A Dynamic Back End of the Line Customization Technique for Yield Improvement

Ardavan Aryanpour

A Thesis

In

The Department

of

Electrical & Computer Engineering

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Applied Science (Electrical Engineering) at

Concordia University

Montreal, Quebec, Canada

July 2010

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Ardavan Aryanpour

Entitled: A Dynamic Back End of the Line Customization Technique for Yield Improvement

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Sheldon S.Williamson Chair

Dr. Terry Fancott Examiner

Dr. Shah M.Jahinuzzaman Examiner

Dr.Glenn E.R Cowan Supervisor

Approved by _____
Chair of Department or Graduate Program Director

_____
Dean of Faculty

Date _____

Abstract

A Dynamic Back End of the Line Customization Technique for Yield Improvement

Ardavan Aryanpour, M. Sc

July 2010

As CMOS technology evolves and transistors get smaller, although chip manufacturers benefit significantly from being able to fit more transistors in a smaller area and also producing chips with lower power dissipation, they have to confront newer problems that are directly related to the size of transistors and the thickness of the deposited layers on a wafer.

Smaller transistors are faster and dissipate less power, but the smaller the technology becomes, the harder the fabrication process is to control. Thin silicon, metal and oxide layers must be accurately deposited because any variation in the thickness will cause unexpected behavior in the device.

These variations affect many parameters in CMOS. Any slight change in temperature, doping density, deposition timing, etc., can cause a significant change of characteristics of a CMOS device and the variation caused by these changes is called Process Variation (PV).

In this thesis, two circuits are taken into study in order to understand how process variation impacts the electrical specifications of a circuit example. The first example is a tapered buffer chain and the second example is a sense-amplifier flip flop. The idea is to propose a technique to decrease the loss percentage (Increase the yield). Basically for one specific design a few variant circuit layouts with different power-speed specifications are implemented and based on the results of the mid fabrication measurements on the test circuits that are deposited throughout the wafer, one of them is chosen with the means of choosing a proper masking sequence. The electrical characteristics of the

test circuits are correlated with the devices in the main circuits inside the chip. The alternative masking arrangement with the same sequence but different blocking masks will give a new design with different electrical parameters to correct the unwanted changes caused by PV. If the results of the on-chip test circuit do not meet the predefined specifications, the masking sequence is changed in order to choose another implemented design that we know will perform as desired. These two designs are simulated in CMOS 90nm technology and for each, both delay and power dissipation specifications are applied.

Simulation results showed 20.8% yield improvement for buffer chains and 19.6% for Flip-Flop's when both delay and power dissipation specification were applied. A comparison between the proposed technique and other current techniques shows a higher possibility to dynamically decrease the number of the chips that would not meet the predefined specification after they were manufactured. As opposed to other static techniques such as binning the products by their supply voltage and their delay, the proposed technique can offer an opportunity to save those products that could not be utilized when both delay and power dissipation specifications were applied during the manufacturing process.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Glenn E.R. Cowan for introducing me into the fascinating world of nanometer technology and providing me with the opportunity to do research with him. His patience and also eagerness to enlighten my path through discovering the unknown is remarkable. Not only he has supervised and corrected me to understand the concepts of this thesis, but also has taught me how not to deviate from the main axis of the research's subject. I will always be grateful for that and it would have not succeeded in continuing this project without his supports, guidance and specially his super friendly and encouraging attitude.

Secondly I would like to thank Mr. Ted Obuchowicz for his help, support and his tutorials on understanding CMOS technology softwares and their applications. His efforts in providing us supporting documents and supplementary materials on Cadence Tools and VLSI design are greatly appreciated.

I would also like to specially thank Dr.Asim Al-Khalili for being such a wonderful professor and for his distinguished classes and lectures on VLSI design. His conduct completely changed my view on Electrical Engineering and I am grateful for that.

I would like to thank my brother Dr.Karan Aryanpour for his help to edit and modify this thesis in terms of arrangement and literacy.

Lastly, I wish to acknowledge Mr. Jaro Pristupa's kind cooperation in allowing me to use the University of Toronto's laboratories and computer facilities in order to proceed in my research. Without his assistance, further continuation of this project would have obviously not been possible.

Table of Contents

Chapter 4: Examples for Implementation of Mid-Fabrication Test Based Mask Selection Technique

List of Figures

List of Tables

Chapter 1- Introduction

1.1 Overview of Process Variation

In ASIC design, manufacturers confront challenges at both the macroscopic and microscopic levels. Here in this thesis, we study the impact of microscopic issues that cause variations during the fabrication process. Process variation is the result of uncontrolled changes in the physical conditions of the fabrication line that causes fluctuation in the electrical characteristics of the transistors and other devices on the surface of a silicon wafer, and this in turns eventually lead to variations in propagation delay, power dissipation, supply voltage drop, threshold voltage, etc.

Silicon chips consist of millions of transistors. Depending on the area of chip, a wafer approximately contains from 50 up to 100 chips (also named Dies) on its surface. They are cut and diced into individual chips once the fabrication sequence reaches to its final step. Normally wafers are inserted into a lot and each lot usually contains from 50 up to 200 wafers depending on the technology and the diameter of the wafer.

Process variation does not impact all the dies and wafers to the same extent. Variations that are detected between different lots are more significant than those in dies (on-chip process variations) on the same wafer of the same lot, however some slight change in any electrical parameter on a single wafer can cause some of the chips on that wafer to fail meeting the pre defined specifications and therefore they are discarded. Usually the larger the dies are the more variation in the propagation delay of the transistors is detected across the die [1].

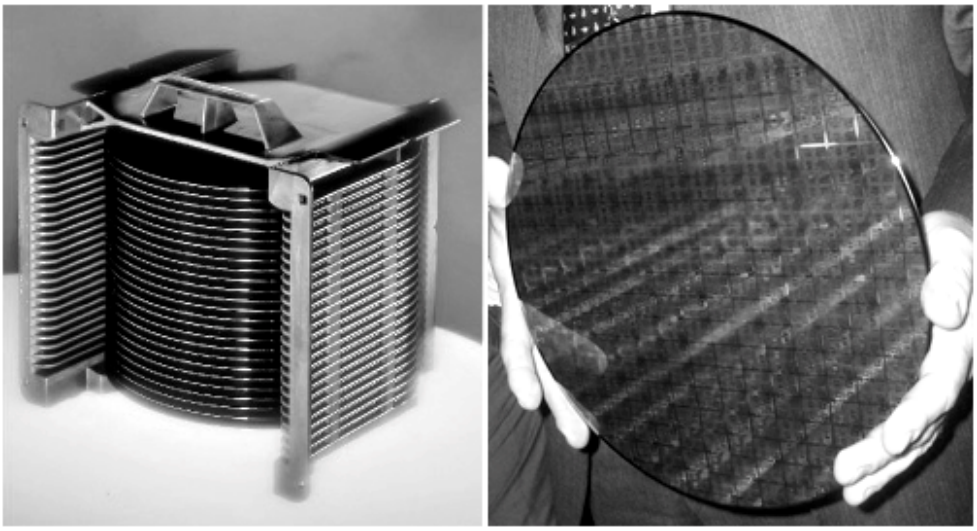

Figure 1-1: A fifty wafer lot and an AMD 45nm Wafer [2].

Process Variations can be classified in two main categories:

1. **Front-End of Line (FEOL) Variation:** These variations are caused by a change in the physical layout of the transistors in the circuit, meaning that the dimensions of the deposited and implemented regions and the oxide thickness are affected. Front-End variation occurs at the transistor level and eventually will change the behavior of the circuit. FEOL variations can be observed after the transistors are deposited on the wafer by measuring the corresponding device characteristics. For example, one of the most challenging types of FEOL variations is well proximity effects (WPE) that are caused by laterally scattered atoms that are embedded in the silicon near the N-Well and P-well. This causes variation in electrical characteristics of the MOSFET's. The distance of the scattered atoms to the wells varies and so do the MOSFET's electrical characteristics.

2. **Back-End of Line (BEOL) Variation:** Back-End variation occurs within the layers of metal and inter-layer dielectric

(ILD) for the interconnection of the devices. This type of variation could be detected in several points of the interconnection between different blocks and circuits throughout a chip. For example, metal thickness or width can have variations. Therefore variation in capacitance and resistance of all these interconnects can ultimately affect the performance of the circuit in terms of speed, bandwidth and power dissipation.

For each main category there are four possible sub-categories that process variation can be detected under:

1. **Within-Die Variation:** or Intra-die variations occur within one single die, meaning that the electrical parameters of a device that is used in different locations of the same chip vary due to the process variation. For example, transistor mismatch occurs within a die. Mismatch is caused by time-independent random variations in a physical quantity of identical devices. Transistor mismatch is caused by dopant fluctuations that change the threshold voltage and eventually the gain of the transistor for a specific threshold voltage. Figure 1-2 shows how drain current could vary for the same gate voltage in identical transistors.

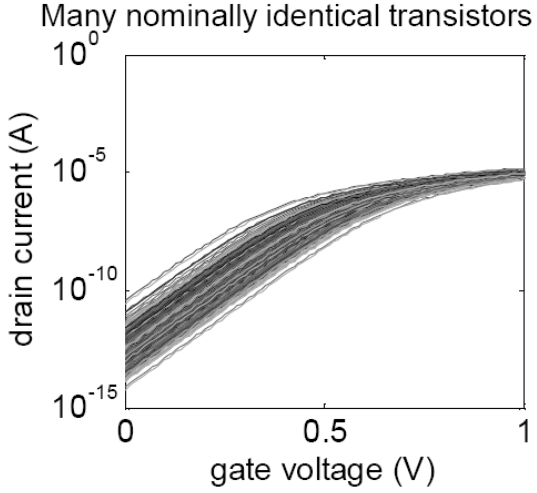

Figure 1-2: Drain current variations in identical transistors with different gate voltage [3].

2. **Die-to-Die Variation:** This variation is detected from one die to another. All the chips on a wafer have the same design; however process variation makes one single device have different electrical parameters in different dies throughout a wafer.
3. **Wafer-to-Wafer Variation:** This variation is detected from one wafer to another. All the wafers in a lot have the same design and are processed under same manufacturing sequence and physical conditions, yet variation in from one to another is detected.

4. **Lot-to-Lot Variation:** Each lot consisting of 50-200 wafers can also be subject to process variation. The entire set of wafers in a lot has similar electrical characteristics but variations are detected from one lot to another from time to time.

If all these four variations occur in the fabrication of a complete lot, meaning all sub level variations from a single die all the way back to the lot happen, we should detect a large magnitude of variations from lot to lot. Depending on the cause, source, and the level in which the process variation occur one can comment on the severity of the magnitude of the process variation on each level [4].

1.2 Problems Caused by Global Process Variation

Process variation can impact the oxide layer thickness, diffusion depth and impurity concentration. These all happen because of changes in temperature, pressure and the dopant concentration. Variation in oxide layer thickness and eventually the cross sectional area of the current path will cause a change in the sheet resistance and also makes the threshold voltage vary. Moreover, limitations in making high resolution lithographic masks can result in variations in the dimension (W/L) of the transistors. All of these changes can alter the electrical characteristics of the transistors, which in turn give rise to the variation in the propagation delay of logic circuits. Sometimes variations

occur by some random change of parameters during the fabrication process; therefore they cannot be classified as a repeatable pattern. They are called random process variations and the radius of these patterns is the same as each device's dimension on the chip which makes them vary individually and independently (could fall into mismatch category). As a result, the behavior of the circuit is affected and these variations can present significant changes in the circuit characteristics, such as propagation delay, bandwidth, power dissipation, etc. For many applications, some electrical specifications are defined for a chip and the chip must meet them in order to be properly utilized. Those chips that do not meet these specifications will fail to pass the quality control; therefore the yield of the production line will decrease. Today, some solutions are being used to decrease the impact of process variation on chips to increase the yield.

1.3 Overview of Fabrication Process and Current Static Solutions for Process Variation

The manufacturing sequence starts with "*fabricating*" the wafers (FAB). That is where the silicon layers form the circuit blocks and electronic devices on a wafer by doping impurities, emitting UV light and depositing layers. This part is one of the longest parts of fabrication. After circuit blocks are built and all layers are deposited on the surface of the wafer, they are sent to the second sequence called *"Die Level Cherry Pi cking"* (DLCP) and that is where all dies are characterized by the electrical specification such as power consumption, speed, etc., and then the wafer is sent to "*Assembly Die Inventory*" (ADI) which consists of the following steps. Wafers are taken to the "*Assembly Test Manufact uring*" (ATM) to be tested. This is where the dies are cut and separated from the wafer after they pass the testing procedure and if not, they are thrown out as defective chips, hence it is a very important sequence since depending on the behavior and the electrical characteristics of the chips they are classified in different categories. The testing machines will test the chips individually and then "Bin" them upon their performance. Binning chips is done when the test machine evaluates the normal operation speed of the product and the decision is made here if the chip is classified as a

fast chip or down-binned as a slow chip because of all the variations that might have occurred during the fabrication. Binning products can be a suitable solution for multipurpose chips or IC's that can be used in many different systems. Based on the requirement or the application of where they are used, they are selected according to their speed of operation, power dissipation, supply voltage, etc. Therefore some of the products are classified as slow and they can be sent to the market with a lower price and be used in circuits where high operation speed is not required. The same scenario is to be applied for more dissipative products and they can be used in applications where power is not an issue. Work voltage is another binning parameter. As we know increasing V_{DD} reduces the propagation delay. Now a chip used in a particular device that has strict specifications, binning could not be a suitable solution since as soon as the product is out of the specification range, it must be discarded. For example, if there are limitations for power dissipation of a circuit or supply voltage is limited due to using batteries, we have to find another alternative to reduce the number of discarded chips [5].

Problems that are associated with process variations become inescapable as the CMOS technology gets smaller and variations with more significant impacts on the performance of circuits appear in a more unpredictable way. Therefore migrating to smaller technologies requires understanding the source of these variations. It is obvious that non-estimated variations can increase the loss percentage, cost and time of fabrication processes. Soon controlling ion implantation is tightened down to counting atoms as the technology goes under 45nm, then controlled process methods or designing adaptive circuits are in need as dynamic solutions to reduce the impact of these variations today[6].

One of the techniques as a dynamic solution to overcome the well proximity effect is to use heuristics to conservatively guard-band devices. This technique offers a trial-and-error evaluation of layout modification until the SPICE simulations yield the desired performance of the circuit, and then the final layout is introduced to the manufacturing sequence. In this technique, the area of the circuit increases as the designer uses guard-band for the size of N and P wells and the area around them for the post layout simulations.

Figure 1-3 depicts the algorithm of the first technique to decrease the well proximity effect.

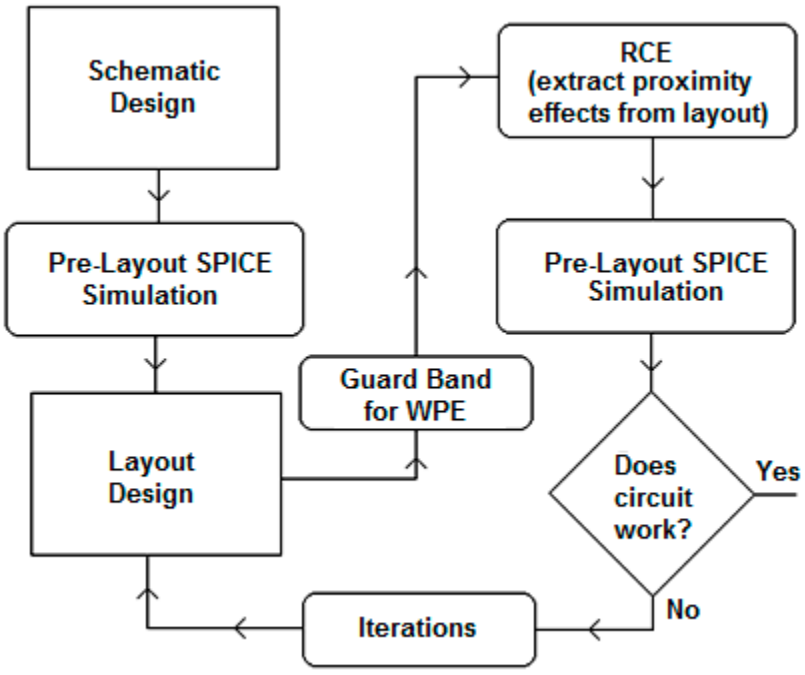

Figure 1-3: Current solution for well proximity is expensive in terms of time and area [7].

Another technique to decrease the impact of WPE is to extract netlists by running post-layout simulations and if there are WPE problems with the layout, one needs to redesign the layout again and do another simulation until the best layout configuration is achieved. Since so many iterations might be needed to be done in this technique, the lengthy iterative process time that the designer has to deal with maybe regarded as the disadvantage of this approach.

Static solutions such as binning are mainly focusing on categorizing the products for different applications with their own specifications and can be useful to increase the yield. However, binning cannot bring back the products into the specification range that is defined for a particular purpose. That means that if a certain set of electrical characteristics are mandatory for a chip and the chip does not meet the specifications due to some process variation, binning cannot help and the chip becomes useless.

1.4 Process Corners and the Proposed Dynamic Solution

The two major design parameters that are affected by the process variation are the speed and power dissipation. Increase in the leakage current that occurs due to variations in the thickness of silicon oxide layer causes an increase in power dissipation for design corners with low speed meaning although the frequency is low, the power dissipation is too large. Figure 1-4 shows a normal distribution of process corners for NMOS and PMOS transistors. The gray area is called *Spice Box* and it provides all possible values from S (slow) to F (fast) corners.

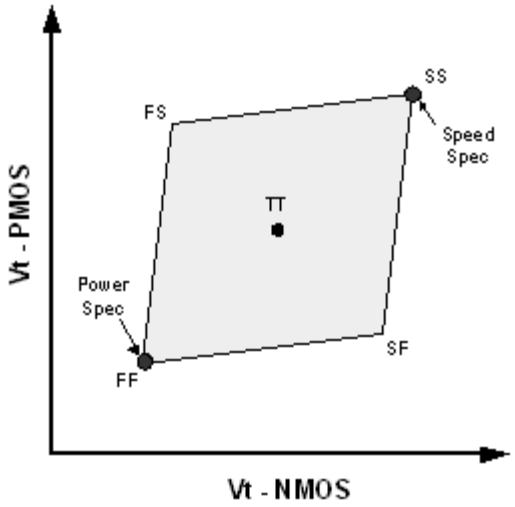

Figure 1-4 NMOS and PMOS Process corners, fast and slow distribution

The TT corner is the typical value of the threshold voltage for both NMOS and PMOS. FF corner corresponds to the manufactured chips with the highest speed and therefore highest power dissipation. It is the opposite for the chips that are designed for SS corner in which the power dissipation and the speed have their minimum values. At the design time, if the designers know how the process variation can change the process corner so that for example the circuit that is designed for FF corner will not meet the specifications, they can design another variant circuit associated with a different process corner (SF for example) that would meet the specs. This way a few variant circuit designs are implemented on the wafer and ready to be utilized upon the results of the test circuits. Figure 1-5 shows how the spice box can be divided into several regions for variant designs [8].

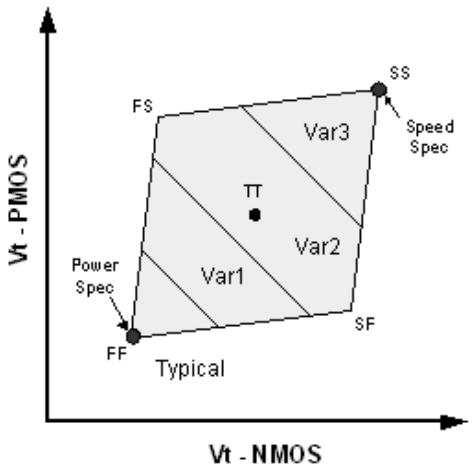

Figure 1-5 NMOS and PMOS Process corners, several regions for variant designs

In Figure 1-5 three variant designs are imagined for three process corners. If due to process variation, the Var3 that is designed for slow circuits becomes too slow and cannot pass the delay specification, Var2 is chosen to remedy the slowness of the corner. On-die mid-fabrication tests are already being done to determine many process parameters such as layer thickness and doping density while the layers are being deposited on the wafer. These measurements are made at various phases in IC fabrication processes. For example, it is common to measure transistor characteristics once the first level of metal (Metal 1) has been deposited. Normally, for cases in which the transistors are operating outside of some specification, the fabrication of that wafer may be stopped and the wafer discarded. The goal is to reduce the number of the discarded wafers in which the test circuit shows undesirable results by modifying the layout during the fabrication process.

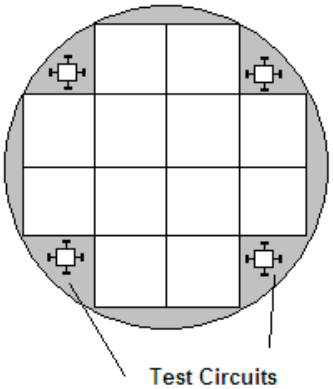

Figure 1-6: A wafer with 12 reticles and 4 test circuits in the corners

Figure 1-6 shows a wafer with 12 reticles on its surface. The darker areas on the wafer are used for the test circuit. For example, if our test circuit is a buffer chain with the same design that is used in each chip, after the Metal_1 has been deposited the pads of the test circuit are exposed to the test computer to make proper contact.

