

Variability Analysis and Yield Optimization in Deep-Submicron Mixed-Signal Circuits

Katerina Papadopoulou



Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-10

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-10.html>

May 1, 2019

Copyright © 2019, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

**Variability Analysis and Yield Optimization in Deep-Submicron Mixed-Signal
Circuits**

by

Aikaterini Papadopoulou

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

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Borivoje Nikolić, Chair

Professor Tsu-Jae King Liu

Professor Jasmina Vujić

Spring 2017

The dissertation of Aikaterini Papadopoulou is approved.

Chair

Date

Date

Date

University of California, Berkeley
Spring 2017

Variability Analysis and Yield Optimization in Deep-Submicron Mixed-Signal Circuits

Copyright © 2017

by

Aikaterini Papadopoulou

Abstract

Variability Analysis and Yield Optimization in Deep-Submicron Mixed-Signal Circuits

by

Aikaterini Papadopoulou

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Borivoje Nikolić, Chair

Scaling of CMOS technology into the deep-submicron regime has made superior device performance and high density possible. However, achieving extreme performance is often limited by an increase in variability due to the aggressive shrinking of dimensions. As variability continues to rise, statistical modeling methods like Monte-Carlo and corner simulation that have been extensively used to predict circuit yield in the past, now become insufficient. It is now evident that models can no longer be developed solely by device variation measurements, but they need to exhibit flexibility and tunability to a specific design.

In this work, we address the problem of rising variability and insufficient variability modeling in two ways. Firstly, by characterizing variability in a deeply-scaled technology node, and secondly by developing a methodology for simple, fast model tuning for design-specific yield optimization.

Technology characterization is achieved by designing a set of dedicated test structures in a 28nm FDSOI technology. Test structures include both device characterization as well as high-speed comparator characterization, and focus on design-dependent, layout-dependent and topology-dependent sources of variation. Worst-case measured within-die device variation goes up to 11 % while a 46% of current variation is measured across different dies. Layout-dependent systematic effects do appear to be significant in this technology. Several comparator topologies are also measured, showing a direct link between comparator sensitivity and measured offset.

Yield optimization is achieved by model customization to a specific design. A methodology that uses backward propagation of variance and sparse regression techniques is developed in order to achieve this. The methodology is shown to have the ability to tune models to variability structure measurements, decreasing the estimated prediction error from 30% to <4%.

Σαν έξαφνα, ώρα μεσάνυχτ', ακουσθεί
αόρατος θιάσος να περνά
με μουσικές εξάισιες, με φωνές -
την τύχη σου που ενδίδει πια, τα έργα σου
που βγήκαν όλα πλάνες, μη ανωφέλετα θρηνήσεις.
Σαν έτοιμος από καιρό, σα θαρραλέος,
αποχαιρέτα την, την Αλεξάνδρεια που φεύγει.
Προ πάντων να μη γελασθείς, μην πεις πως ήταν
ένα όνειρο, πως απατήθηκεν η ακοή σου
μάταιες ελπίδες τέτοιες μην καταδεχθείς.
Σαν έτοιμος από καιρό, σα θαρραλέος,
σαν που ταιριάζει σε που αξιώθηκες μια τέτοια πόλι,
πλησίασε σταθερά προς το παράθυρο,
κι άκουσε με συγκίνησιν, αλλ' όχι
με των δειλών τα παρακάλια και παράπονα,
ως τελευταία απόλαυσι τους ήχους,
τα εξάισια όργανα του θιάσου,
κι αποχαιρέτα την, την Αλεξάνδρεια που χάνεις.

