technique based on self-consistent electrical performance evaluation." *SPIE, vol.5756, no.1, 2005, pp.426-33. USA.*

[33] Shiying Xiong, Bokor J, Qi Xiang, Fisher P, Dudley I, Paula Rao, Haihong Wang, En B. "Is gate line edge roughness a first-order issue in affecting the performance of deep sub-micro bulk MOSFET devices?" *IEEE Transactions on Semiconductor Manufacturing, vol.17, no.3, Aug. 2004, pp.357-61.*

[34] Shiying Xiong, Bokor J. "A simulation study of gate line edge roughness effects on doping profiles of short-channel MOSFET devices." *IEEE Transactions on Electron Devices, vol.51, no.2, Feb. 2004, pp.228-32.*

[35] "BSIMSOI 3.1 MOSFET MODEL User's Manual", http://www-device.eecs.berkeley.edu/~bsimsoi/

[36] Akers LA, Sugino M, Ford JM. "Characterization of the inverse-narrow-width effect." IEEE Transactions on Electron Devices, vol.ED-34, no.12, pt.1, Dec. 1987, pp. 2476-84. USA.

[37] Shah SS, Gupta P, Kahng AB, Sylvester DM. " Modeling of non-uniform device geometries for post-lithography circuit analysis." *SPIE, vol. 6156, no. 54, 2006.*

[38] Wang CT. "Three-dimensional threshold voltage expressions for both the LOCOS and deep trench isolated MOSFETs." IEEE Comput. Soc. Press. 1985, pp. 283-5. Washington, DC, USA.

[39] K.K-L. Hsueh, J. J. Sanchez, T. A. Demassa, and L. A. Akers, *"Inverse-Narrow-Width Effects and Small-Geometry MOSFET Threshold Voltage Model*", IEEE transactions on Electron Devices, Vol. 35, No. 3, March, 1988, pp. 325-338

[40] ITRS 2006

[41] Pawloski AR, Acheta A, Lalovic I, La Fontaine BM, Levinson HJ. "Characterization of line-edge roughness in photoresist using an image fading technique,"SPIE, vol.5376, no.1, 2004, pp. 414-25. USA.

[42] Finders J, Dusa M. "Matching multiple-feature CD response from exposure tools: analysis of error sources with their impact in low-k1 regime." SPIE, vol.5754, no.1, 2004, pp. 164-76. USA.

[43] Zhang G, Terry M, O'Brien S, Soper R, Mason M, Won Kim, Changan Wang, Hansen S, Lee J, Ganeshan J. "65nm node gate pattern using attenuated phase shift mask with off-axis illumination and sub-resolution assist features." SPIE, vol.5754, no.1, 2004, pp. 760-72. USA.

[44] Cain JP, Spanos CJ. "Electrical linewidth metrology for systematic CD variation characterization and causal analysis." SPIE, vol.5038, 2003, pp. 350-61. USA.

[45] Friedberg P, Cheung w., Spanos C. "Spatial modeling of micron-scale gate length variation" Proc. SPIE 6155, p.61550C (2006). USA

[46] Dusa M, Arnold B, Fumar-Pici A. "Prospects and initial exploratory results for double exposure/double pitch technique." IEEE-ISSM. 2005, pp. 177-80. NJ, USA.

[47] Robins GC, Dusa M, Kye J, Neureuther A. “Interferometric-probe aberration monitor performance in the production environment.” SPIE, vol.5377, no.1, 2004, pp. 1971. USA.

[48] Juliet ring defocus

[49] Cain JP, Spanos CJ. “Electrical linewidth metrology for systematic CD variation characterization and causal analysis.” SPIE, vol.5038, 2003, pp. 350-61. USA.

[50] Stonebraker M, Wong E, Kreps P, Held G. The design and implementation of INGRES. ACM Transactions on Database Systems, vol.1, no.3, Sept. 1976, pp. 189-222. USA.

[51] DeWitt DJ, Katz RH, Olken F, Shapiro LD, Stonebraker MR, Wood D. Implementation techniques for main memory database systems. SIGMOD Record, vol.14, no.2, June 1984, pp. 1-8. USA.

[52] Pawloski AR, Acheta A, Lalovic I, La Fontaine BM, Levinson HJ. “Characterization of line-edge roughness in photoresist using an image fading technique,”SPIE, vol.5376, no.1, 2004, pp. 414-25. USA.

[53] ITRS 2003

[54] Paul Friedberg’s thesis 2007

[55] Kawachi T, Fudo H, Iwata Y, Matsumoto S, Sasazawa H, Mori T. “Highly sensitive focus monitoring on production wafer by scatterometry measurements for 90/65-nm node devices.” IEEE, vol.20, no.3, Aug. 2007, pp. 222-31. Publisher: IEEE,USA.

[56] Ina H, Oishi S, Sentoku K. “Focus and dose measurement method in volume production.” Japanese Journal of Applied Physics, vol.44, no.7B, July 2005, pp. 5520-5.

[57] P. Niedermaier and T. Roessler , “Structural failure prediction using simplified lithography simulation models” SPIE 6521, 652107 (2007)

[58] Izuha K, Asano M, Fujisawa T, Inoue S. “Novel in-situ focus monitor technology in attenuated PSM under actual illumination condition.” SPIE, vol.5040, no.1, 2003, pp. 590-9. USA.

[59] Pang LT, Nikolic B, “Measurements and Analysis of Process Variability in 90nm CMOS.” ICSIC, 2006

[60] K. Gonzalez-Valentin “Sources Due to Layout Practices,” M.S. Thesis, MIT (2002).