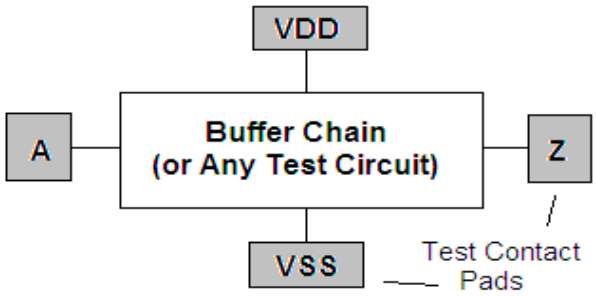

Figure 1-7: Test Circuit with 4 pads after Metal_1 deposition

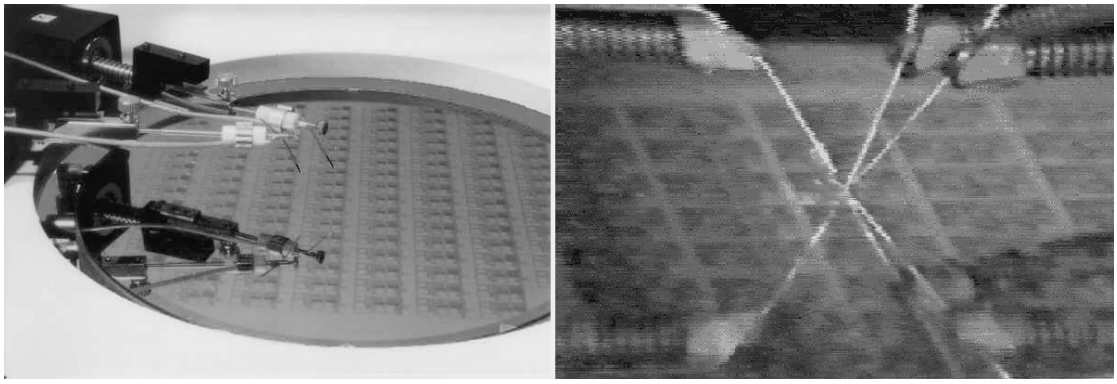

Figure 1-8: The chips are still on the wafer and are functionally tested using a test fixture with hundreds of needles that contact tiny metal pads on the surface of each chip. The probes send and measure signal responses from the chips. Chips that fail can sometimes be repaired; otherwise they are marked as failed and discarded after the dicing process [9].

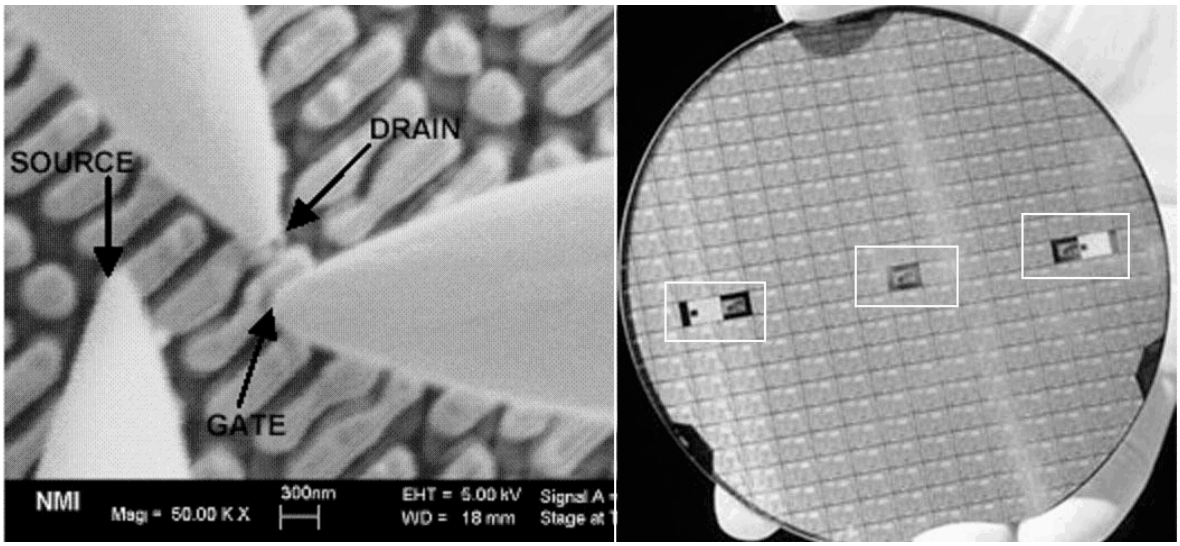

Figure 1-9: Test circuits on the surface of a wafer [9].

Now the test computer measures the delay between pads A to Z and based on the result one of the variant designs are selected. As was shown in Figure 1-5 at the design time several variant circuits are designed by optimizing the design over different process corners. Therefore a collection of different masks for one or more processing steps is generated to implement the designs. Based on these results, the masks corresponding to the most appropriate design optimization are selected and used by the operator for the rest of the fabrication. Ideally, in the proposed technique, several sets of masks are available to connect the proper design (faster) to the output. Thus, after an unacceptable delay is measured, the alternative masking sequence will allow us to choose the faster design that normally would be bypassed because of its higher power dissipation. Now that the wafer is about to be discarded, having a lower delay at the cost of higher power dissipation by means of connecting the reserved circuits to the output is justified.

This procedure is provided if the process variations only occur from wafer-to-wafer, therefore, if the chips from the same wafer suffer from variations too (for example due to non-uniform depositions, Ion implantations or inaccuracy in L_{eff} and W_{eff}, etc…) the test circuits must be inserted between reticles to be able to detect the die-to-die variations in all locations throughout a wafer. The goal of the proposed customization is to enhance the yield while achieving maximum performance and minimum design margin.

In this thesis, the impact of the global process variation on the production yield of two different circuits will be studied. For each circuit one or a few variant designs are proposed to replace the default design when applying delay and power dissipation specifications limits the number of the products that pass them, due to the variance that occurs in their electrical characteristics that is generated by the global process variations. In both circuits the delay-power dissipation spread has a negative gradient indicating that a smaller delay is correlated with higher power dissipation. Hence when the delay of the circuit is not in the allowed range, dissipating more power will bring it back into the acceptable range.

For both circuits, both the default and the variant designs have the same delay vs. power dissipation trend. The variant design has higher average power dissipation with a lower average delay in both circuits; viz. the variant design is always faster and more dissipative than the default design for a given set of process parameters. If the process variation behavior patterns in a chip manufacturing machine are known and based on the results of mid-fabrication measurements it can be verified that the default design does not meet the delay specification, depending on how tight the specifications are, it could be replaced with an alternative design.

1.5 Contributions of the Thesis

As the CMOS technology gets smaller, the process variation grows. There are a number of techniques to control the process variations as they will be discussed in the following chapters. Here in this thesis, we are trying to show the advantages of modifying the layout of a design before the fabrication is finished. The decision whether the technique can be used or not will be based on the results of the measurements that are done after the Metal_1 layer has been deposited. Applying the proposed technique will bring back the electrical parameters of the design to the acceptable range that were predefined for the circuit and eventually increase the yield of the production line. In this thesis, two circuit examples are studied. The first example is a chain of buffers that is calculated and designed for four different per-stage gains. All the chains are connected to the similar load capacitance and the transistors are sized in order to have similar output waveforms. For each design 500 iterations of Monte-Carlo process variation simulation are run and the results are compared to investigate the impact of process variation on each chain. Then delay and power dissipation specifications are applied to calculate the yield of each design individually. The yield of the proposed technique is also calculated and compared with other yields and 20% of yield improvement is detected. To show that this technique could be applied to other circuits as well as buffers, a sense-amplifier flip-flop (SAFF) is examined. One conventional SAFF consisting of NAND gates for the output driver and one Modified SAFF that is fast but more power-dissipative are taken and connected to the similar load capacitance. The same simulations are done for both circuits but this time 1000 iterations are simulated for SAFF. Delay and power dissipation specifications are introduced to both designs and the yield of each design is calculated. After applying the proposed technique the results show a 19.6% yield improvement.

This thesis was admitted for a poster presentation for the *IEEE Midwest Symposium on Circuits and Systems* (MWSCAS 2009) in Cancun, Mexico in August 2009 [11].

1.6 Outline of the Thesis

In *Chapter-2*, we begin with understanding the meaning of "yield" of a production line, a general outlook of how yield varies regarding the number of the products and what happens to this distribution once a set of products specifications is applied to the final products in order to control their quality. We will see that depending on the range of the specifications, the yield will decrease significantly once the specifications are applied and the number of the quality-controlled-passed-products drops.

In *Chapter-3*, we discuss the masks, their sequential implementation steps and how they are used to deposit different layers on the wafer. Then a possible implementation of extra masks is introduced to inspect how the proposed technique could be applied to the normal fabrication process in order to modify the design for a higher yield.

In *Chapter-4*, two circuit examples are taken into simulations. First a buffer chain constructed by inverters is shown. Then the way they are sized and also possible alteration in their design in terms of gain which would change the number of the buffers used in each is calculated and discussed. Then by using Cadence Tools suite (90nm STMicroelectronics), Monte Carlo simulations are performed for all the circuit examples, the results are compared and then the idea of mid-fabrication test based mask selection technique is applied to the simulations and the obtained new set of results is compared with all of the designs. The same sequence is done for a sense amplifier flip flop, and the results of each different design are again compared with the proposed design that is implemented by mid-fabrication test based mask selection technique.

Finally in *Chapter-5*, the thesis is summarized and an overview of possible future work and the requirements for further is provided in *Chapter-6*.

Chapter 2: The Production Line and Global Process Variation

2.1 Yield of the Production Line

In general, when a product is mass produced in a manufacturing company, the finished goods are taken into a quality check procedure to examine all their important specifications of the product and if these specifications meet the requirements that are defined by the application of the product before the product is ready to be sent to the market. When a product is manufactured in large numbers, due to the change in the physical parameters of the production environment and the machineries, some of the products will not meet the specifications and need to be discarded. The percentage of the products that pass the quality control section is call the *Yield* of the production and it can vary for different products.

Below, we assume that the product that is mass produced in a line is an IC with nominal delay and power dissipation and for this design, the faster the IC can operate, the more power is dissipated. If the quality control procedure defines two limits, one for the delay and the other for the power dissipation, we have criteria to filter those products that are not operating within this range and therefore, have to be discarded. This obviously introduces a financial loss percentage to the production of that specific IC.

Figure 2-1 shows a representative distribution for the delay of an IC's critical path taken from ICs over many wafers, spanning different lots of wafers. On the right hand side of the distribution are the ICs with a delay that is out of specification. On the left hand side of the distribution are the ICs that are very fast, but also consume excessive power causing them also to be outside of specification. The region in the middle contains the usable chips. Ideally there would not be sections of the distribution outside of this region.

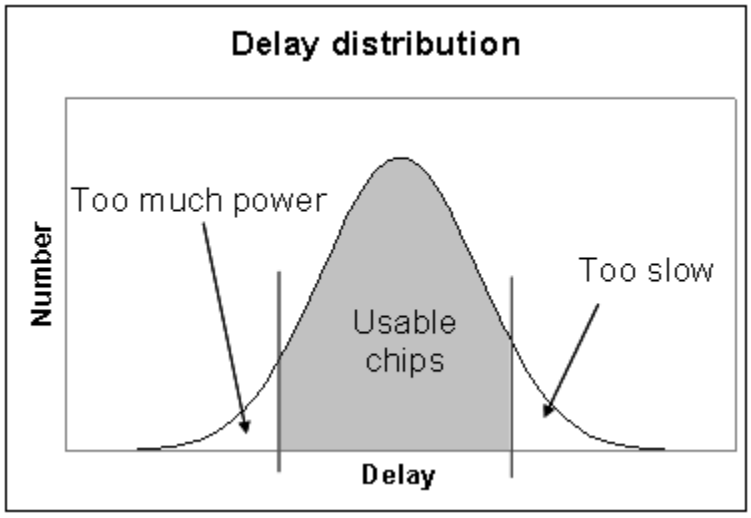

Figure 2-1: Critical path delay distribution showing ICs that do not meet cycle time specification or power dissipation specification

2.2 Products and Their Specifications

Ideally, in order to reduce the number of the chips that are outside the two vertical lines, we need to decrease the number of the slower and less dissipative chips while decreasing the number of very fast and more dissipative chips. With doing that we have increased the number of the IC's that are between the specification lines and as a result the yield of the production[12].

In Figure 2-2, the right half of the distribution corresponding to the slower-than-average chips has been shifted to the left, i.e., reducing the delay. Due to the same technique, the left hand half of the original distribution corresponds to the faster-than-average chips has been shifted to the right (i.e., increasing the delay). It is assumed that by slowing down the circuits from the faster wafers, those circuits also reduce their power dissipation. From Figure 2-2 it can be observed that a larger fraction of the chips is within the specifications.

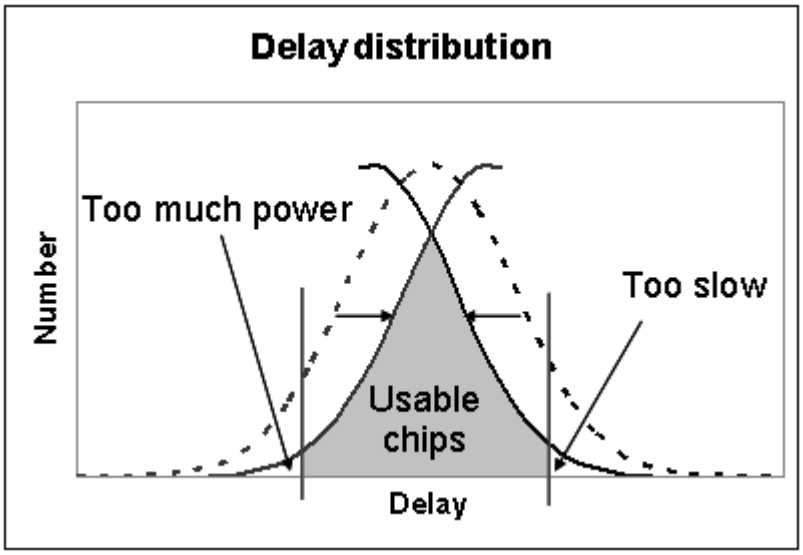

Figure 2-2: Critical path delay distribution showing ICs that fewer chips are out of specification

When specifications are applied to chips, depending on how tight the specifications are the number of chips that meet the specifications may vary. Efforts in manufacturing chips are focused on increasing the yield of the production line and finding solutions to avoid high loss percentages when specifications are applied.

Here, two specifications for electrical characteristics of the products which are delay and power dissipation are used to determine the production yield of the given circuit examples. The histogram in Figure 2-3 shows the distribution of number of products vs. their delay for an inverter chain. If specifications are applied to the samples in Figure 2-3, the qualified products will be limited to the left of the vertical line Delay Specification.

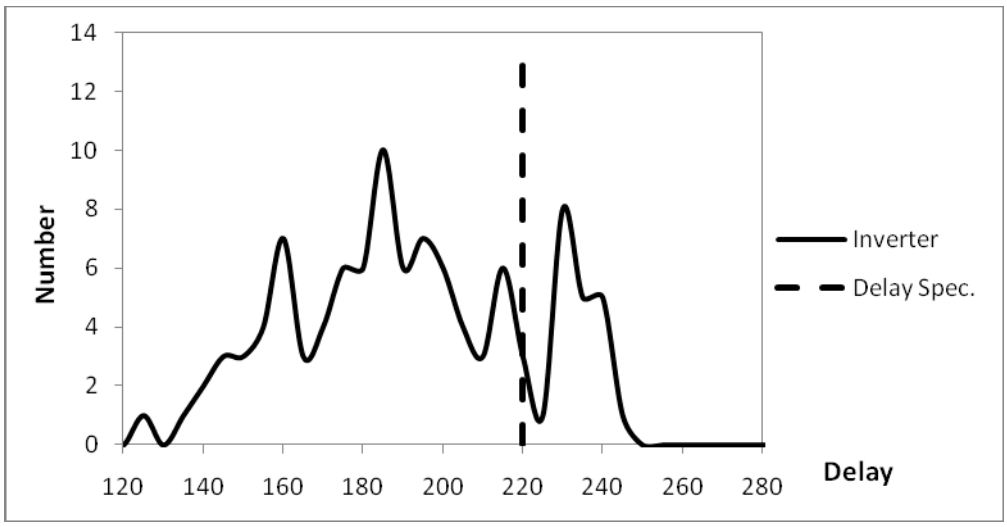

Figure 2-3: Histogram of the number of products vs. delay of an inverter chain, an assumed delay specification line is drawn

Those inverters to the right of the vertical line do not meet the delay specification and must be discarded. Since it is not possible to show how applying two specifications will limit the yield in a 2D graph, 3D graphs are used in the next chapters to show a more detailed explanation of how a 2-sepcification filtering affects the production yield. Generally the faster an inverter is, the higher its power dissipation will be. Therefore if a power dissipation specification is defined as well, it will limit the number of the products on the left side of the histogram which means a certain percentage of the fast inverters.

2.3 Yield Analysis and Statistics

In this thesis, the factor that changes the electrical parameters of the circuit examples and gives a range of number of productions vs. delay or power dissipation vs. delay is the *Monte Carlo Process Variation Si mulations* [12]. By simulating a circuit example for a given number of iterations between 500-1000, a data set of delay and power dissipation is obtained and then the specifications are applied to them to calculate the yield. The ratio of those samples that pass the specifications to the entire set will give a percentage that is equal to the yield of that data set.

The specifications that are used are hypothetical numbers within the range of delay and power dissipation of the circuit examples only to evaluate the applicability of our proposed technique and compare the results of the variant designs for each circuit example. For each circuit example in this thesis, there are one or more variant designs and the Monte Carlo process variation simulations are done for all of them individually. The results are compared and the hypothetical specifications are applied. After applying the specifications a percentage of the chips for each design fail to meet the specifications. This percentage shows how many chips cannot be used for the application for which the specifications were defined. The proposed technique offers a selective approach algorithm and this algorithm generates a new dataset of delay and power dissipation which again is taken into account for calculating the yield. The new dataset shows a higher yield and the number of the chips that fall within the specifications lines is larger.

Chapter 3: Introduction to Mid-Fabrication Test Based Mask Selection Technique

3.1 Mid-Fabrication Test Based Mask Selection from the Masks Perspective

It was briefly shown in earlier chapters that there are two keys to make the proposed technique applicable. First we have to be able to design a few modified variant circuits based on having different process corners. The designer has to know that if because of the process variation one of the process corners (for example the least dissipative corner) cannot provide a high yield for that specific design, there is another variant process corner that most likely can improve the yield to a higher value. The second key is implementing test circuits on the wafer that have the same power-delay behavior as the main circuit so that when their electrical parameters are measured, the results can relatively provide the facts the designer needs for the main circuit in order to make a decision which process corner to choose.

In one implementation of the technique, the circuit is optimized at different process corners. Each wafer is processed to the earliest point at which its process corner can be determined. This is typically after Metal_1 has been deposited which means taking the wafer into a set of test procedures. Assuming that measurement can only be done when Metal_1 has been deposited, different Via1 (the layer that connects Metal_1 to Metal_2) masks are designed for each of the above design optimizations. After process measurement the appropriate mask is selected for Via1 with all subsequent masks being the same for all process corners. The reason we intend to change Via1 mask only is that it is necessary to have Metal_1 layer deposited so that the test circuit pads are exposed to the testing probes. Due to the cost of masks it is best to limit the number of masks that are required for each different optimization. Ideally, there is only one main Via1 mask with a Via1 arrays on it and a set blocking masks with possibly lower lithographic resolution to block the unwanted vias on the Via1 mask to achieve different layouts.

Since Metal_1 layer has already been deposited on the substrate to run the circuit tests and verify if the sample meets the requirements, then it is not possible to reverse the process and delete the deposited Metal_1 layers. However it is possible to have different masks for Metal_2 and have the two paths deposited and etched beside each other, and then choose between the two paths by connecting proper inputs and outputs with the means of utilizing the proper Via1 mask that will eventually connect the preferred Metal_1 path to Metal_2. For example, simply imagine two inverters with different sets of N and P sets of transistors in terms of size with path1 and path2 deposited by Metal_1 as their inputs. If only one of them is to be connected to the signal path, the proper Via1 mask is chosen to connect the input of the preferred inverter to the signal and the other inverter is bypassed by grounding its input to VSS.