[Κ. Π. Καβάφης]

Contents

Contents	ii
List of Figures	v
List of Tables	ix
Acknowledgements	x
1 Introduction	1
1.1 Motivation	1
1.2 Research goal	3
1.3 Thesis organisation	4
2 Variability and statistical modeling	5
2.1 Variability sources and effects	5
2.2 Traditional statistical modeling techniques	7
2.3 Advances in statistical modeling	9
2.3.1 Yield optimization through circuit performance modeling	10
2.3.2 Yield optimization through device performance modeling	12
2.4 Summary	13
3 Customized model generation	14
3.1 Customized models	14
3.1.1 Base models	14
3.1.2 Base model tuning	16
3.1.3 Parameter screening and system constraints	18

3.2	Application examples	23
3.2.1	Customized models for device characterization	23
3.2.2	Customized models for comparator characterization	28
3.3	Summary	34
4	Test structure design for data extraction	42
4.1	Overview of test structures and goals	42
4.2	Test structure design	43
4.2.1	Offset characterization array	45
4.2.2	Impulse sensitivity function characterization array	46
4.2.3	Device characterization structures	48
4.3	Characterization and comparison of comparator topologies	49
4.4	Characterization of random and systematic variability	51
4.4.1	Random variability test structures	52
4.4.2	Systematic variability test structures	54
4.5	Chip overview and testing	55
4.6	Summary	57
5	Experimental evaluation of customized models	58
5.1	Characterization of device variability in 28nm FDSOI	58
5.1.1	Matching performance	59
5.1.2	Within-die variability	60
5.1.3	Die-to-die variability	62
5.2	Characterization of comparator variability in 28nm FDSOI	64
5.2.1	Layout-related effects of variability	67
5.3	Variability-aware comparator design in deeply-scaled technologies	69
5.3.1	Variation-driven comparator topology selection	69
5.3.2	Variation-driven comparator clocking scheme selection	73
5.4	Design-specific model customization	74
5.4.1	Base model performance	74
5.4.2	Model customization overview	75
5.4.3	Customized model performance	77

5.5	Summary	82
6	Conclusions	83
6.1	Key contributions	83
6.2	Future work	84
6.3	Conclusions	85
	Bibliography	86

List of Figures

1.1	Effects of scaling in digital circuit design.	3
1.2	Effects of scaling in mixed-signal circuit design.	4
2.1	Corners and Monte-Carlo scatter plot of NMOS and PMOS V_{TH} (normalized).	8
2.2	Illustration of design centering.	9
2.3	Block diagram of circuit design process and yield optimization through parameteric performance modeling.	11
2.4	Simplified illustration of a customized corner.	12
3.1	Illustration of principal component analysis using a 3-dimensional space as an example.	15
3.2	Ridge regression, produced with artificially generated data. Coefficients are kept small and converge to their final values.	20
3.3	LASSO solution, produced with artificially generated data. For decreasing t , more and more coefficients are forced to zero.	21
3.4	Elastic net solution with $\rho = 0.5$, produced with artificially generated data. The solution combines the properties of ridge regression and LASSO, depending upon the value of ρ	22
3.5	Root mean square error for ridge, LASSO and elastic net regression.	22
3.6	Two-dimensional contour plots of the ridge (black), LASSO(light grey) and elastic net (dark grey) penalty terms for a hypothetical 2-parameter system. The dotted line indicates the valid solution space.	23
3.7	Simulation circuit for IV curve extraction of NMOS and PMOS.	24
3.8	Scatter plots of a subset of the parameters used.	25
3.9	Comparison of extracted histograms (a) without adding physical constraints and (b) with physical constraints. In both cases blue denotes the original model and red the customized model. All data are normalized.	26