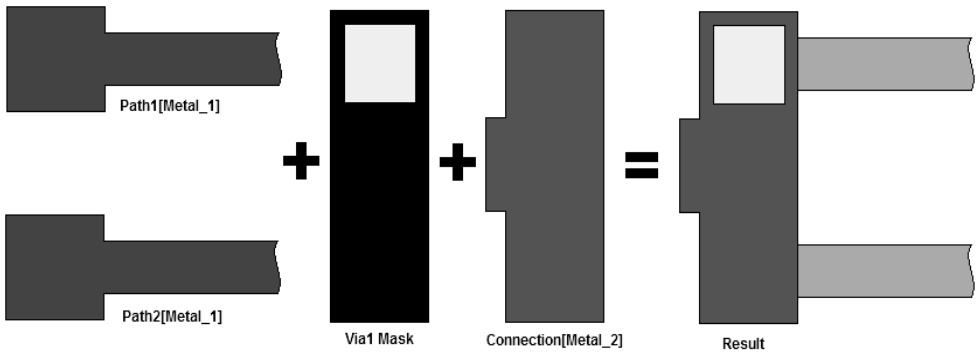

Figure 3-1: Using Via1 mask to connect Metal_1 and Metal_2, Path2 is excluded

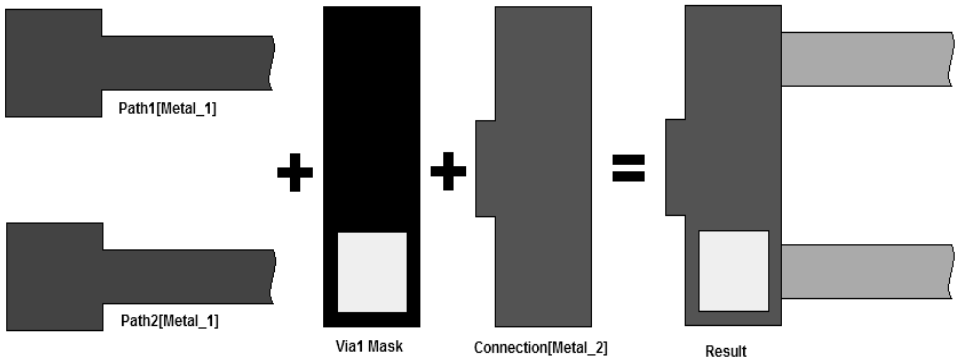

Figure 3-2: Using Via1 mask to connect Metal_1 and Metal_2, Path1 is excluded

Figure 3-1 and Figure 3-2 depict how different layouts of Via1 masks can change the signal path. Generally via masks are expensive due to the need of their lithographic accuracy. Vias have the smallest areas on the wafer and designing multiple via masks can make the cost of the proposed technique unreasonable. In order to decrease the cost of designing another via mask, we can use Blocking Masks that will not require the same precision and are easier to design. By blocking the unwanted Via1 location on the Via1 mask, we can implement the desired wiring between the circuit blocks. Figure 3-3 depicts how a via mask is blocked by a blocking mask.

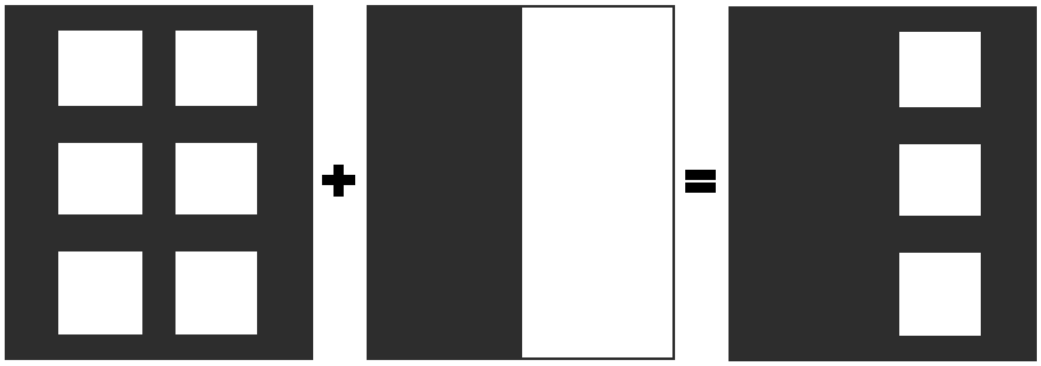

Figure 3-3: Left: Via1 Mask, Middle: Blocking Mask, Right: New Via1 Mask

This technique is applicable when two or more devices are selectively chosen to build a signal path. The same technique could be used for one device to change its electrical characteristics. For the same given inverter example, the designer can add (remove) fingers to increase/decrease the W_{eff} of the P and N transistors and consequently the gain and the propagation delay of the inverter. This decision can be innovatively made by the designer based on the results of the simulations and also the predefined power-delay specification constraints. Figure 3-4-a depicts how having a different number of fingers in a transistor can change the effective area. Different Via1 mask configuration allows us to connect and disconnect circuits from one to another. Figure 3-4-b depicts two different possible configurations of vias and how the effective width can change by choosing via masks.

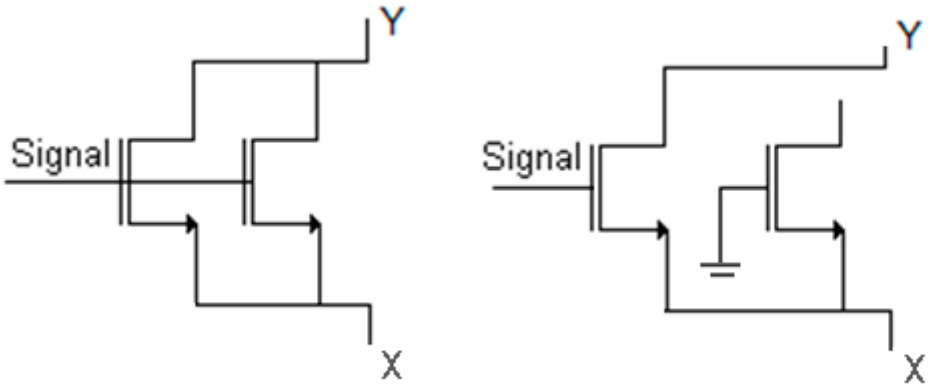

Figure 3-4-a: Different number of fingers in the schematic view, Gates must be connected to the ground to assure the transistor remains off, Drains can stay floating

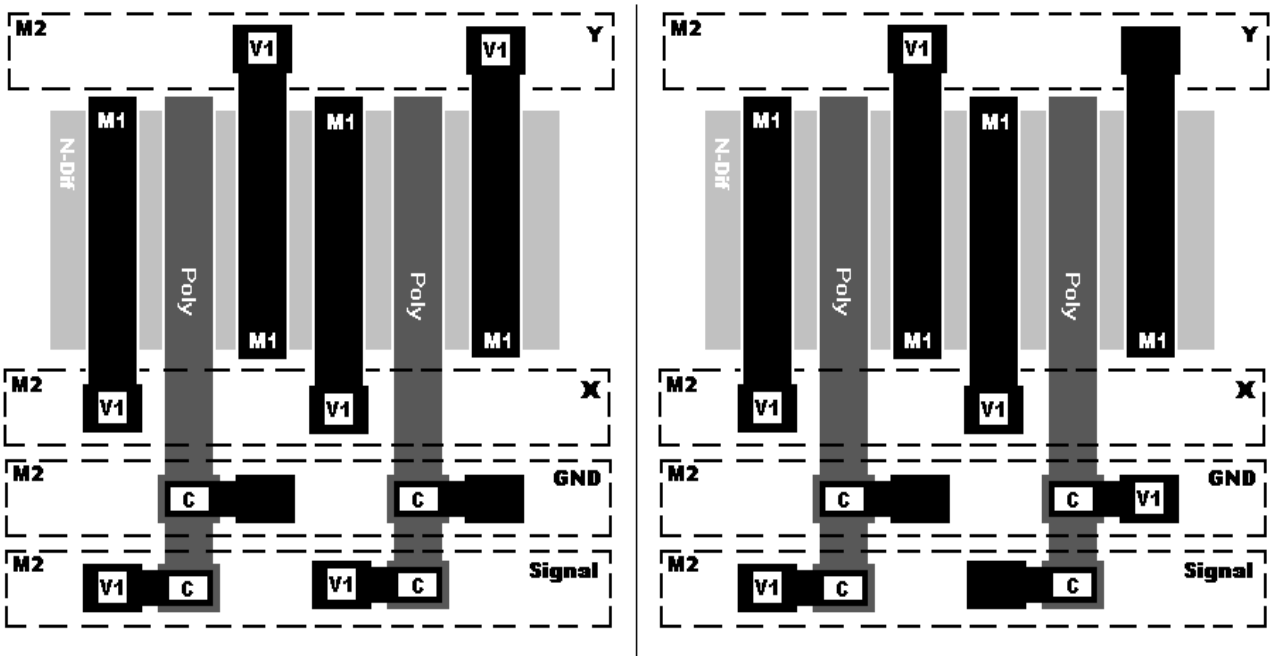

Figure 3-4-b: Different number of fingers for the same N-well will change the W_{eff}, poly's are selectively connected to the desired points by choosing the correct Via1 mask

3.2 Masking Mechanism

As mentioned before, the masking technique will allow us to select which design to use during the fabrication process based on the decision that is made

by analyzing the results of a test circuit that is mounted on the wafer for this specific purpose. In order to be able to relate the impacts of process variation on test circuits to main circuits, the test circuit must have the similar delay-power behavior to its counterpart in the main circuit when the process variation occurs. Test circuits have large pads and measurements are done by contacting the test probes to these pads after the deposition of the Metal_1 layer is completed.

In the automatic masking machine, all the masks are lined up in a vertical rack and the programmed computer will run a normal sequence to deposit the desired layout on the wafers until the entire lot is finished. Here, to apply the proposed technique, the normal sequence is followed by the computer up to Metal_1 layer and then once the deposition of Metal_1 layer is finished, the wafer is taken for measurements. As was discussed in the previous chapter stopping the deposition sequence for measurements is regularly done for quality control purposes. Therefore the proposed technique is not forcing extra steps to the manufacturing process.

Figure 3-5 depicts the fabrication process in more detail. Characterization of the test circuitry following the deposition of Metal_1 gives insight into the wafer's or reticle's process parameters, which in turn are utilized to select the optimal variant design. The inter-space area between two chips is called *kerf* area that can be used for implementing test circuits. First transistors in the kerf area test circuitry and the variant designs (A and B) within the final product circuitry are built. Based on the decisions that are made by testing the wafer, only Circuit A or B is to be chosen and connected to the final output. Variant designs are denoted by Circuit A and B. For example, circuit B is the selected variant design; therefore connections between Circuit B and the output are made through the Via1 layer. Subsequent processing steps are identical for all variant designs.

As can be seen in Figure 3-5, the layout of Metal_2 (Metal_2 mask) layer is designed to be able to make contacts with both variant designs once the measurements are done.

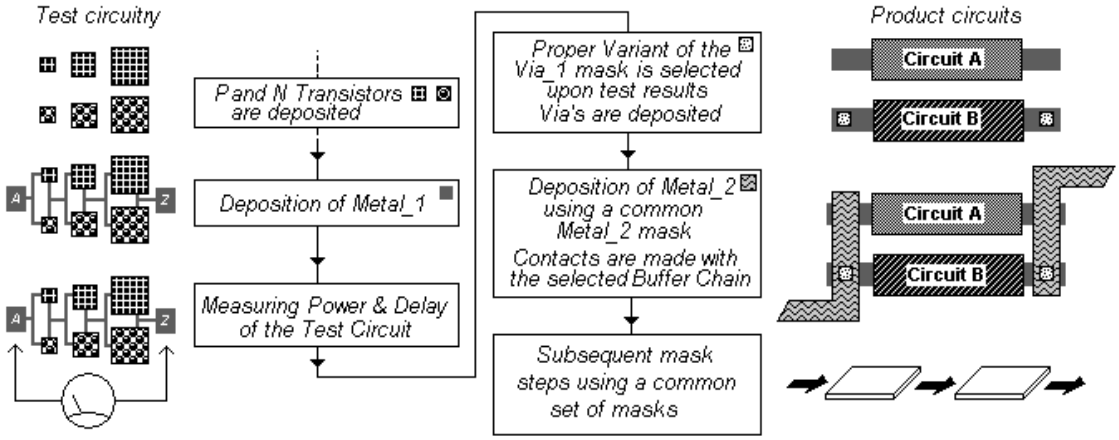

Figure 3-5: Flow chart of the proposed processing steps.

By assuming that the process variation patterns of the fabrication machine are known by previous measurements, a proper sequence can be programmed in the computer. If the results show uniform variations for all the wafers in a lot, and these variations are always following the same pattern, the computer easily selects the same masking sequence for all of the wafers in the lot. This means the operator knows that it is sufficient to only run the test procedure for one wafer and apply the proper masking sequence for all of the wafers in that lot. If this is not the case and the results vary from wafer-to-wafer, then each wafer must pass the test steps and the desired masks are chosen for them. In case of variations from die-to-die or on-chip variation (OCV), the masking technique will become complicated. This means each chip must be tested and depending on the results, and several test circuits are mounted on different coordinates on the surface of the wafer. Then based on the location in which the highest variation occurs, a new masking sequence is chosen by the operator and programmed into the computer or done manually. In a normal masking procedure the computer that runs the photolithography knows how many dies are located on the wafer and the dimensions of each are programmed in it. When a mask is picked by the machine, the pattern of the mask is applied to each die located on the wafer separately. Figure 3-6 illustrates a normal masking procedure on a wafer [14].

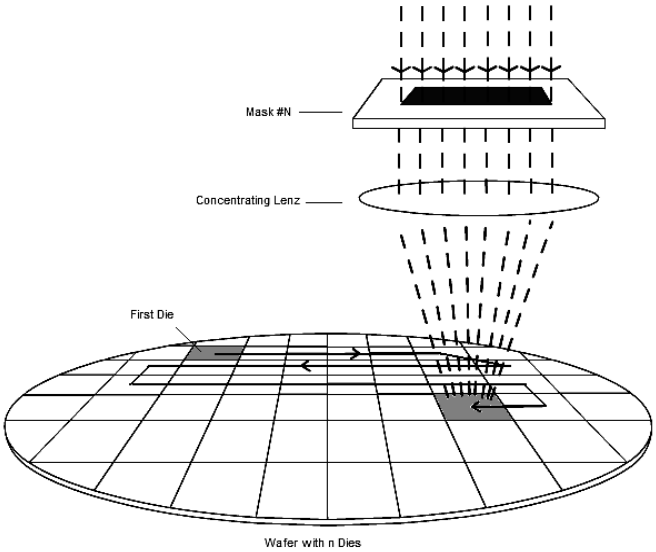

Figure 3-6: Normal masking procedure

When all dies are exposed to only one layout pattern, the lithography computer starts exposing the pattern to the dies one after the other towards the end of the wafer and then picks up another wafer and continues until a lot is fully finished. Now if different masks must be used for one wafer, the chips that have to be exposed to a different mask must be programmed in the computer so that the lithography computer locates the right chips by having their coordinates and exposes them to the new mask pattern[15]. It is shown in Figure 3-7.

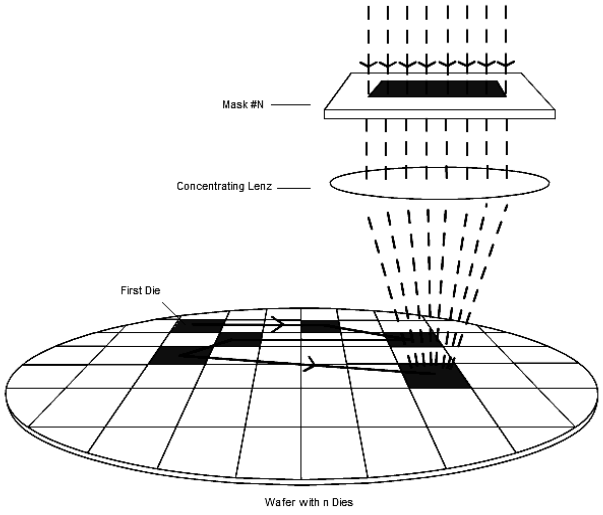

Figure 3-7: Exposing Slower Chips to Alternative Pattern

3.3 The Effect of Process Variation on Metal_1 layer

Since the results of any simulation in this thesis are obtained by designing and simulating the circuit in the schematic view, it is useful to inspect if the Monte Carlo Process Variation Simulations affect the sheet resistance or the Metal_1 layer characteristics. In order to evaluate any change in Metal_1 layer's electrical characteristics, a thin long wire of Metal_1 was taken to run a simple test and measure the process variation impacts on it. To make the proper contact in the layout mode, a resistance (G3≈2Ω in Figure 3-8) was used to avoid short circuit in the simulation. A 1pF capacitor was connected to point "B" and 1ns long pulse with 50% duty cycle was applied to point "A" and 1000 Monte Carlo simulations were run to measure the impact of process variation on the current and the delay between point "A" and point "B". The PLS strategy is set to *Single Extraction* and we want to know if the sheet resistance shows any variation in this layout extraction method. If variations in the sheet resistance are detected then we expect to see variations in the current and the delay between the injected pulse and the respond.

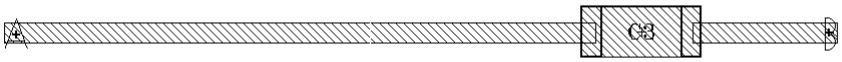

Figure 3-8: A narrow Matel_1 layer resistance sheet connected to a resistor

After simulations were finished, since no significant change in current and delay was detected, the area of the wire was increased by 132 times to inspect if a large area of metal sheet can possibly introduce more variations and then

run the simulations again. Figure 3-9 illustrates the new shape of the Metal_1 layer. The new sheet had an area 133 times larger than the first sheet; however the results still did not show any significant change in the variance of the current or the delay between the input pulse and the response, except that the average delay was smaller and the current larger.

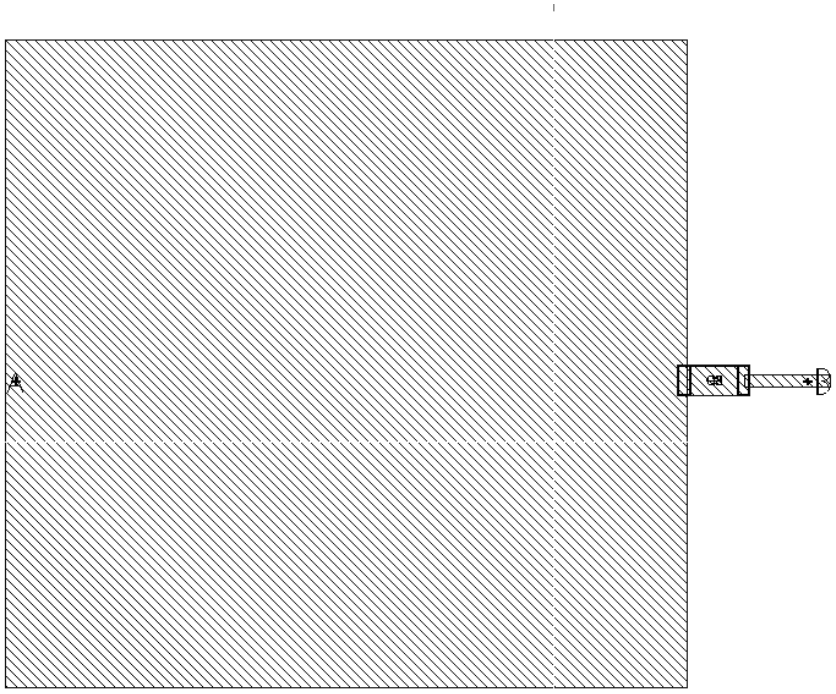

Figure 3-9: A wide Matel_1 layer resistance sheet connected to a resistor

Table 3-1 presents the results for 1000 Monte Carlo simulations:

Area of Wire	Load	Avg. Delay	Delay Std. Dev.%	Current Std. Dev.%
7µm x 50 nm	*1pF*	*8.0416ps*	*0.08%*	*≈0*
7µm x 6.62 µm	*1pF*	*0.9862ps*	*0.02%*	*≈0*

Table 3-1: The results of 1000 Monte Carlo process variation on a Metal_1 sheet

The results shown in Table 3-1 indicate that the deviations in the delay and the current of the circuit above are insignificant. Hence simulating circuits in schematic view where the wires are ideal and the sheet resistance variations in the design layout are neglected should have the same results as the layout

mode (different layout extraction methods such as Max., Typical and Min. will yield a variety of process variations). This will simplify finding alternative designs for any circuit example and as a result, only arrangement of transistors and their sizes will be taken into account. Although the device characteristics can change by BEOL variations, here it is shown that these variations are not accounted for in the Monte Carlo simulations. One should know that the schematic view does not always yield accurate results as the layout plays a significant role in providing the facts about the circuit when it comes to simulating it.

3.4 Layout Alteration

Although all the results and calculations are done for only the schematic view of the circuit examples in the following chapters, one possible layout alteration of two buffer chains with different per-stage gain was implemented in order to study the applicability of the mid-fabrication test based mask selection technique by modifying the Via1 mask.

In Figure 3-10 a possible layout of two buffer chains is shown and each buffer chain is circled with a dashed line. It will be demonstrated how the inverters are sized and aligned with each other. The layout in Figure 3-10 was simulated twice and each time one of the chains was bypassed by grounding its input and output by connecting them to VSS.

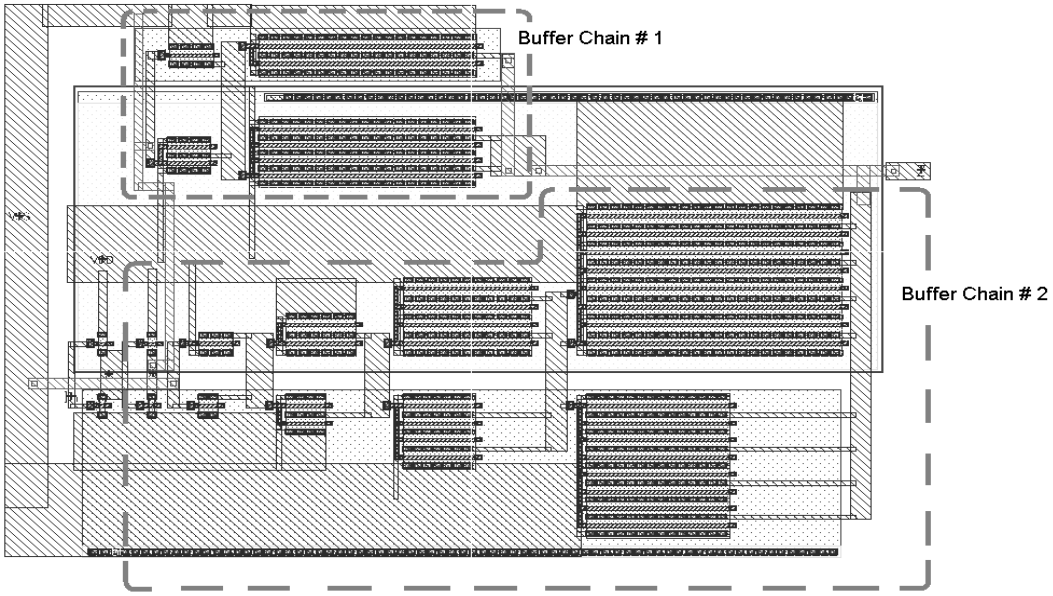

Figure 3-10: Layout view of two buffer chains with different per-stage gains

Figure 3-11 illustrates how buffer chain # 2 is bypassed by using vias to connect its input to VSS.