3.10	Comparison of extracted histograms (a) without adding physical constraints and (b) with physical constraints. In both cases blue denotes the original model and red the customized model. All data are normalized.	27
3.11	Output histograms (top) and quantile-quantile plot (bottom) comparison. In both cases blue denotes the original model and red the customized model. All data are normalized.	28
3.12	Simulated probability of failure using Monte-Carlo and importance sampling.	29
3.13	Circuit of a StrongARM latch including pre-charge devices.	29
3.14	Simulation setup for ISF extraction of a sampler.	31
3.15	Impulse response of the sampler for various clock fall times, compared to an ideal response. Larger clock fall times increase the aperture width.	37
3.16	Extracted spectrum for various clock fall times, compared to an ideal response. Increased aperture width limits the bandwidth of the sampler.	37
3.17	Simulation setup for offset measurement of a comparator.	38
3.18	Simulation setup for impulse sensitivity function measurement of a comparator.	38
3.19	Root-mean square error of elastic net regression for comparator offset.	40
3.20	Comparison of offset prediction between original model and customized model.	40
3.21	Comparison of the distributions and QQ plots between original model and customized model.	41
4.1	High-level chip block diagram, featuring the offset, ISF and comparator arrays and the on-chip memory.	43
4.2	Simplified schematic of the complete chip.	44
4.3	One column of the offset measurement array.	45
4.4	Example of a cumulative probability function of a single comparator, after noise averaging.	46
4.5	One column of the ISF measurement array.	48
4.6	Input signal path consisting of SMA input, PCB trace, bondwire, chip input pad, probepad, wires and the comparator.	49
4.7	Schematic design of the transistor array for Kelvin-sensing measurement of a large number of DUTs.	50
4.8	Monte-Carlo simulation showing the effect of reverse gate bias for leakage reduction.	51
4.9	Part of the device characterization array layout	54
4.10	Layout picture of the complete chip.	56

4.11	Test board setup, including motherboard, daughterboard and Opal Kelly Shuttle LX1 board.	57
5.1	Die photo of the chip, showing the various blocks described in Section 4.2.	59
5.2	The constant-current voltage threshold extraction method.	60
5.3	Distributions of measured V_{TH} and I_{ON} for a 310nm/30nm device across different dies, compared to the simulated distributions.	61
5.4	Pelgrom plot using simulated and measured data.	62
5.5	Measured WID $3\sigma/\mu$ variation for a 310nm/30nm device across different dies. The solid line shows the average WID variation, while the dashed lines mark the best and worst WID variation measured.	63
5.6	Measured WID $3\sigma/\mu$ variation from all dies across different layouts. The blue dots correspond to the mean value.	64
5.7	Comparison of measured V_{TH} distributions of fastest and slowest die for different layouts. Data are normalized to the mean V_{TH} value of the reference device N0 in the slowest die.	65
5.8	Comparison of measured I_{ON} distributions of fastest and slowest die for different layouts. Data are normalized to the mean I_{ON} value of the reference device N0 in the slowest die.	66
5.9	(a) Noise cumulative distribution function of a single SA comparator instance and (b) offset distribution of SA comparator within a die	67
5.10	Colormaps of measured WID comparator offsets for different comparator layouts.	68
5.11	Colormaps of measured comparator offsets for different comparator topologies within a die.	70
5.12	Comparison of all single clock-phase topologies.	70
5.13	Illustration of sensitivities for each device of each comparator type. Colors correspond to the maximum sensitivity of the device.	71
5.14	Scatter plots of simulated I_{ON} and V_{TH} with respect to model parameters. Data are normalized to zero mean and unit variance. Each plot title shows the corresponding Pearson correlation coefficient.	72
5.15	Comparison of comparator offset in all two-stage topologies.	73
5.16	Mean absolute percent prediction error across different supply voltages for all topologies.	75
5.17	(a) Conventional design process steps and (b) proposed design process steps.	76
5.18	Strong-arm latch used for design centering. The greyed-out devices were not assigned statistical parameters.	77

5.19	Comparison of (a) offset distribution and (b) corresponding quantile-quantile plot for the strong-arm comparator at nominal supply voltage.	78
5.20	Comparison of (a) offset distribution and (b) corresponding quantile-quantile plot for the strong-arm comparator at scaled supply voltage.	79
5.21	Comparison of offset standard deviation across (a) supply voltage and (b) input common-mode, when the customized model is calibrated at each voltage step.	79
5.22	Comparison of offset standard deviation across supply voltage, when the customized model is calibrated only at nominal supply.	80
5.23	Comparison of offset standard deviation across supply voltage when the customized model is calibrated at each voltage step for various comparator topologies.	81
5.24	Comparison of mean absolute percent prediction error across different supply voltages for all topologies.	82