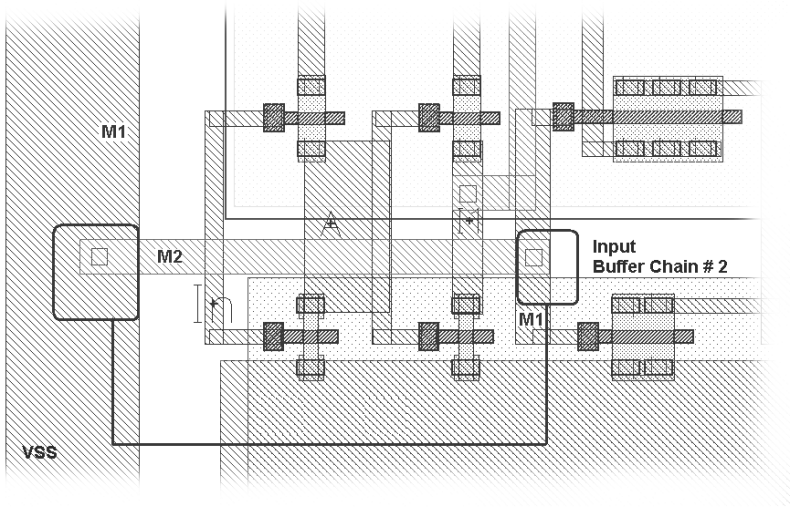

Figure 3-11: Connecting input of buffer chain # 2 to VSS, buffer chain # 1 is chosen

Vias that are circled in Figure 3-11 and Figure 3-12, connect Metal_1 layer to Metal_2 layer. Figure 3-12 shows how buffer chain # 1 is connected to Z (output pin) by using vias.

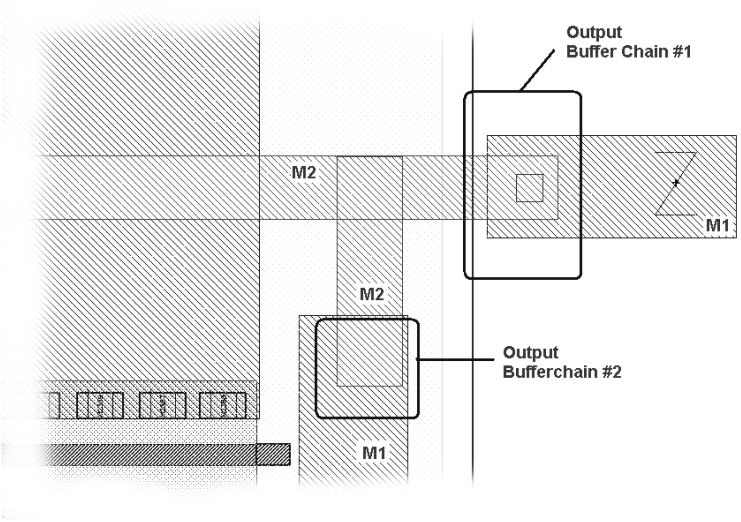

Figure 3-12: Connecting output of buffer chain # 1 to Z, buffer chain # 2 is bypassed

3.5 Limitations of the Proposed Technique

One limitation is how to estimate the cost of making extra masks and also the extra time duration of the mask changing sequence that will be added to the overall manufacturing time. This is possible when the process variation patterns are always the same and the delay and power dissipation variances are constant. Then the extra cost could always save a fixed percentage of the non-qualified products. However if for example one of the process corners in which the chip's specifications are within the acceptable range randomly repeats itself more than 100 times, then the number of wafers that do not meet the specifications is not high enough to justify the cost of making the extra masks. In such a case the loss percentage drops and the yield improves without applying the proposed technique and the cost of extra masks will not be economically reasonable. In other words, when a circuit example is simulated in the Monte Carlo Process Variation Simulator, it gives a data set of delay and power dissipation for the given number of iterations (i.e. 1000). Each sample of the dataset has unique values for the delay and power dissipation and none of iterations is duplicated in the dataset. However, in reality, if the physical conditions of the manufacturing process stay constant

for example for two different lots, then we may have higher number of acceptable chips than expected during simulations. Moreover, if reticle-to-reticle variations occur, the time that is required to apply mid-fabrication test based mask selection technique to the reticles will significantly increase, and that needs to be evaluated if being economically beneficial.

Another limitation is the area of the reserved transistors for the alternative design. Regardless of using the reserved transistors or not, they have to be deposited on the substrate in order to be connected to the Metal_1 layer by vias if needed. This effectively increases the area of the chip and also limits one's freedom in designing the best circuit layouts for designers. The best design would be a combination of the typical circuit and the variants so that they share maximum possible parts.

It will be demonstrated in Section 5.1 that if the delay and power dissipation variations are large enough to enhance the loss percentage, then the cost of applying the mid-fabrication test based mask selection technique might be reasonable and this technique could be a proper dynamic solution to increase the production yield. The cost of the proposed technique needs be estimated by the manufacturing company. For that, the process variation parameters in a manufacturing line must be measured and registered so as to estimate the production loss and then justify the extra cost of the proposed technique. Since every design has its unique sensitivity to the process variation, these estimations could vary from design to design.

3.6 Frequency and Supply Voltage Binning

As it was discussed before, process variation can cause unwanted changes in the products specification such as power dissipation and frequency. When products are too slow or too dissipative, they cannot be sent to the market with their nominal specifications; therefore they must be set aside and discarded and the yield of the production will consequently plummet. One of the static solutions for yield improvement being used today is binning the products by their electrical characteristics, such as binning by supply voltage,

frequency, etc. , which amount to classifying the products by such electrical characteristics and the sending them to the market. When some of the products are classified as slow, they can be sent to the market at a lower price and be used in circuits where high operation speed is not required. As an example, *Intel* sells two CPU's that are identical in terms of architecture and technology at different prices because of their different clock frequencies (Quad Core Xeon 2nd Processor E5335, 2x4MB Cache, 2.0GHz 1333MHz FSB, PE1950 for $400.00 and Quad Core Xeon 2nd Processor X5355, 2x4MB Cache, 2.66GHz 1333MHz FSB, PE1950 for $1170.00). Classifying chips by their clock frequency and sending them to the market at different prices is called *Frequency Binning* [16]. In this case the product is considered *AS IS* and a higher frequency cannot be applied to the chip or the chip will not operate properly. The same scenario can be applied for more dissipative products and they can be used in circuits for which power is not an issue. If a dissipative chip is used in a circuit that cannot provide the power that is necessary to operate the chip, the circuit will fail to run.

Adjusting the supply voltage of the chip in order to bring the electrical characteristics back into their desired values is called *Voltage Binning* [17]. For slower chips it is possible to remedy the low clock frequency by raising the supply voltage V_{DD}. Voltage binning is also utilized for leaky or highly dissipative chips. Since both dynamic and static power dissipations are exponentially proportional to V_{DD}, lowering down the V_{DD} can decrease the overall power dissipation of the chip. After all the chips are tested, they are divided to several voltage groups (bins) with values such as 1V, 1.2V, etc... .

A buffer chain was taken as an example in order to investigate how increasing V_{DD} can affect the delay. For example, a square wave with T=1ns was fed to a buffer chain with the gain of 3.9 (G3.9) with C_L=300fF and then 3 different values of V_{DD} were applied for the circuit to study the impact of increasing supply voltage on the delay and also the power dissipation of the buffer chains under the same process variation conditions. Figure 3-13 below presents the result for the three sets of data each including 500 iterations of Monte Carlo simulations at V_{DD}=1V, 1.2V and 1.4V.

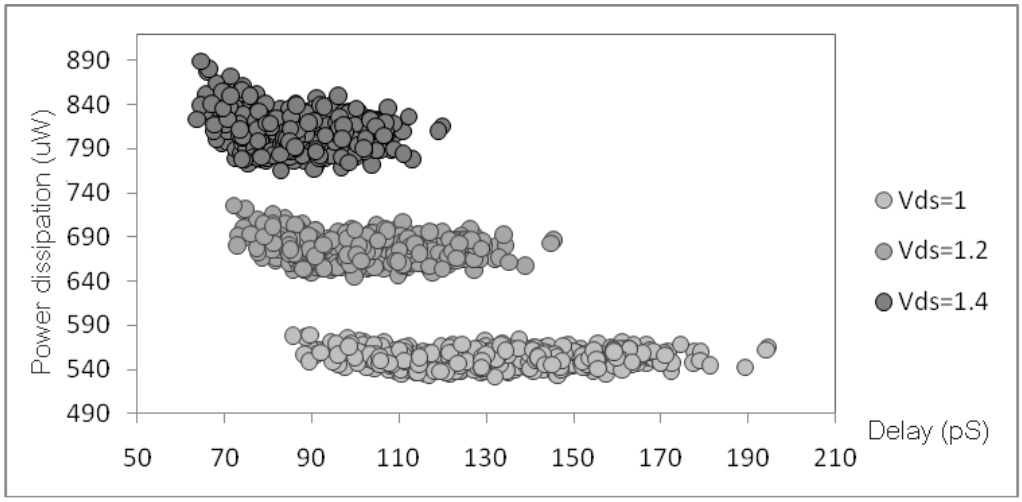

Figure 3-13: Delay vs. Power dissipation for a buffer chain with gain of 3.9 when three different V_{DS}=1V, 1.2V and 1.4 V were applied for each 500 Monte Carlo Simulations.

It illustrated in Figure 3-13, higher V_{DD} decreases not only the delay but also the delay variance percentage. However all of these are at the cost of a significant increase in the power dissipation of the buffer chain. According to Table 3-2, when there was a 20% increase in $V_{DD,}$ a 21.20% improvement in the speed of the buffer chain was detected, and the average current increased up to 22.30% from its initial value at V_{DD} =1V. The third row in the table also presents 34% improvement in speed and 46.2% increase in the average current for a 40% increase in $V_{DD.}$

Voltage(V)	Avg. Delay	Avg. Current(μA)	Max Delay	Max Current	Delay Decrease %	Current Increase %	Delay STD%
1V	131.8	552	194.5	579	---	---	16.05%
1.2V	103	675	145	725	**21.20%**	**22.30%**	**14.11%**
1.4V	87	807	120	889	**34.00%**	**46.20%**	**12.89%**

Table 3-2: Comparison between the results of a buffer chain with G3.9 when three different V_{DS}=1_V, 1.2_V and 1.4_V were applied for each 500 Monte Carlo Simulations,

Generally if voltage binned IC's are decided to be used, several adjustable voltage regulators are implemented in a circuit in order to provide different

supply voltages for different sections. This can be easily done for larger scale circuits where area is generally not a concern and higher supply voltage can be directed to those chips that are binned for higher voltages (PCB boards) and in doing so, one requires extra voltage regulators. Thus, in regard to all the above mentioned complications, our proposed technique becomes a considerable competitor to binning:

1. If only a few chips of the entire circuits that are binned for higher voltage are connected to a higher supply voltage, there will be voltage matching issues when different blocks with different supply voltage are electrically connected to each other, for example one block is biased with V1 =1V and another with V2=1.2V, therefore the output of the block with higher supply voltage will be higher and will not match the input of the next stage.

2. For some circuits, there might be limitations on having higher amounts of voltage, meaning if the battery cannot provide higher voltages, the voltage binning will be impossible.

3. It is known that voltage regulators dissipate a considerable amount of power ($P_D = I_{OUT} (V_{IN} - V_{OUT}) + V_{IN}I_{Q,}$ where I_Q is no-load current), therefore having more regulators amounts to dissipating more power while binned IC's with a higher supply voltage already dissipate extra power which can make the designers to reconsider using voltage binned products.

4. As the metal and oxide layers get thinner in smaller technologies, more leakage current and therefore more variation in frequency and power static power dissipation is detected, so for smaller technologies voltage binning with the purpose of enhancing the clock frequency may not be advisory.

5. Based on an experiment that was done on three industrial chips [18] increasing supply voltage can enhance the leakage current again and that can in turn cause the clock frequency to drop. Therefore, the there are limitations on increasing the supply voltage. Figure 3.14 demonstrates that depending on the chip that is being tested, after a certain amount of supply voltage, yield starts to decrease as the supply voltage continues to increase.

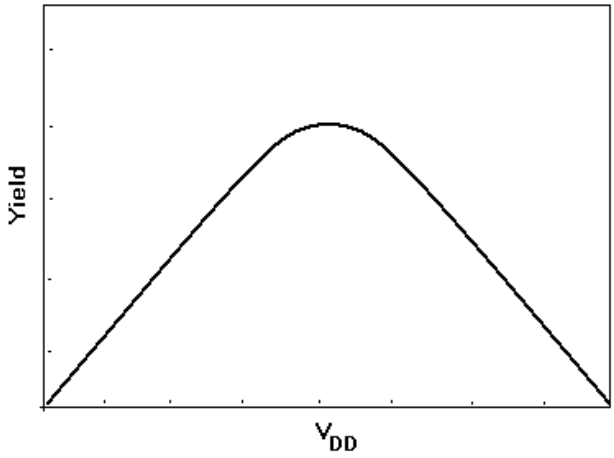

Figure 3-14: Yield vs. VDD

In an overall view, voltage binning seems to be a good technique to enhance the yield of the production for smaller designs that are not supposed to operate along with other IC's with different supply voltage because of voltage matching problems. If a specific circuit is to be used in a chip or a particular device that has strict specifications, binning cannot be a suitable solution since as soon as the product is out of the specification range, it must be discarded. For example if there are limitations for power dissipation of a circuit or supply voltage is limited due to using batteries, we must find another solution to reduce the number of discarded products. For such purposes, mid-fabrication test based mask selection technique can offer a solution to increase the yield of the production line without having to deal with matching problems and extra tunable voltage regulators and extra external pins**.**

Chapter 4: Examples for Implementation of Mid-Fabrication Test Based Mask Selection Technique

4.1 Tapered Buffer Chains

A tapered buffer chain in a synchronous system was taken as a validating example of the proposed technique. A typical chain is depicted in Figure 4-1 where each inverter in the circuit drives successively a larger inverter, allowing the output signal, v_{IN} of a minimum-sized inverter to drive the large load capacitance, C_L. The indices above the inverters correspond to the ratio between the widths of their transistors and those in the smallest inverter used in the design. That is, 1 corresponds to an inverter with W_N=0.5µm and W_P=0.9µm. This buffering scenario would occur when on-chip circuits synchronously drive an off-chip load and C_L represents the capacitance of the off-chip environment. Alternatively, this scenario also exists when a particular signal has a very large fan-out on chip.

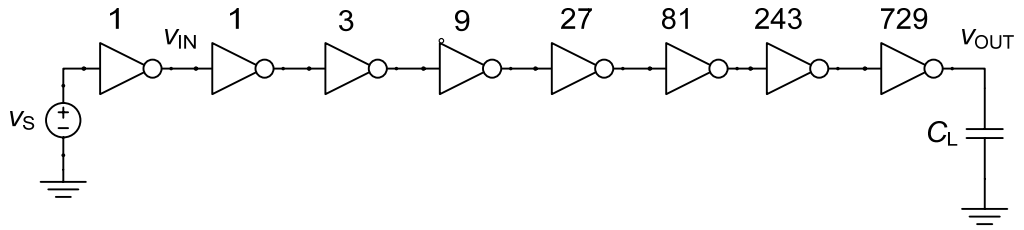

Figure 4- 1: Tapered buffer chain with a per-stage gain of 3

In some cases, the loading might be so large that even under nominal process conditions, and optimal buffering, the propagation delay through the buffers exceeds one clock cycle. In this case, one or more flip flops must be inserted in the buffer chain, thereby splitting the buffering over multiple clock cycles. This enhances the latency and is therefore undesirable.

4.1.1 Gain and Size Calculations

In order to calculate the dimensions of the transistors, the gate oxide thickness is needed. Depending on the technology the oxide layer thickness varies. Here in CMOS90NM from STMicroelectronics technology the nominal gate oxide layer thickness for N-Channel and P-Channel equals 1.77nm and 1.804nm respectively.

$$Cmos90nm \Rightarrow t_{ox\text{-}n} \approx 1.77nm,\ t_{ox\text{-}p} \approx 1.804nm$$

With knowing t_{ox} we can simply calculate the oxide layer capacitance:

$$C_{ox\text{-}n} = \varepsilon_o / t_{ox\text{-}n} = 0.345/17.7A = 19.5\ fF/\mu m^2$$
$$C_{ox\text{-}p} = \varepsilon_o / t_{ox\text{-}p} = 0.345/18.04A = 19.1\ fF/\ \mu m^2$$

With calculating the input capacitance of each inverter it will be possible to obtain the total fan-out of the buffer chain and thereby for each case, the number of inverters (N) in the buffer chain is chosen and the per-stage gain is obtained. Here in the calculation it is assumed that there are five inverters in the buffer chain:

$$C_{in} = C_{ox\text{-}n} \times W_n \times L_n + C_{ox\text{-}p} \times W_p \times L_p$$
$$C_{in} = 19.5\ fF/\mu^2 \times 0.5\mu m \times 0.1\mu m + 19.1\ fF/\mu^2 \times 0.9\mu m \times 0.1\mu m$$
$$C_{in} = 0.975 + 1.719 = 2.7\ fF$$

If the number of inverters is 5 (N=5) then we have:

$$\text{fan-out} = \frac{C_L}{C_{in}}$$

C_{in}=2.7fF (output capacitance of IVSTX1H)

$$\Rightarrow \quad \text{Fan-out} = 5pF/2.7fF \approx 1851$$
$$\Rightarrow \quad S = \sqrt[5]{1851} \approx 4.6$$
$$\Rightarrow \quad C5 = 5pF\ /\ 4.6 \approx 1.08pF\ ,\ \text{if } W_p = 1.8W_n$$
$$\Rightarrow \quad W_n = 265, Wp = 477$$

Then Wn and Wp of each inverter is 4.6 times smaller than the following inverter so it reaches to the first inverter that has the same sizing as IVSTX1H (Cadence Tools, CMOS90NM technology Library), this means the gain between IVSTX1H and the first inverter equals one. The same calculation applies to another case when N is taken as 3. Therefore S ≈ 13.

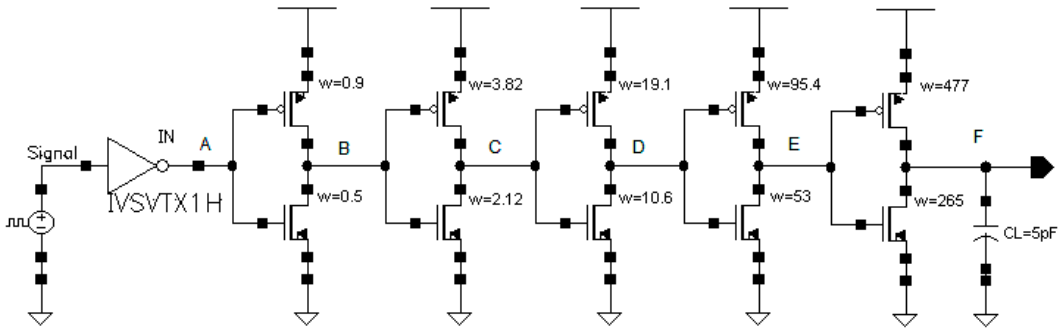

Figure 4-2: Transistor of buffer chain construction.

In order to make sure that each inverter is loading the previous stage with the same fan-out for the next stage, the buffer chain with S=4.6 is simulated to inspect the intermediate waveforms. If the rise-time and fall-time of the intermediate nodes are equal, then the calculation of sizing is correct. Figure 4-3 presents the intermediate signals associated with the intermediate nodes. Equality of the rise-time and fall-time of a signal at each node with the one at its following node, verifies that the calculations are indeed correct.