List of Tables

3.1	First step of parameter selection	19
3.2	FDSOI 28nm PSP model statistical parameters for Monte-Carlo (MC) and Fixed-Corner (FC) simulation	35
3.3	Percent sensitivities to parameters	36
3.4	Parameter standard deviations in original model (OM) and customized model (CM)	36
3.5	Proposed search algorithm	36
3.6	Comparator offset sensitivity to statistical parameters for input devices and cross-coupled pair devices at nominal supply voltage.	39
4.1	Time step $\Delta\tau$ for various combinations of f_{ref} and N for ISF characterization.	47
4.2	Comparator topologies included in the test-chip	52
4.3	Layouts and geometries of devices under test	53
5.1	Comparator layout configurations	68
5.2	Percent offset shifts of comparators with respect to SA0	69
5.3	Comparison of simulated and measured variation for a 310nm/30nm device	74
5.4	Comparator percent offset sensitivity to statistical parameters	78
5.5	Mean absolute prediction error comparison at nominal supply voltage	81

Acknowledgements

I wish to acknowledge the contributions of the students, faculty and sponsors of the Berkeley Wireless Research Center, wafer fabrication donation of STMicroelectronics, and the support of the Center for Circuit and System Solutions (C2S2) Focus Center, funded under the Focus Center Research Program, a Semiconductor Research Corporation (SRC) program.

More specifically, I would like to acknowledge Brian Zimmer, Amy Whitecombe and Vladimir Milovanović for their contributions to this work through discussions and assistance in circuit and board design. I would like to thank James Dunn for doing the layout of our boards, as well as Brian Richards and Ubirata Chavez Coelho for all technical support. Last but not least I would like to thank my advisor Borivoje Nikolić for his patience and support throughout my time in grad school.

On a personal level, I would like to thank Evi Kitsou and Dimitra Bavelou, as well as the "good" ones, Ioannis Protonotarios, Iris Safaka, Marina Kordoni and Antonis Moros for their endless love and support. Infinitely many thanks to Mike Lorek because his heart is made of gold, and also to Georgia Gkioxari for being a friend and a sister.

Chapter 1

Introduction

Scaling of CMOS technology into the deep-submicron regime has brought on significant changes in all aspects of circuit design, from modeling and simulation to manufacturing. Along with scaling, the introduction of new technologies like ultra-thin body devices and FinFETS has introduced new challenges as well.

Over the past few years research focus has shifted towards optimizing these technologies at deeply-scaled nodes, in order to enable high yield design. This thesis attempts to characterize and quantify device and circuit variability with a focus on ultra-thin body silicon-on-insulator (SOI) technology, as well as present a modeling optimization methodology for mixed-signal circuits. This is achieved through the design, measurement and analysis of a 28nm fully-depleted SOI testchip. In this introductory chapter, the motivation and research goals will be discussed and an overview of the thesis will be given.

1.1 Motivation

In the recent years, rapid technology developments in the metal-oxide-semiconductor industry have lead to dimensional and functional scaling of CMOS processes down to the sub-20nm regime. Transistor gates are projected to scale down to less than 12nm by 2020 [1], resulting in significant changes in all stages of circuit design process, from device manufacturing up to product design.

As a result of aggressive technology scaling, the spectrum of applications has now broadened and improved. Superior device performance and high density have given new perspective to circuit design, ranging from digital processors and memory to high-speed analog front-ends. Figure 1.1 shows the effects of scaling for digital processors and memory, based on data published by Intel [2], illustrating how digital circuit design has been smoothly following Moore's Law over the past years. Figure 1.2 shows the effects of scaling in mixed-

signal design, using recently published time-interleaved successive approximation register (TI-SAR) analog-to-digital converters as a representative circuit [3]. Although analog and mixed-signal circuits do not scale as well as digital, the benefits are obvious as the integrated circuit community has been consistently pushing the boundaries of performance, demonstrating multi-GS/s designs over the past few years.

While achieving extreme performance is now possible, it is often limited by an increase in variability due to shrinking of dimensions into the deep submicron regime. Both within-die and die-to-die variations are shown to increase when scaling from 90nm to 45nm [4] and are expected to increase even more as devices scale to sub-10nm. Traditional sources of variation, like random dopant fluctuation (RDF), line-edge roughness (LER) and gate work-function variation (WFV) [5] become more pronounced as variances fail to track scaling of the mean, which makes any slight deviation from the nominal value to have an amplified effect on transistor electrical performance. Additionally, with the introduction of new technologies and materials, new sources of variation arise like silicon thickness (T_{si}) variation, which is present in thin-body devices, or fin width variation which affects FinFET performance. Device variability is increasingly becoming a bottleneck in achieving high-performing and high-yielding circuits, and therefore it is now a necessity to address it in all stages of circuit design, from manufacturing to device modeling to final design.

Traditionally, device models introduced a large set of model parameters, used to predict typical device performance. As variability started to rise, statistical modeling methods were developed to help designers improve yield and, therefore, decrease cost.

Monte-Carlo simulation is one of the most common ways to account for variability in device models, and it involves extracting variation data from devices, and assigning distributions to the model parameters such that the simulated device performance tracks the measured one. Another commonly used method for yield prediction is corner modeling. Similarly, variation data are extracted from a set of devices. Instead of distributions, the model parameters are assigned deterministic shifts to their nominal value, such that the device performance is pushed n standard deviations away from its mean value. Different combinations of those shifts push the device into different points in the design space, known as corners. As variability rises, more corners become available in new models, including corners for analog design, based on device I-V curves, and corners for digital design, typically based on delay-chain measurements.

Although corner modeling and MC simulation have been extensively used to predict circuit yield at the design stage in the past, extreme device scaling makes the nature of variability much more complex now, and renders these two techniques insufficient. As relating device-level variation to the circuit-level variation in an efficient and accurate manner becomes more challenging, research increasingly focuses on improving statistical modeling techniques [6, 7]. It is now evident that a vertical approach must be taken in dealing with variability; models can no longer be developed solely by device variation measurements, but they need to exhibit flexibility and tunability to a specific design. Accurate, design-specific, high-yielding statistical modeling is still an open problem.

Overall, device scaling delivers both the promise for exceptional circuit performance and