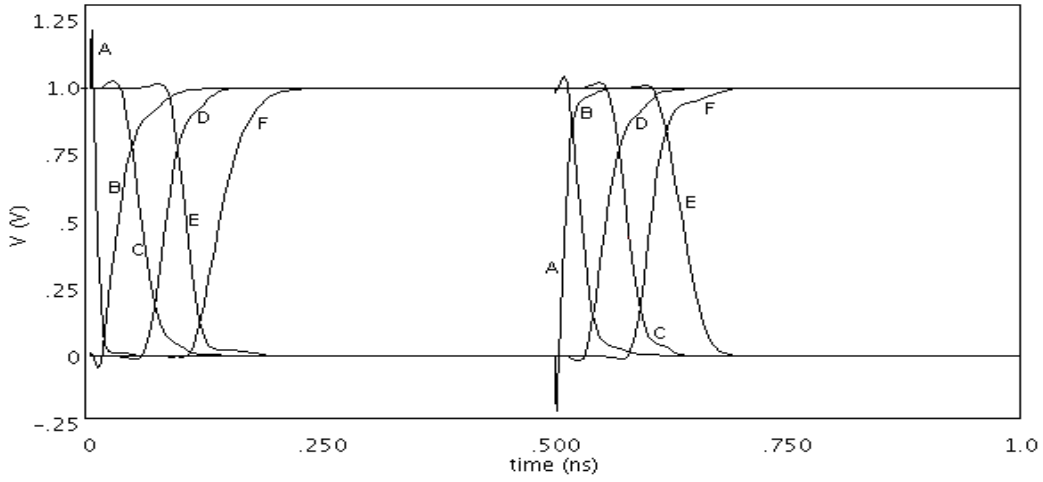

Figure 4-3 Intermediate waveforms, points A to F are corresponding to the intermediate connections in Figure 4-2

4.1.2 Simulations and Results

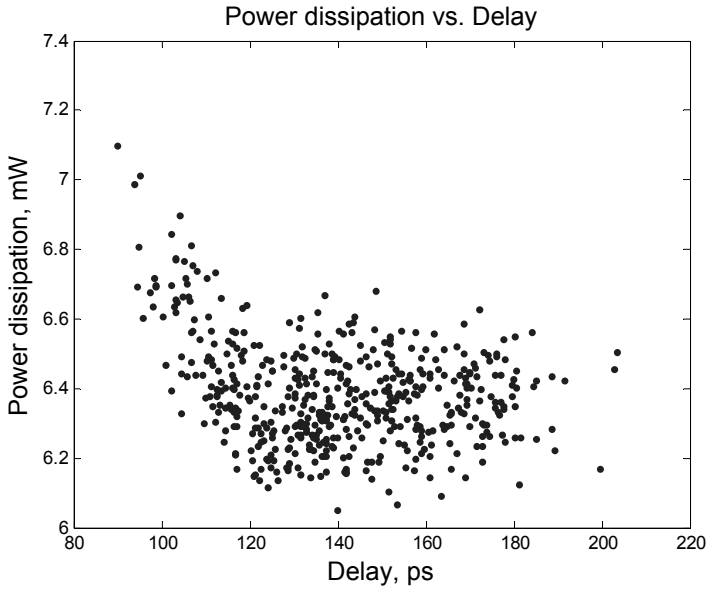

Figure 4-4: Power dissipation vs. delay for buffer chain with 1 GHz input

In this design example, we consider the case in which under nominal process conditions, buffering can occur within one clock cycle (F_{clock}=1GHz), and with an acceptable power dissipation. The circuit of Figure 4-1 is employed since its per-stage gain is close to the known delay optimal design for which the per-stage gain is *e* [19]. Figure 4-4 depicts a scatter plot of delay and power dissipation for this circuit, taken over 500 Monte Carlo process variation simulations. Suppose that system specifications require that the buffer's propagation delay be below 225 ps. All 500 versions of this circuit meet the specification, with an average power dissipation of 6.4 mW and a maximum power of 7.1 mW. The average delay is 140 ps with a standard deviation of 23 ps.

In this example we consider possible improvements in power dissipation if, based on mid-fabrication process measurements, alternative buffer configurations can be selected. In particular, we explore buffer chains with per-stage gains of 3, 4.66, 6.84, and 13.0, corresponding to 4, 3, and 2, intermediate buffer stages, respectively, as shown in Figure 4-5.

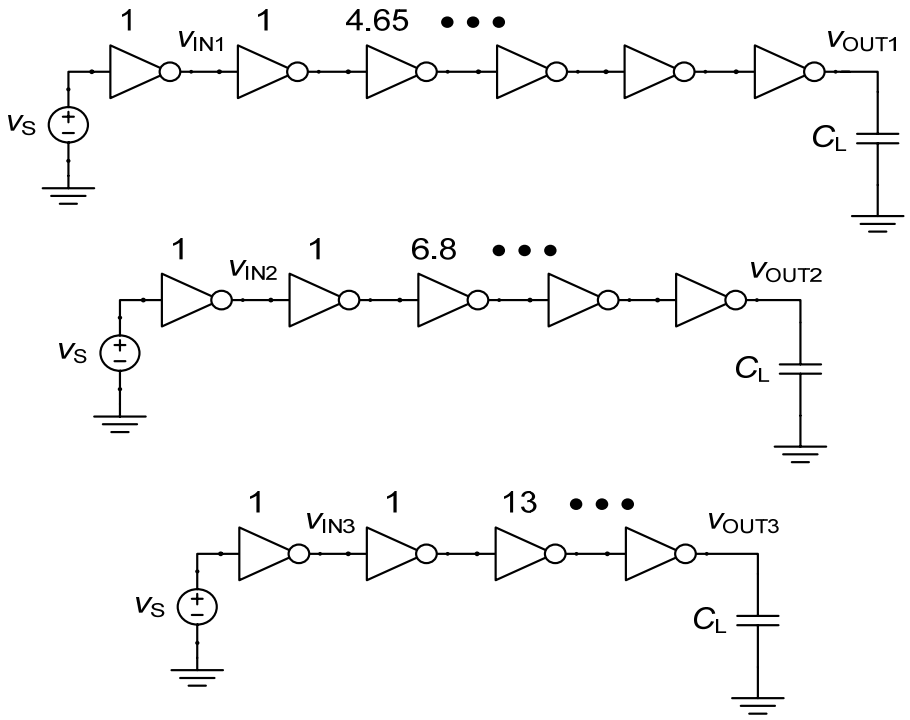

Figure 4-5: Buffer chains with per-stage gain of 4.65 (top), 6.8 (middle), and 13 (bottom)

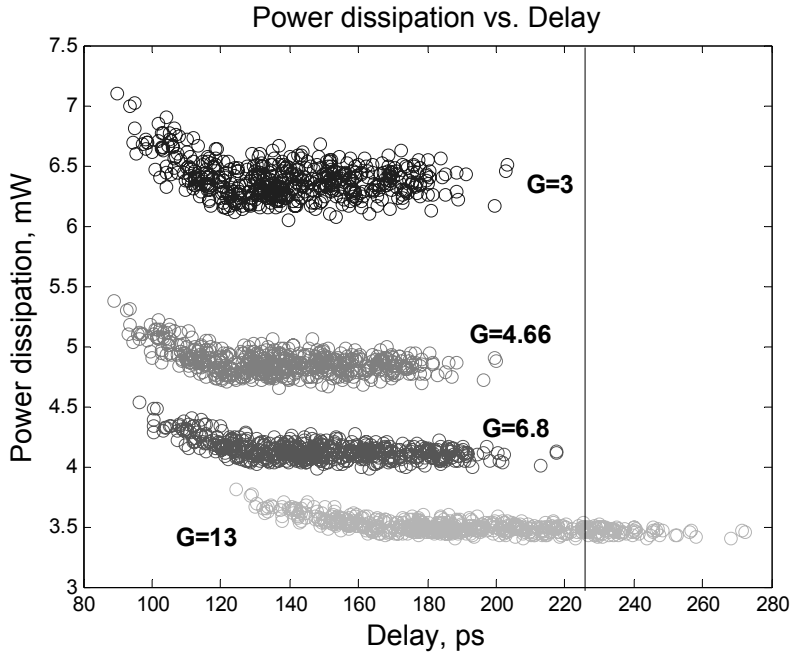

Figure 4-6: Power dissipation vs. delay for buffer chains with 1 GHz input, per-stage gains of 3, 4.66, 6.8, and 13.

Buffer index	Gain = 3	Gain = 4.66	Gain = 6.8	Gain = 13
1	*0.9, 0.5*	*0.9, 0.5*	*0.9, 0.5*	*0.9, 0.5*
2	*2.7, 1.5*	*4.19, 2.33*	*6.15,3.42*	*11.68,6.49*
3	*8.1,4.5*	*19.5,10.8*	*42.1,23.4*	*151.6,84.2*
4	*24.3, 13.5*	*90.78, 50.4*	*287.8,159.9*	*Not used*
5	*72.3, 40.5*	*422.6,234.7*	*Not used*	
6	*206.9,121.5*	*Not used*		
7	*620.7,364.5*			

Table 4-1: Device sizes in inverter chains (W_P, W_N (μm))

Table 4-1 lists the sizes of NMOS and PMOS transistors for each buffer chain for the chosen gain.

4.1.3 Applying Delay and Power Specifications

4.1.3.1 One Specification Analysis

Figure 4-6 presents the scatter plots for the power dissipation and delay of the four candidates for the buffer chain. Under process variation, it can be seen that all iterations of the chains with gains of 3, 4.66, and 6.8 satisfy the 225 ps specification. Only 421 of 500 iterations for the per-stage gain of the G13 buffer chain meet the specification. Notice that for all four chains, shorter delay is correlated with higher power dissipation. When process parameters yield short delays, excess power is dissipated. For example, the fastest of the G6.8 chain has a delay of less than 100 ps, but a power dissipation of 4.5 mW. Should a wafer be fabricated with those process parameters, a G13 buffer chain would suffice, still meeting specification with a significant margin, while dissipating less than 4 mW.

In the following, we assume that we can make process parameter measurements that accurately predict delays thereby allowing us to select masks that determine which buffer chain is implemented. It is further assumed that the design can be altered only by selecting a variation of a Via1 mask.

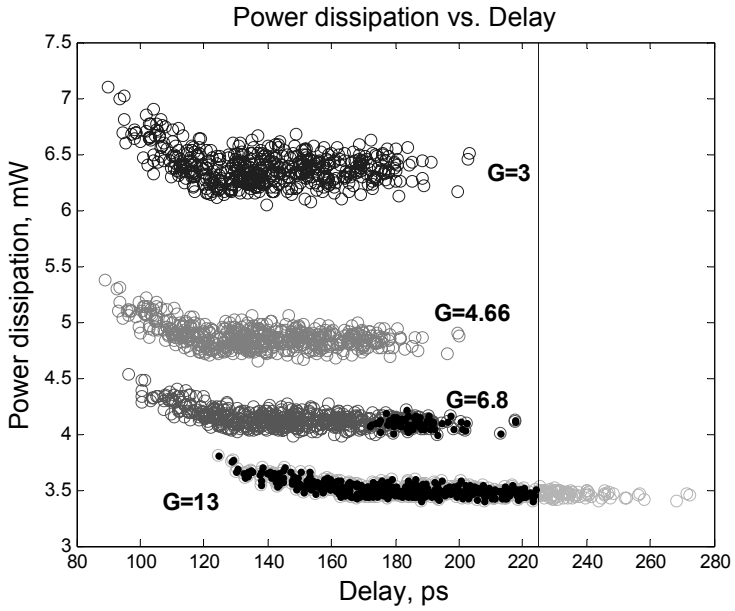

Figure 4-7: Power dissipation vs. delay for buffer chains with 1 GHz input and per-stage gains of 3, 4.66, 6.8, and 13. The black dots represent the selected chain as per our scheme

Figure 4-7 presents the results from our proposed technique. If process parameters indicate that a wafer is slow, the G6.8 chain is fabricated, ensuring that the circuit meets the specification. However, to save power, the G13 chain will be fabricated when the wafer is fast enough. Following this approach (instead of always utilizing a G6.8 buffer chain) means that all iterations meet the specification, while reducing the average power dissipation of the 500 iterations from 4.14 to 3.61 mW or 13%, while the worst case power dissipation decreased from 4.54 to 4.21 mW or 7.8%.

This technique has been investigated for various delay specifications with results summarized in Figure 4-8 and Figure 4-9. Figure 4-8 demonstrates how the yield increases from 0 or near 0 out of 500, to 500 or close to 500 as the specification is relaxed from 100 ps to 250 ps. Buffer chains with G4.66 and 6.8 reach a yield of 500 of 500. The aggregate yield, if any of the three chains can be chosen is shown as the curve G=any. This approach has the same yield as the G4.66 curve, but will offer lower average and maximum power dissipation. Figure 4-9 presents average and maximum power dissipation for each of the three candidate chains, as a function of delay specification. As the delay specification becomes more relaxed, higher number of iterations for each fixed implementation meet the specification and the statistics are based on a larger sample that includes less dissipative (and

slower) iterations. The maximum power dissipation (for each fixed implementation) stays constant as the delay specification increases since the first iteration obeying the specification is typically the most dissipative.

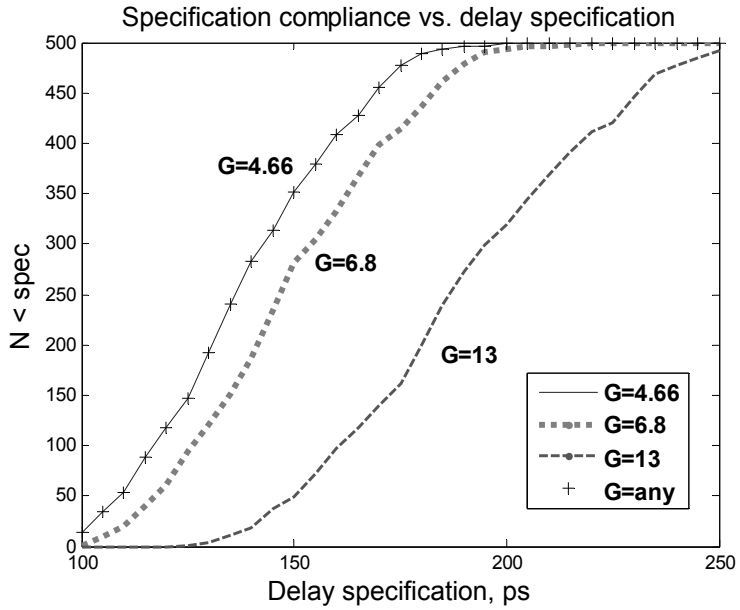

Figure 4-8: Number of delay chains that meet specification vs. delay specification. G=any refers to when any of the three chains can be selected

Figure 4-9 illustrates that by fabricating the most appropriate buffer chain, average and maximum power dissipation can be both reduced, while maintaining high yield. As a second specification example, suppose that the chain needs to have a delay less than 180 ps. From Figure 4-8 it can be inferred that for this delay, to have a high yield from a single design one needs to fabricate the G4.66 chain (N=489 pass). This chain has a maximum power dissipation of 5.4 mW and an average dissipation of 4.9 mW. However, if we can select between the buffer chains, maximum power dissipation drops to 5.0 mW while the average power dissipation drops to 4.0 mW, without sacrificing the yield.

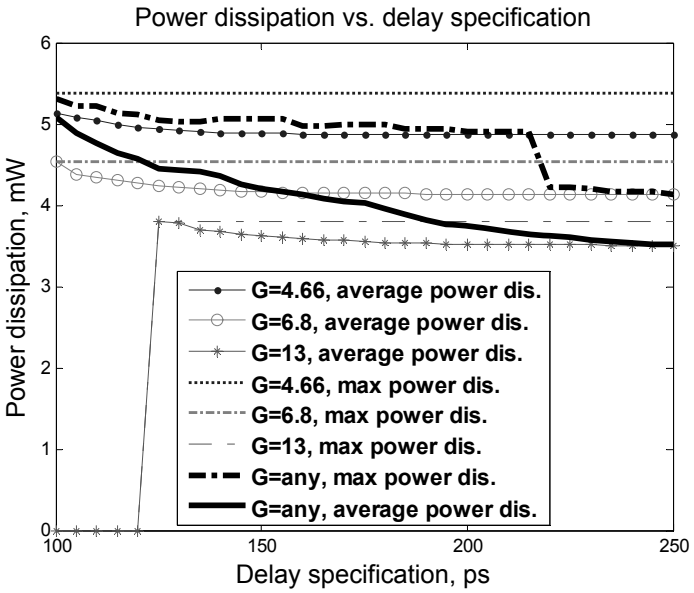

Figure 4-9: Average and maximum power dissipation for G=4.66, 6.8, 13, and "any" buffer chains. Any refers to optimal selection among the three

4.1.3.2 Two Specifications Analysis

The circuit example from the previous section is re-examined considering both power and delay as specifications simultaneously.

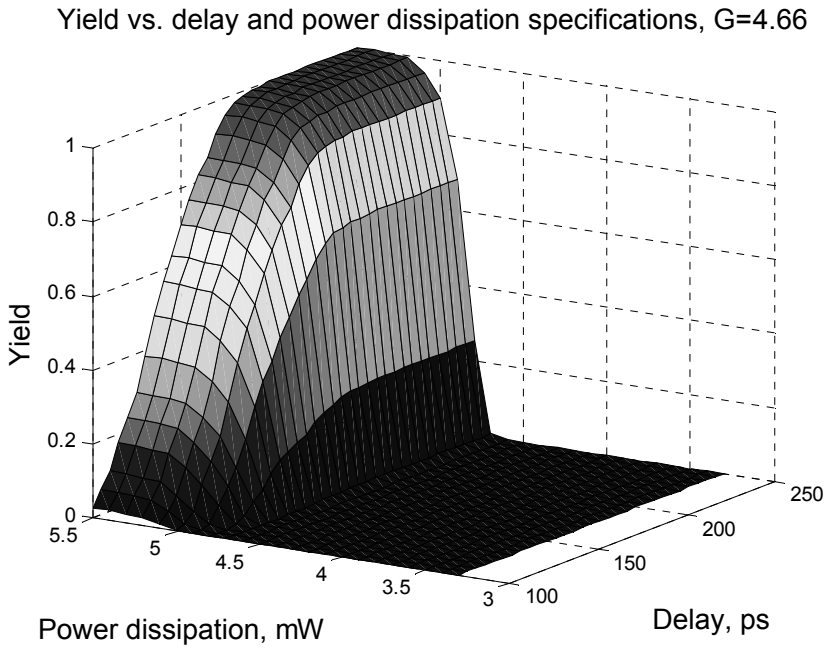

Figure 4-10: Yield vs. delay and power dissipation specifications for G = 4.66, 500 Monte Carlo simulations

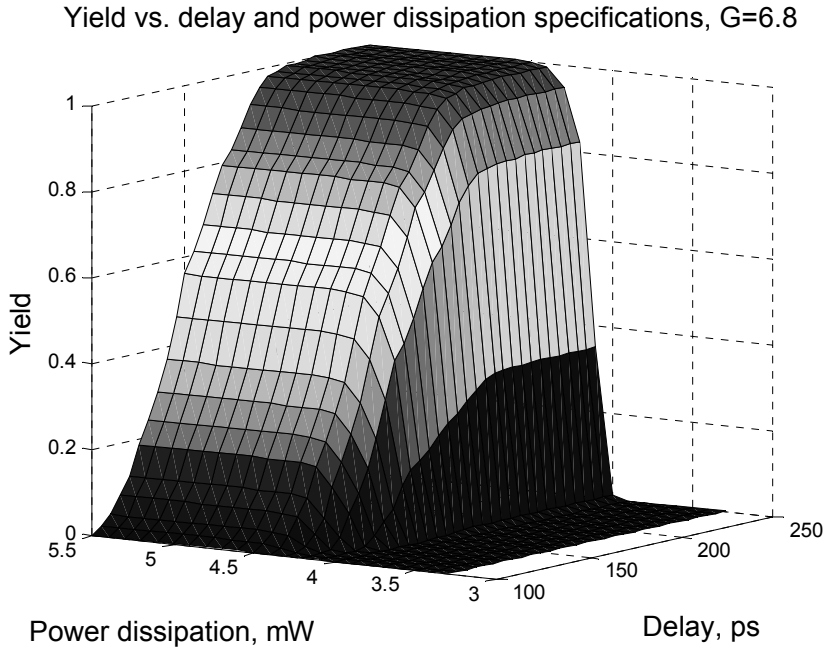

Figure 4-11: Yield vs. delay and power dissipation specifications for G = 6.8, 500 Monte Carlo simulations

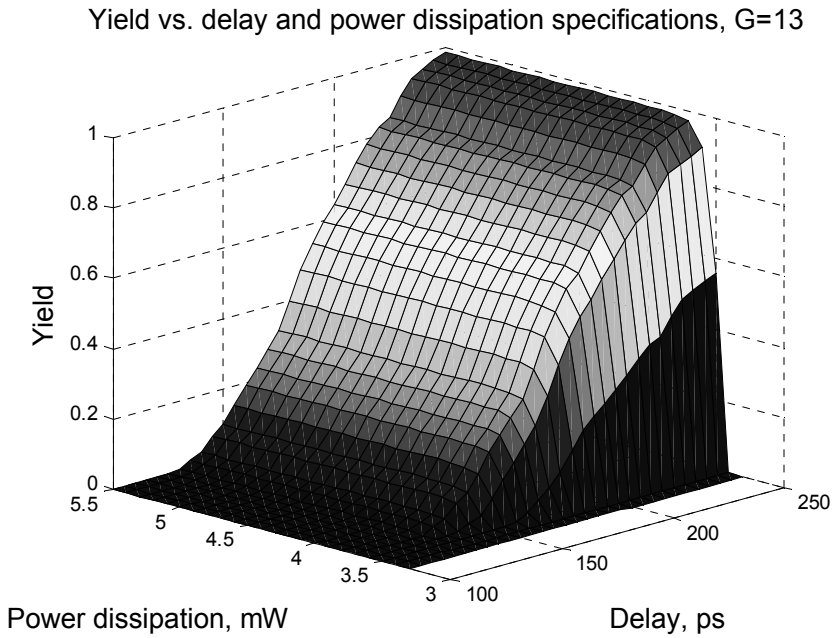

Figure 4-12: Yield vs. delay and power dissipation specifications for G = 13, 500 Monte Carlo simulations

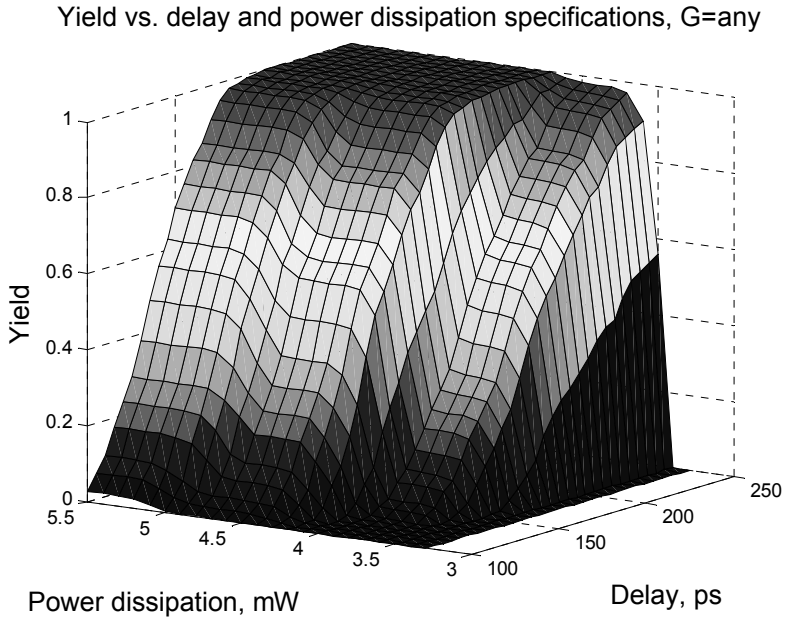

Figure 4-13: Yield vs. delay and power dissipation specifications for G = any, 500 Monte Carlo simulations

Figure 4-10 through 4-12 present the fraction of buffer chains that meet the specifications when both a delay and a power dissipation specification are considered simultaneously for buffer chains with gains of 4.66, 6.8, and 13, respectively. The shape of each surface reflects the power-delay trade-off in each design. Figure 4-13 demonstrates the fraction of buffer chains that met the two specifications if the gain of the implemented buffer chain were selected based on mid-fabrication measurements. This surface has a larger region of near-complete yield.

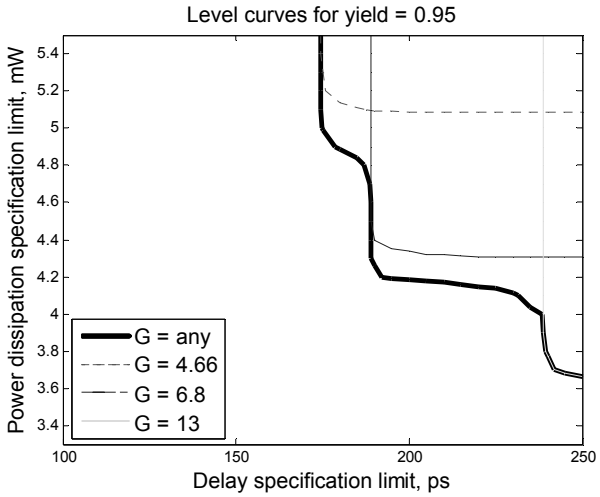

Figure 4-14: Level curves for G = 4.66, 6.8, 13, and "any", Yield = 0.95

One can assess the effectiveness in this technique in several ways. Figure 4-14 depicts level curves taken at Yield = 0.95 from each of the surfaces in Figure 4-10 through Figure 4-13. The thick line denotes the region in the power-delay specification space in which the dynamic implementation (G=any) achieves 95% yield. This region is larger than the combination of the level curves for each of the candidate implementations, indicating that there are combinations of power and delay specifications for which this technique can achieve 95% yield, while no static implementation could. One does not have to conclude, however, that the proposed technique lacks utility should the specifications be such that a fixed implementation also gives 95% yield. This is because the choice of buffer chain gain relies heavily on accurate modeling of the technology. It is known that during processor/technology co-design, the "typical" process characteristics may be known with only limited accuracy, making the gain (or any other timing/power related choice) difficult to choose. Using the technique proposed above allows the circuit implementation to be selected only after technology parameters are known,

thereby alleviating many of the uncertainties in the circuit/technology co-design. Even if only one buffer chain gain were selected, the proposed technique has merit when the "typical" process parameters are unknown at the design time.

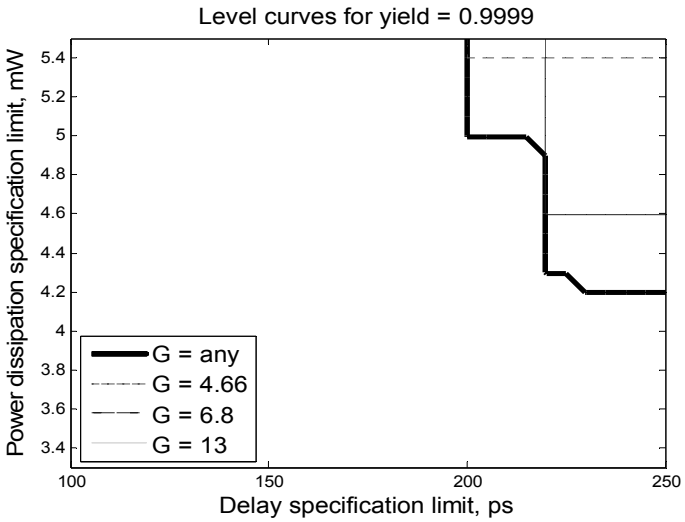

Figure 4-15: Level curves for G = 4.66, 6.8, 13, and "any", Yield = 0.9999

Figure 4-15 shows a similar level curve for a yield of 0.9999. In this example, with a sample size of 500, this corresponds to all 500 cases of meeting the specification. In this example, the G = "any" approach greatly increases the size of the region.

Another means by which the proposed technique can be evaluated is to examine the yield improvement across the power-delay specification space. Figure 4-16 demonstrates yield improvement over the power-delay specification space. If the implementation were selected dynamically based on mid-fabrication process measurements, this method compares the yield to the one for the best fixed implementation for a given combination of power and delay. Though not obvious in the 3-dimensional plot, the largest improvement in yield is 20%, where the delay specification is 170 ps and the power dissipation specification is 4.1 mW. Here, the best fixed implementation is G = 13, which gave 28 % yield, while the dynamic, G = "Any" technique gave a yield of 49 %.

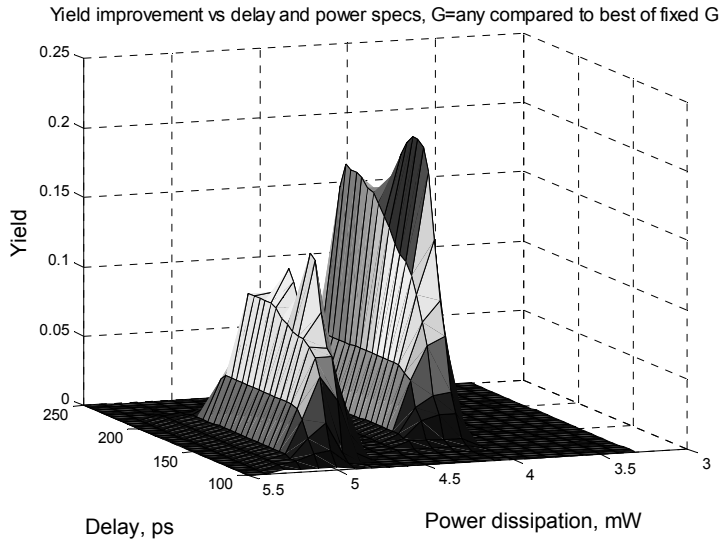

Figure 4-16: Yield improvement vs. delay and power specifications. G = "any" approach compared to the best of any of the fixed gain implementations

4.1.4 Possible Implementation

The design in Figure 4-17 allows for the realization of variant designs G4.7 and G6.8. The switches represent connections made using the Via1 layer. Those labeled "*f*" are closed on fast wafers and open on slower wafers, while those denoted by "s" are closed on slow wafers and open on fast wafers. Together, these switches realize the G4.7 variant design on slow wafers, and the G6.8 variant design on fast wafers. The G4.7 variant design connects inverters in parallel to create larger inverters.

The chains are connected to complementary outputs of the flip-flop because the G4.7 chain has an extra signal inversion compared to the gain in the G6.8 chain. The total transistor width used to implement either variant design is the same as that used to implement only the G4.7 variant design, indicating that the technique imposes no overhead in terms of transistor width.

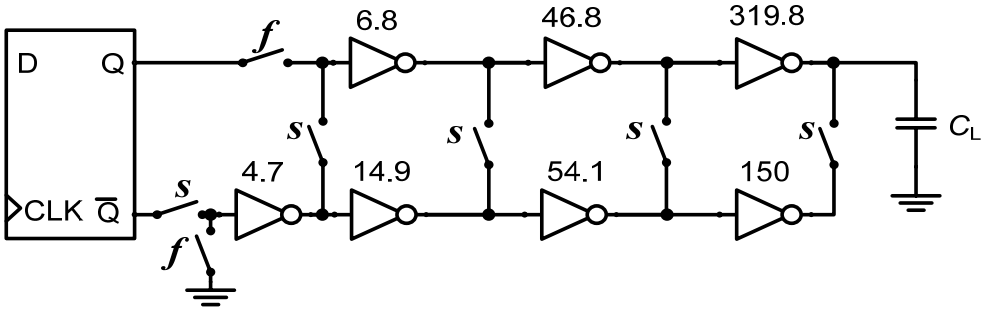

Figure 4-17: Implementing Gain=4.7 and Gain=6.8 variant designs

Figure 4-18 demonstrates one possible implementation of the switches shown in Figure 4-17, Closed and open switches differ only by a Via on the Via1 mask. By generating a variant of the Via1 mask for each variant design, the most appropriate variant design can be selected for each wafer, or even each reticle.

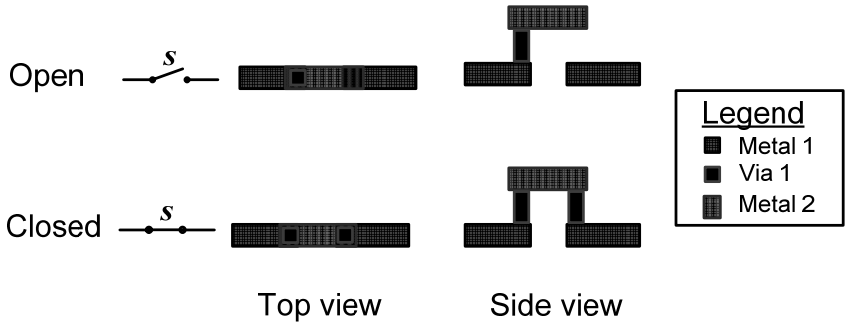

Figure 4-18: Implementation of switches in layout

4.1.5 Yield of the Production Line

In this section we intend to show how the mid-fabrication test based mask selection technique will change the yield of the production line in more detail when a set of specifications is applied to the fabricated chips. For that, we take two approaches into account, one when only delay specifications are applied and power dissipation of the chips is not limited and the other when both delay and power dissipation are defined with a specific value. Ideally, the manufacturers want to produce the fastest chips possible, so when power dissipation of the chip is not a subject in controlling the product line, the design in which the delay is always minimum would be taken as the default design. Here, in the first example, the buffer chains, the design with G3 is always the fastest design and it also dissipates maximum power. Now we assume that there is no power limits defined but the chips are preferred to have the lowest power dissipation and as long as they meet the delay specifications, the one with lower power dissipation is accepted. Neglecting power dissipation may not be an option in reality but here in this section we assume that there is a preference to choose the least dissipative design. According to Figure 4-19, the buffer chain with G13 is taken as the default design because it is the least dissipative design for our circuit example and if

the delay exceeds the delay specifications the design is switched to the first possible design in which the delay is within the accepted.

4.1.5.1 Yield and One Specification Algorithm

Assume that the maximum accepted delay for the buffer chains is 190ps. The selection algorithm will be as follows:

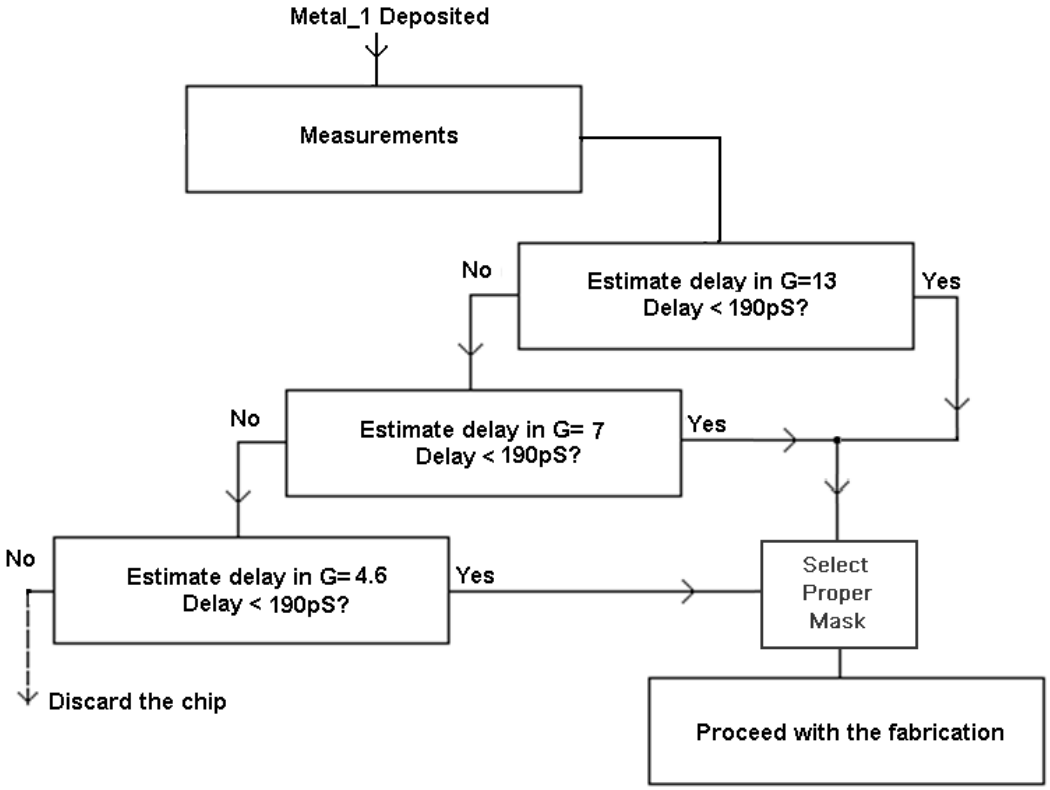

Figure 4-19: Algorithm of 1-specification selection

The result of this selection is shown in the following Figures. In Figure 4-20-a the delay-power spread shows the dispersion of each design and the result of the 1-Spec selection is highlighted. The histogram in Figure 4-20-b illustrates the result of this selection.

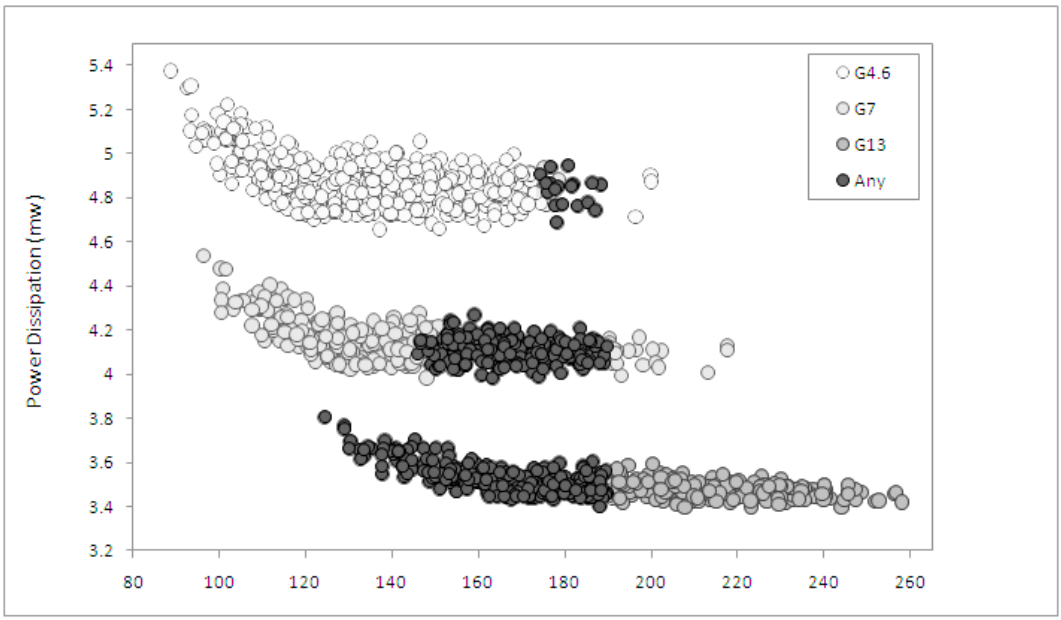

Figure 4-20-a: Power dissipation vs. Delay spread of each buffer chain

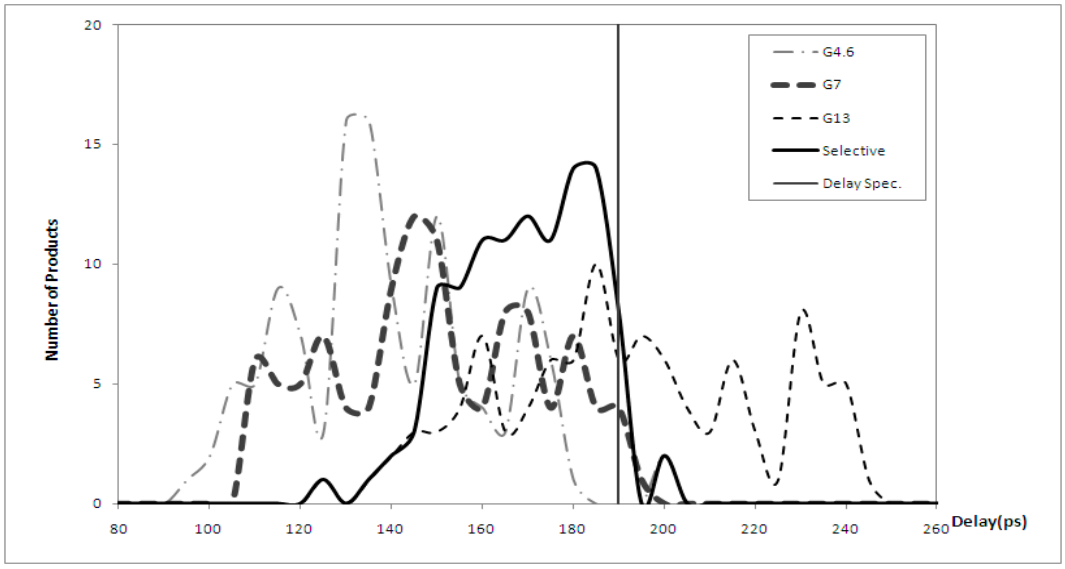

Figure 4-20-b: Number vs. Delay of each buffer chain after applying the delay specification

In Figure 4-20-a the selected samples are the darker dots that are filtered by the algorithm shown in Figure 4-19. The diagram in Figure 4-20-b in bold-black line is the result of the selective approach which shows a higher number of samples in which the delay is close to 190ps. The other three diagrams (G=4.6, 7, 13) show the concentration of the samples before any specification is applied. By applying the selective approach we can increase the number of

the samples that meet the delay specifications with lower power dissipation (see Figure 4-20). Table 4-2 lists the statistics and the yield of each design after only delay specifications are applied

1-Spec	Discarded	Loss	Yield	Avg-D	Avg-P
G4.6	3	0.60%	99.40%	166.5076	4.869
G7	21	4.20%	95.80%	146.69	4.14
G13	228	45.60%	54.40%	166.54	3.53
Selective	3	0.6%	99.40%	167.51	3.81

Table 4-2: Results of 1-specification selection on each design

As shown in Table 4-2, when delay specifications were applied to each design, a specific number of the samples in each design did not pass and had to be discarded. This leads to a decrease in the yield of the production line depending on how tight the limits are. Here, after applying a 190ps maximum delay limit, we lost 0.6% of the samples if the G4.6 design was always selected, 4.2% of the samples in G7 and 45.6% in G13. The last row of Table 4-2 shows the result of the selective approach that was offered by the mid-fabrication test based mask selection technique. If the proposed technique is applied during the fabrication process, the number of the discarded samples is reduced down to 0.6% with the average power dissipation of 3.81 mW, whereas the average power dissipation of the G4.6 is 4.87 mW. Although in one specification analysis we do not include power dissipation into filtering process we still prefer to have lower average power dissipation. There is a 21.7% decrease in the average power dissipation of the selective approach compared to only G4.6 as the fastest design. Also a 0.6% increase in the average delay is detected. Considering that the yield is the same in both selective and G4.6, the proposed technique can decrease the average power dissipation from 4.87 mW to 3.81 mW with keeping the yield intact.

4.1.5.2 Yield and Two Specification Algorithm

In this section, the selection algorithm has two criteria to filter the manufactured chips. This time delay (190ps) and power dissipation (4.9mW)

specifications are applied to the testing procedure as it is demonstrated in Figure 4-21.

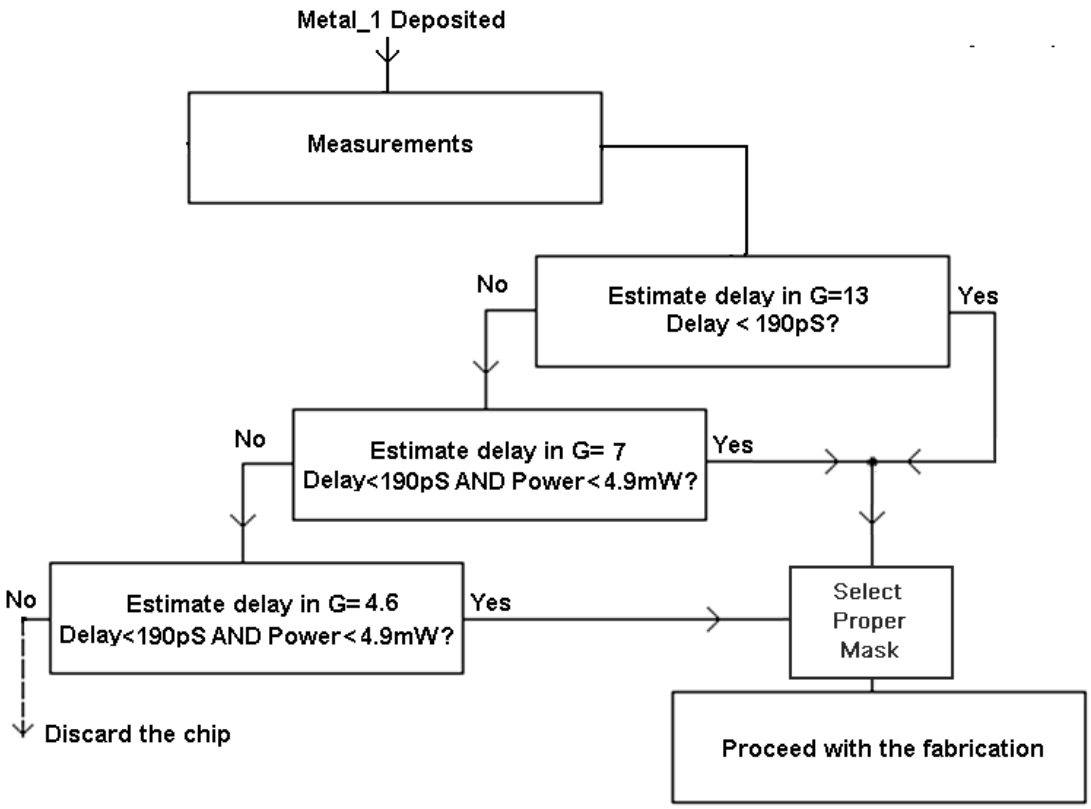

Figure 4-21: Algorithm of 2-specification selection

Once again if the delay in the test circuit in which the process variation is correlated with G13 is less than 190ps, G13 is selected and the wafer is sent to the next sequence of the fabrication. The reason the power dissipation for G13 is not measured is that all the samples of G13 in Monte Carlo simulations have power dissipations less than 4.9mW. If the delay is more than 190ps then the test circuit correlated with G7 is selected and tested. If the delay is below the accepted limit the power dissipation is measured. If the power is below 4.9mW then G7 is selected and if not, G4.6 is selected. When both delay and power dissipation are below specifications, G4.6 is selected and if not, the wafer is discarded and the next wafer is taken into the test procedure.

Figure 4-22 illustrates the selected samples on top of all G4.6, G7 and G13 samples in a 500 run Monte Carlo simulation.

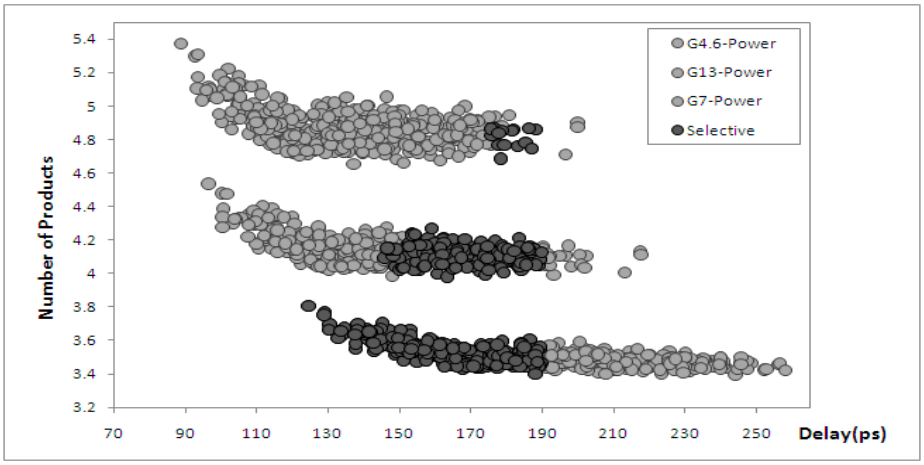

Figure 4-22: Delay-power spread of G13 and G7. Darker dots are the selected chips after two specifications were applied

In Figure 4-23 three histograms are presented. Once again three unfiltered data set of each buffer chain are compared with the result of the selective approach. The histograms show an increase in the number of samples that meet the specifications, in the "Selective". The statistics are discussed below.

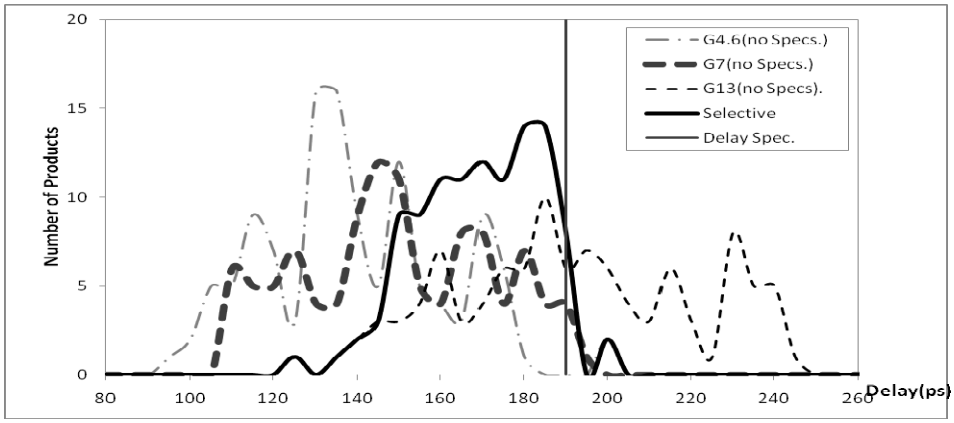

Figure 4-23: Histograms of number vs. delay of two buffer chains after 2-specifications. Selection is applied; "Selective" shows the results of mid-fabrication test based mask selection technique

When both delay and power specifications are applied, 160 samples out of 500 in G4.6 do not meet the specifications and have to be discarded. The yield of G4.6 when delay spec is 190ps and power spec is 4.9mW is 68% whereas it is 95.8% for G7 and 54.4% for G13. In the diagram for the

selective approach the 6% loss is the part that stays outside the vertical line (delay specification).

2-Specs.	Discarded	Loss	Yield	Avg-D	Avg-P
G4.6	160	32%	68%	142.64	4.81
G7	21	4.20%	95.80%	146.7	4.14
G13	228	45.60%	54.40%	166.54	3.53
Selective	6	1.20%	*98.80%*	167.45	3.81

Table 4-3: Statistic of the 2-specification approach

Applying the mid-fabrication test based mask selection technique offered 98.8% of yield for the tapered buffer chains if 2-specification selection was applied. This is the highest yield in all cases as reported in Table 4-3.

By analyzing the results table of both 1-specification and 2-specification algorithms we understand that defining the delay and power dissipation specifications plays the most important role in calculating the yield of the selective approach. It means the selective design will always have the highest yield with a better average power dissipation percentage since the design with lower power dissipation is taken as the default design in the algorithms. For example, in Table 4-3, G7 has a high yield close to the selective approach, so to answer the question that why do we choose to apply the proposed technique when G7 can provide a high yield we can answer that the yield of G7 can simply vary by changing the specifications but since the selective approach offers the highest number of samples possible of all designs can always offer a higher yield as well as lower average power dissipation and this is when delay of all samples is still in the accepted range.

4.2 Sense Amplifier Flip Flop

4.2.1 Two Designs of the Sense Amplifier Flip Flop

A Sense Amplifier Flip-Flop (SAFF) is taken as an example of the proposed technique. Generally conventional flip-flops are built in two sections: the pulse generator (PG) and the slave latch (SL). The inputs of the PG section are the Data and Clock and the results that are generated depending on the timing of the input signals and the design conditions, are forwarded to the latch stage. A sense-amplifier flip flop is widely used in integrated circuits[20]. It consists of a Sense-Amplifier and a Set-Reset latch (Figure 4-24). SAFFs have more than one implementation that can trade-off delay and power dissipation. Here the proposed technique is applied to two of them: a conventional design (Figure 4-24) in which the RS latch is constructed by two Nand gates and a modified design (Figure 4-25) which has a clock-controled latch following the sense-amplifier [21].

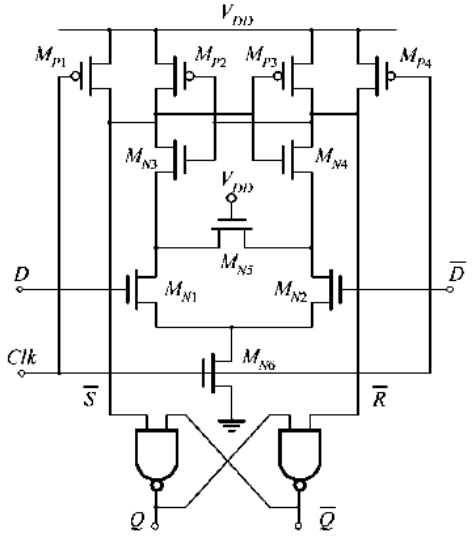

Figure 4-24: A SAFF with two Nand gates constructing the RS Latch

The two designs were sized to have equal rise and fall times. As depicted in Figure 4-26 the modified design has a faster response to the same load (C_L=10fF) in both rising and falling edges of Q to a CLK.

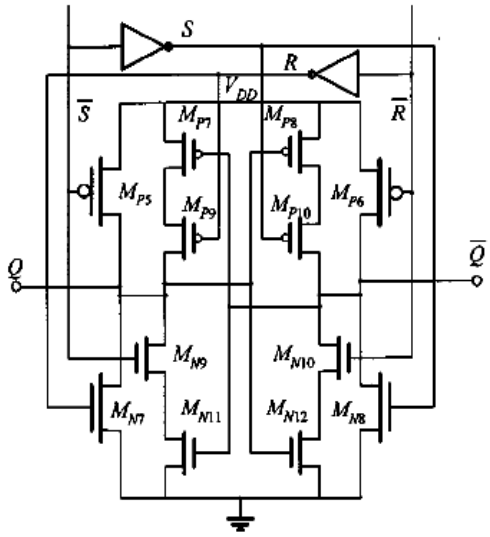

Figure 4-25: The proposed clock controlled latch

Both designs were simulated with 1000 Monte Carlo process variation iterations. The delay between the CLK and Q of each design is illustreted in Figure 4-26. The average delay in the modified design is 12% less (rising edge of CLK to rising edge of Q) than the conventional one whereas the power dissipation is 14% higher. The rise-to-fall response of the modified design was faster than the rise-to-rise response. However, to have the least optimistic comparison between the two designs, the rise-to-rise delay of CLK to Q response was taken as the data set in all calculations.

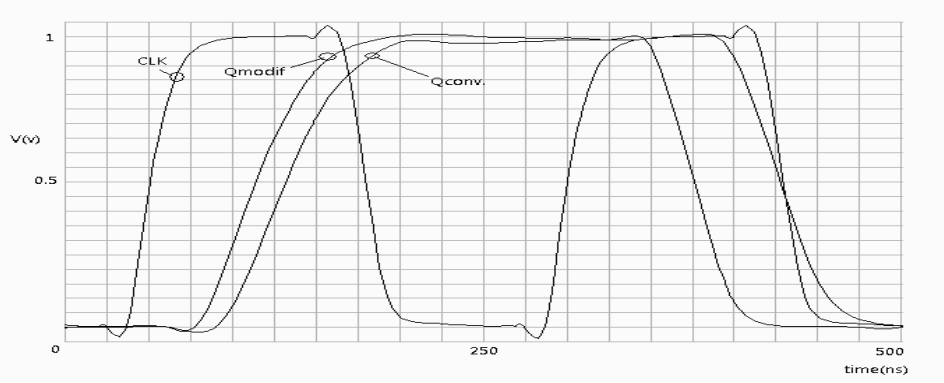

Figure 4-26: Waveforms of $Q_{conventional}$ and $Q_{modified}$ compared to CLK

4.2.2 Applying Specifications

The results of 1000 Monte Carlo process variation simulations for each design are presented in Figure 4-27. In the SAFF example, according to the data sets the modified design is always faster and dissipates more power for a given process corner.

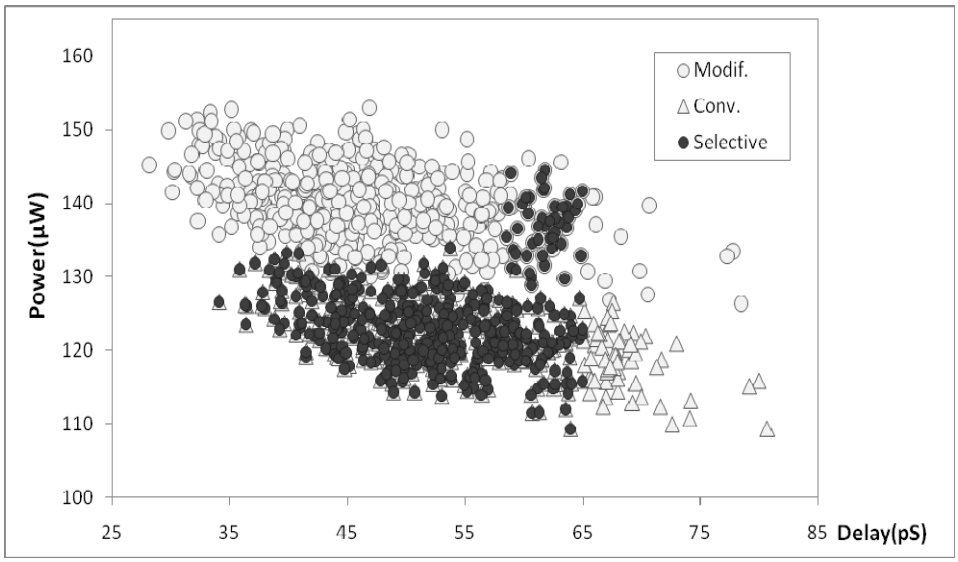

Figure 4-27: Delay vs. power dissipation distribution for both (modified and conventional) designs, the black dots "Selective" show the accepted samples for the selective approach

Again if the 2-specifications selection algorithm is applied (maximum delay of 65ps and maximum power dissipation of 145µW), 11.5% of the products when the conventional flip flop is selected have to be discarded since they do not meet the specifications. The same value is 22.2% for the modified flip flop when it is taken as the default circuit. The algorithm of the mid-fabrication selecting process is the same as the one in the 2-specifications selection in buffer chains. Table 4-4 lists the results of the selection algorithm for each design. The row with "Selective" shows the results of the mid-fabrication test based mask selection technique.

The same algorithm as in Figure 4-24 is chosen for the "selective" approach, in which the conventional design is the default design and then the delay of the test circuit that is correlated with the conventional design is measured. If

the delay is less than the delay specification then the power dissipation of the test circuit is measured or if the power dissipation is less than the allowed value then the conventional design is chosen. Otherwise, if any of the two conditions are not satisfied then the test circuit in which the delay and power dissipation are correlated with the modified SAFF is taken for measurements. The same algorithm is applied and if any of the two conditions in not satisfied, the wafer is discarded otherwise the modified design of SAFF is chosen and the manufacturing process continues. As it is listed in Table 4-4, the loss percentage is reduced down to 3.4% when the selective approach is utilized.

Design	Discarded	Loss	Yield	Avg.-D	Avg.-P
Conventional	*115*	*11.50%*	*88.50%*	*51.8161*	123.14
Modified	*222*	*22.20%*	*77.80%*	*47.64*	*138.55*
Selective	*34*	*3.40%*	*96.60%*	*52.64*	*124.33*

Table 4-4: The results of 2-specifications selection for both designs and the proposed technique

Figure 4-28 presents the comparison between the histograms of each design and also the selective approach. Again, the histogram "Selective" corresponds to the results of those samples to which the proposed technique was applied. The region of the selective diagram beyond the vertical line in the RHS (delay specification) corresponds to the 3.4% loss percentage.

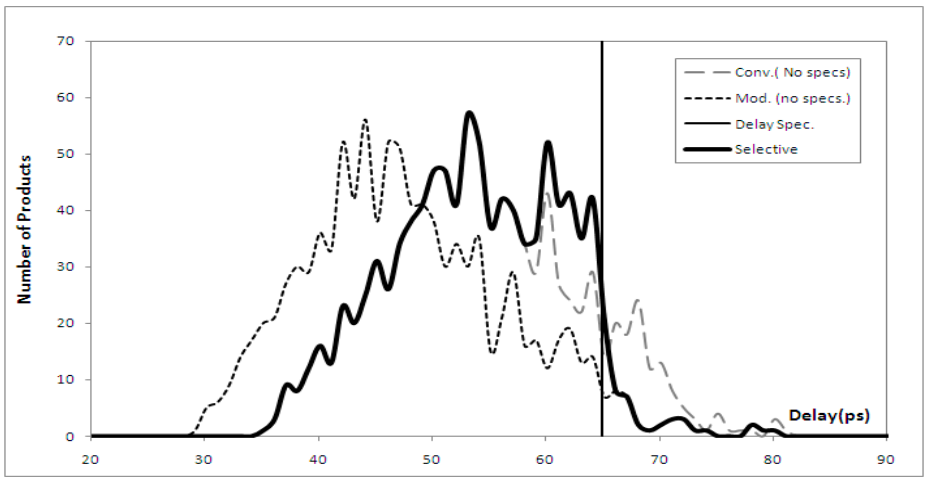

Figure 4-28: Comparison of the number vs. delay histogram for both designs

4.2.3 Possible Implementation

In order to use the proposed technique a common transistor arrangement is necessary so that each variant design can be implemented. In Figure 4-29, one possible way of implementing both designs is demonstrated. By using two different combinations of blocking masks and Via1 masks it is possible to modify the via configuration.

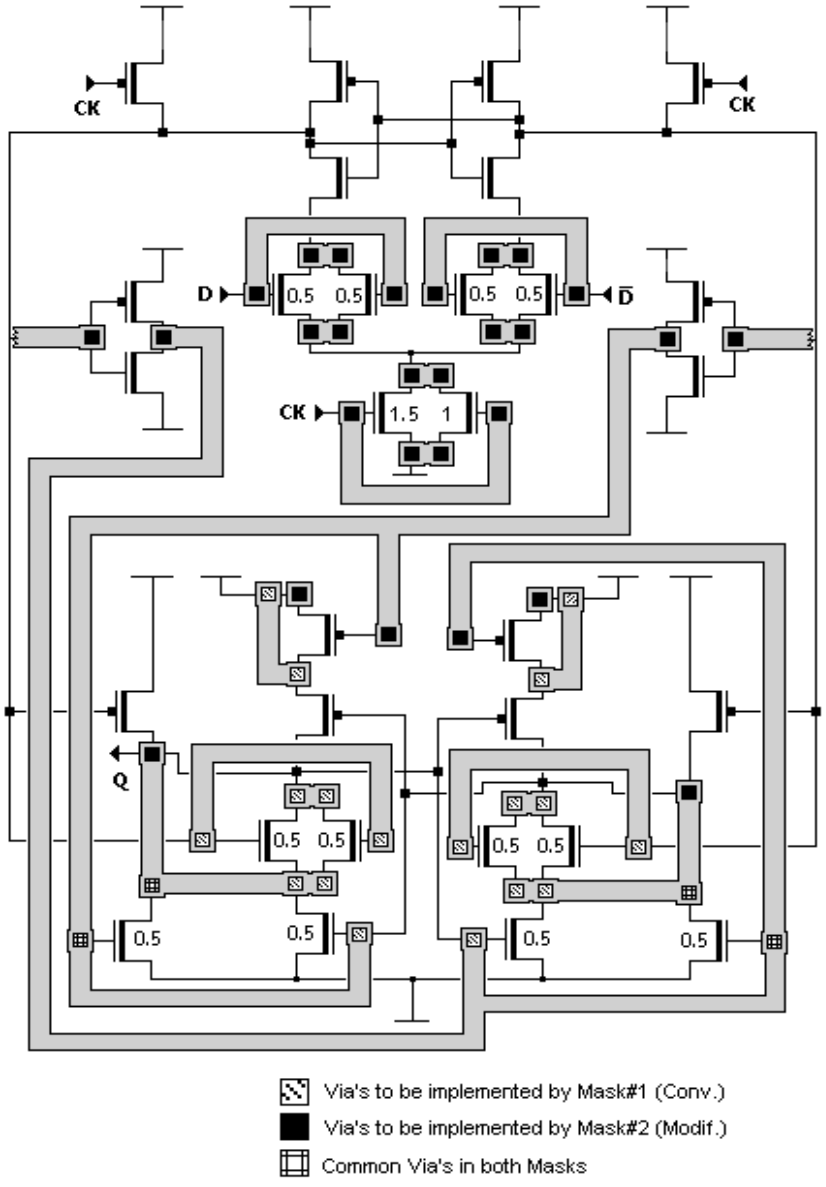

Figure 4-29: Possible Via1 mask change to implement two different designs. For simplicity common wires are drawn in thin lines

The Via1 mask includes all the vias that are used for both designs and by using the two blocking masks that are designed for both modified and conventional SAFF separately, we can include or exclude transistors to obtain the desired layout. Here in the SAFF example, the area of the design, regardless of whether the conventional design is selected or the modifeid design, is the same. By applying the proposed technique, those transistors that are bypassed (short circuited by metal_2 layer) for the conventional design are included in the modified design, therefore the transistors that are placed in parallel will act as a larger transistors.

As was shown earlier, the same circuit is used for the pulse generator in both designs. In Figure 4-29 , those transistors that are connected to each other by fixed wires will remain intact during the application of the proposed technique. To make Figure 4-29 more understandable, the vias that are used to implement the conventional design are shown in the dotted squares(▨). If the modified design is to be used, another mask will then deposit another arrangement of vias (■). Those contacts that remain intact (common for both designs) are shown in checked squares(▣). So, [▣+▨] and [▣+■] are used for the convetional and modified SAFF repectively. For simplicity, those wires that are not changed with changing masks, are drawn in black lines. In order to change the area of the common transistors, two transistors are connected in parallel with those vias that are selectively deposited by the means of the proposed technique.

4.2.4 Results and Dynamic Comparisons

Again in this section, the 3D plots help to have a deeper understanding of how the proposed technique can improve the yield of the production line. Before applying the same delay and power specifications as in the 2D histograms to define the yield improvement, the delay and power curve limits are taken as 80ps and 160μw respectively. Figure 4-30 to 4-33 present the yield of both designs versus delay and power dissipation for 1000 simulation iterations and also the yield improvement for the selective approach.

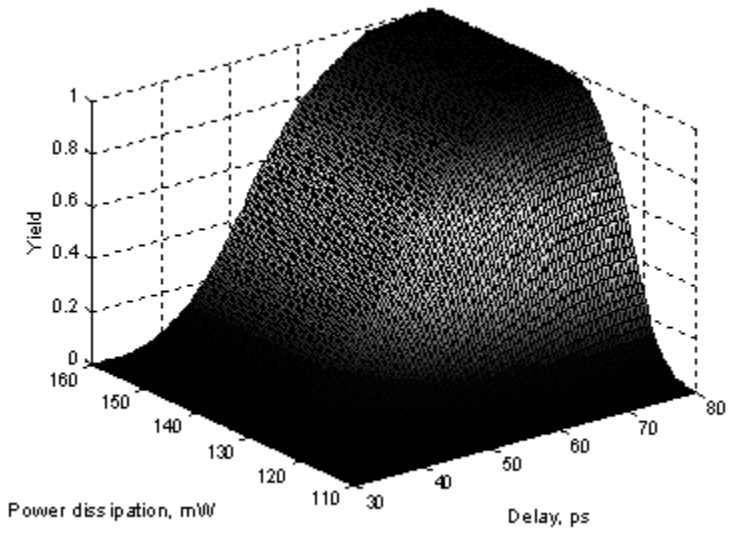

Figure 4-30: Yield vs. delay and power dissipation specifications for conventional SAFF for 1000 Monte Carlo simulations

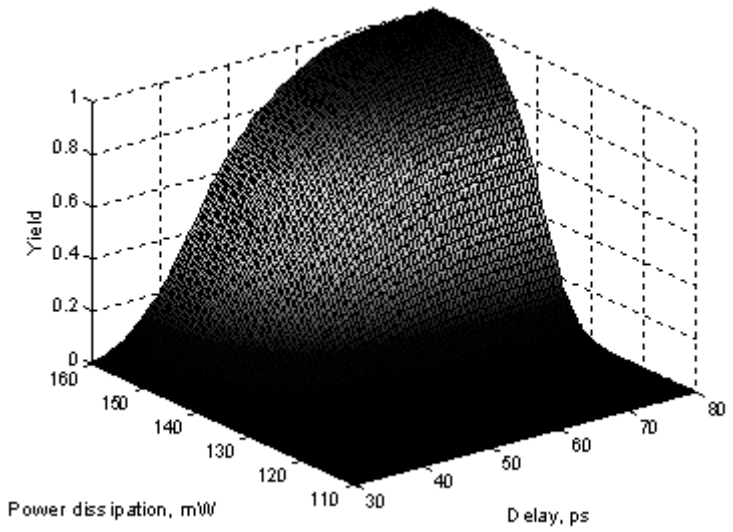

Figure 4-31: Yield vs. delay and power dissipation specifications for modified SAFF for 1000 Monte Carlo simulations

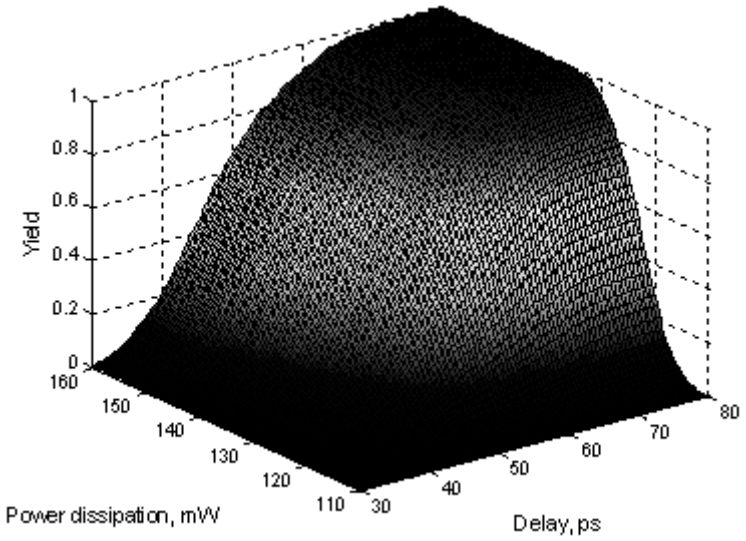

Figure 4-32: Yield vs. delay and Power dissipation specifications (selective design)

As it is displayed in Figure 4-32, the selective approach has a larger surface area near a yield of 100% as expected. The highest yield improvement that the selective approach can offer is for delay of 52ps and power dissipation of 142.5 μw which is 19.6 %. This is shows in Figure 4-33.

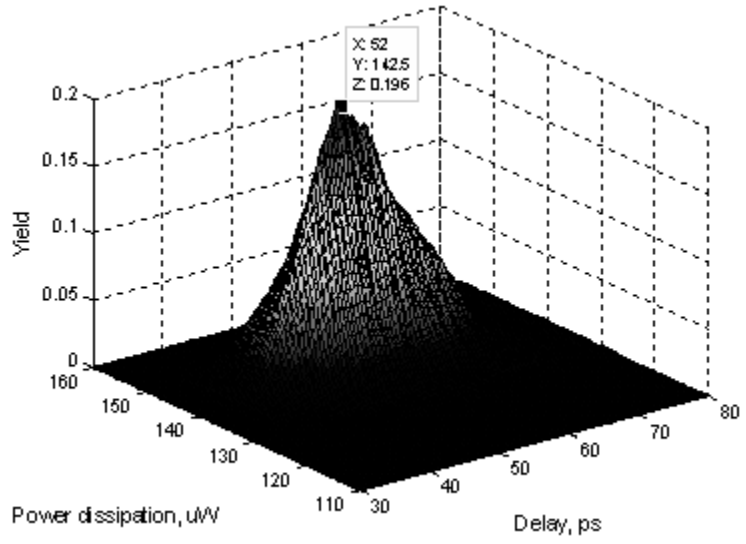

Figure 4-33: Yield improvement vs. delay and power specifications. Design is the "selective" approach compared to the best of any of the fixed gain implementations

Figure 4-34 presents the yield of two SAFF designs and the selective approach as a function of both the delay and power dissipation. The corresponding yield for a given delay and power dissipation specification is shown in each 3D graph. When delay is 52ps and power is 142.5 μW the yield is 44.3% and 63.9% for the conventional and modified designs respectively. The selective approach chooses the highest yield possible and as it is shown below, the maximum yield for this point is again 63.9% which gives the same 19.6% yield improvement. The same point is found on all 3D graphs below.

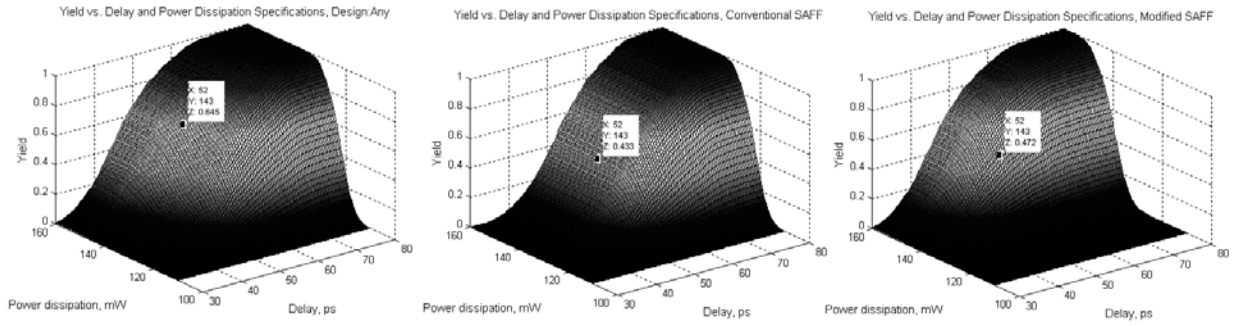

Figure 4-34: Yield of a given point of delay and power specifications, Delay=52ps and power dissipation=142.5μw

From Figure 4-34 it is inferred that selective approach picks the highest yield that is associated with the given point already belonging to one of the designs and then adds it to its dataset. However Figure 4-35 demonstrates a more interesting effect of the dynamic yield improvement for the selective approach. If all the 3D graphs are cut by a plane of Yield=95%, the result will appear as in Figure 4-35.

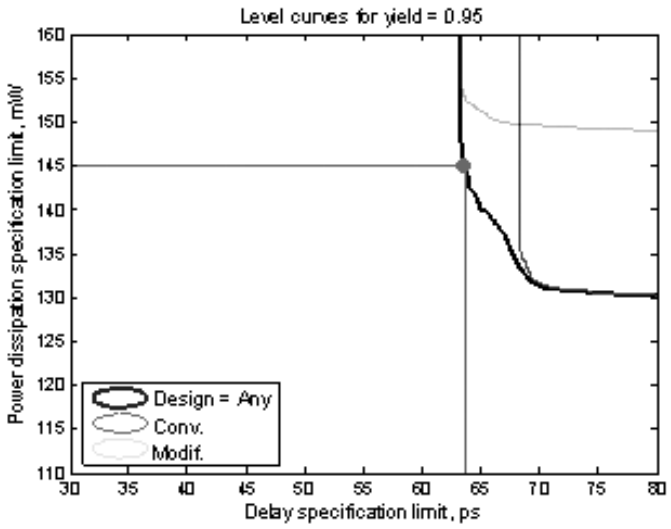

Figure 4-35: Level curves for conventional and modified SAFF, The selective, Yield is 0.95

Figure 4-35 illustrates a cross sectional area of each design's 3D graph and a plane of 95% yield. The area that belongs to the selective approach shows that the yield of the selective approach is not only the addition of the areas of both designs, but also the dynamic yield improvement offers 95% yield that is higher than the yield of both designs for the same given point. As an example for delay of 64ps and power of 145µW, the yields of the conventional and modified design are 76.7% and 87% respectively.

Chapter 5: Conclusion

A technique was proposed to increase the yield of a chip production line when both delay and power dissipation specifications were applied to the final product. Two circuit examples were taken to study to apply the proposed technique. For each circuit example a few variant designs based on different process corners were offered. An algorithm was designed to choose the best possible variant design based on mid-fabrication tests and one of the variant designs for different process corner was selected. The algorithm was applied to a 500-1000 samples of Monte Carlo simulations dataset and a new selective dataset was generated. Then a set of delay and power dissipation specifications was applied to all the possible fixed variant implementations as well as the selective dataset. The yield of the each fixed variant designs and the selective dataset was calculated separately and compared with each other. For a buffer chain a 20.8% and for a sense amplifier flip flop a 19.6% of yield improvement was detected. The limitations of the proposed technique were noted as difficulties of estimating the cost of making extra masks, the need for more area for the reserved transistors of the variant designs and the dependency of the cost of the technique to the type and the design of the circuit.

Chapter 6: Future Work:

As discussed in preceding chapters, in order to globalize and apply the mid-fabrication test based mask selection technique to more widely used circuits, CMOS circuits in which the delay and the power dissipation are inversely proportional should be found and studied. Figure 6-1 illustrates a non-linear diagram of delay vs. power dissipation. We showed that if the designer were to compromise power dissipation for higher speed, then it might be possible to find an alternative design for a specific circuit that for all simulation iteration numbers, delay is lower and power dissipation is higher than the original design. If such design is found, it is possible to replace the alternative design with the original when the original design is too slow and does not pass the pre defined specifications. It is the first basic requirement when a circuit is taken as a candidate for the proposed technique.

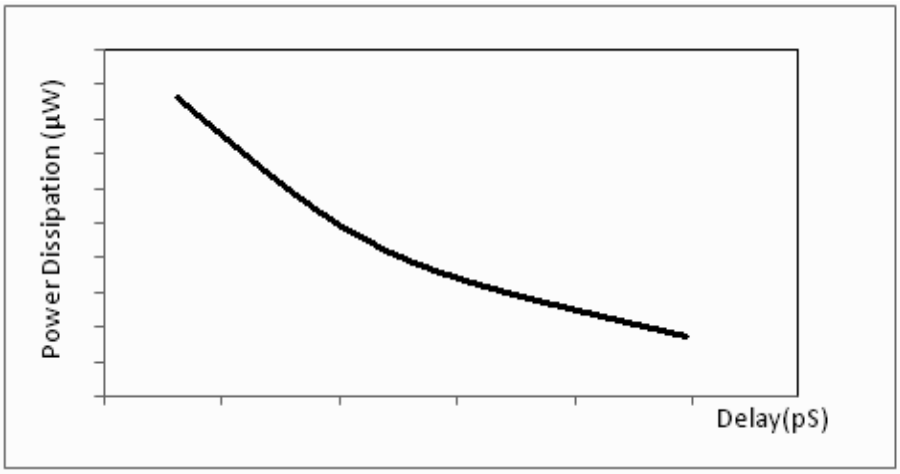

Figure 6-1: A non-linear delay vs. power dissipation behavior

We believe that in order to give better results and explanations of how practical this technique could be in the real manufacturing world, more research needs to be done in:

1. Finding CMOS circuits with the behavior that was shown in Figure 6-1. To improve the production yield of a device with such characteristics, an alternative design is to be found with the same delay-power dissipation behavior, but faster and more dissipative. Figure 6-2 shows an ideal delay-power dissipation spread of two different designs of a CMOS device. A higher gradient of the graph will cause a larger yield improvement.

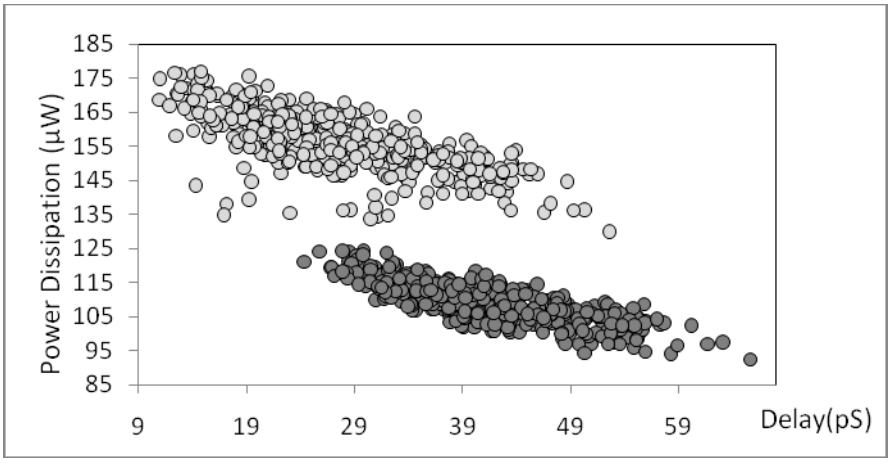

Figure 6-2: Example of Delay vs. Power Dissipation of two designs of one CMOS device

2. Bringing the modified designs into more complicated and advance circuits to verify if the mid-fabrication test based mask selection technique will improve the behavior of the final output.
3. Studying the impact of process variations on the critical path of widely and commonly used devices such as memories, and do research on finding alternative designs for them.

In this thesis, all the presented results and values are related to CMOS90NM technology. As the CMOS technology is getting smaller, it is best to use smaller technologies libraries and softwares for further studies.

References

[1] S.Natarajan, M.A. Breuer and S.K. Gupta, ”Process Variations and their Impact on Circuit Operation “, *Defect and Faul t Tolerance in VLSI Systems* , Nov 1998, pp. 73-81, Presented at 1998 IEEE International Symposium, Austin, TX, USA.

[2] News Release*, “Intel First to Demonstrate Working 45nm Chips* “INTEL’s official website, Jan. 25, 2006. [Online], Available: http://www.intel.com/pressroom/archive/releases/2006/20060125comp.htm, [July 2009].

[3] KOCH Lab, California Institute of Technology, Division of Biology, “Introduction to neuromorphic and bio-inspired mixed-signal VLSI, Mismatch and compensation techniques”, CNS182.[Online]. Available: http://www.klab.caltech.edu/~shih/10-mismatch.pdf, [Accessed: June 21, 2010].

[4] S.Yadala,“Process-Voltage-Temperature (PVT) Variations and Static Timing Analysis”.[Online]. Available: http://asic-soc.blogspot.com/2008/03/process-variations-and-static-timing.html, [Accessed: June 21, 2010].

[5] J.D. Johnson, “Managing Variability in the Semiconductor Supply Chain”, Engineering Systems Division, Massachusetts Institute of Technology, 2005.

[6] E.S. Fetzer “Using Adaptive Circuits to Mitigate Process Variations in a Microprocessor Design”, *Process Variation and Stochastic Design and Test* , Intel, Nov 2006, pp.476-483.

[7] Well Proximity Effect Model, BSIM4.5.0 Manual, UC Berkeley, 2005. [Online].Available: http://www.eigroup.org/cmc/minutes/2q05_presentations/bsim450_chap14n_wpe.pdf.

[8] D.Hillman,“ Integrated Power Management, Leakage Control and Process Compensation Technology for Advanced Processes”, Transmeta Corp., Santa Clara, CA, USA, 2010.

[9] Advanced Motion Controls, 2005.[Online].Available : http://www.a-m-c.com/content/m101/industry_highlight/semiconductor/index.html.

[10] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter Variation and Impact on Circuits and Microarchitecture", *in Proc. of Design Automation Conf.*, June 2003, pp. 338-342.

[11] A.Aryanpour, Glenn E.R. Cowan, "A Circuit Design and Fabrication Approach to Address Global Process Variation", *IEEE International Midwest Symposium on Circuits and Systems, Cancun, Mexico, Aug 2009.*

[12] A. Das, S. Ozdemir, G.Memik, J. Zambreno and A. Choudhary*,* " Mitigating the Effects of Process Variations*", Architectural Approaches for Improving Batch Performance*, Presented in Workshop on Architectural Support for Gigascale Integration (ASGI), Northwestern University. Evanston, IL, USA, 2007.

[13] F.M. Bufler, Y.Asahi, H. Yoshimura, C. Zechner, A. Schenk, and W. Fichtner, "Monte Carlo Simulation and Measurement of Nanoscale n-MOSFETs*",* *IEEE Transactions on electron devices*, Vol. 50, No. 2, FEBRUARY 2003.

[14] M. Nourani, A. Radhakrishnan, " Testing On-Die Process Variation in Nanometer VLSI", University of Texas at Dallas and Texas Instruments , Nov. 2006.

[15] K. Mai, M. Tuckermann , " SPC Based In-line Reticle Monitoring on Product Wafers ", *IEEE/SEMI Advance Semiconducto r Manufacturing Conference,* Germany, 2005, pp. 184-188.

[16] A.Datta, S.Bhunia, J.H.Choi, S.Mukhopadhyay and K.Roy, "Profit Aware Circuit Design Under Process Variations Considering Speed Binning"*, IEEE Transactions on very Large Scale Integration (vlsi) systems,* VOL. 16, NO. 7, July 2008.

[17] V.Zolotov , C.Visweswariah and J.Xiong, "Voltage Binning Under Process Variation", *IEEE/ACM International Conference , Computer-Aided Design - Digest of Technical Papers, 2009. ICCAD 2009.*,

[18] A.Mutlu, K.J. Le, M.Celik, D.S.Tsien, G.Shyu and L.C.Yeh, "An Exploratory Study on Statistical Timing Analysis and Parametric Yield Optimization" , *Proceedings of the 8th International Symposium on Quality Electronic Design*, pp. 677-684, 2007.

[19] A.Agarwal, D.Blaauw and V.Zolotov, "Statistical Clock Skew Analysis Considering Intra-Die Process Variations", *IEEE Transactions on Computer-Aided Design*, vol. 23, no. 8, pp. 1231-1242, Aug 2004.

[20] B. Nikolic´, V.G. Oklobd˘zija, V. Stojanovic´,W. Jia, J.K.S.Chiu and M. M.T. Leung, "Improved Sense-Amplifier-Based Flip-Flop", *Design and Measurements, IEEE Journal of Solid-State Circuit,* Vol. 35, NO. 6, June 2000.

[21] M. Hansson and A. Alvandpour, "Comparative Analysis of Process Variation Impact on Flip-Flop Power-Performance," *IEEE Int. Symp. on Circuits and Systems,* May 2007, pp. 3744 - 3747