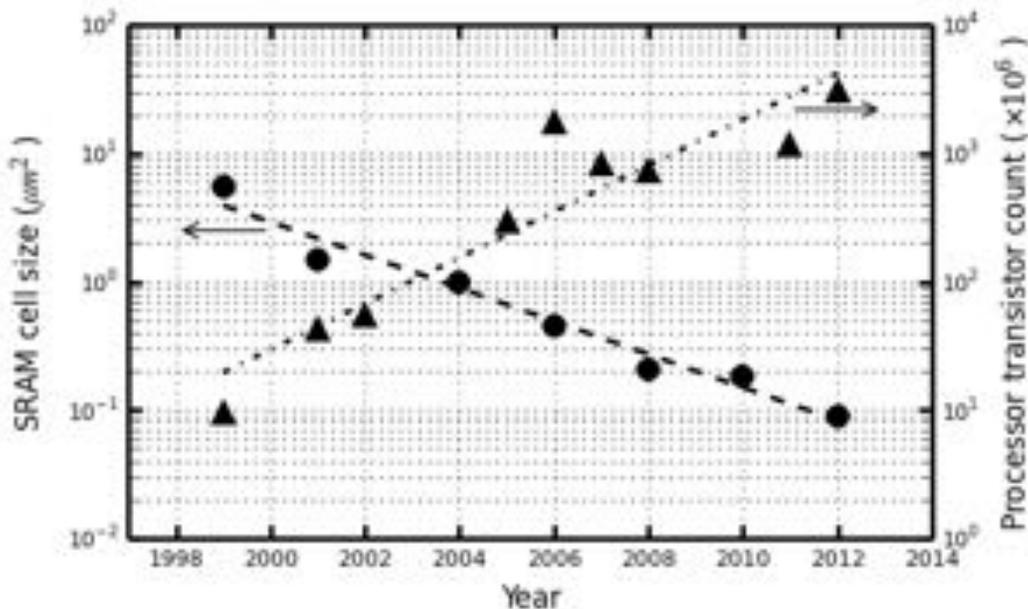


Figure 1.1: Effects of scaling in digital circuit design.

the threat of lower yield and increased cost. As the device manufacturing technology is facing new challenges, research in CMOS device variability as well as in modeling of its effects has the potential to determine the future steps of circuit design.

1.2 Research goal

As devices are scaling in the deep-submicron regime variability rises and becomes more complex, making traditional statistical modeling insufficient. Circuit designers need to account for design-specific effects of variability in order to accurately optimize for yield, which normally requires some statistical modeling expertise. The goal of this work is twofold:

1. To investigate variability in mixed-signal circuit design by characterizing design-dependent, layout-dependent and topology-dependent sources of variation.
2. To present a methodology for simple, fast model tuning for design-specific yield optimization, that is accessible to the circuit designer. The methodology utilizes backpropagation of variance and convex optimization techniques to customize existing statistical model cards to a given design.

Achieving those two goals can shorten the design manufacturing process by providing

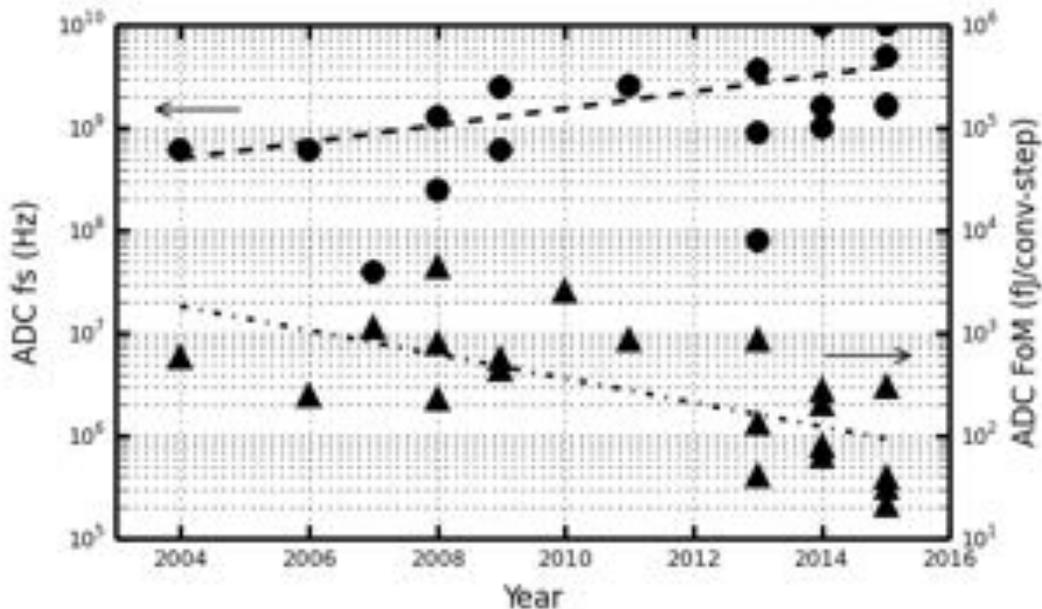


Figure 1.2: Effects of scaling in mixed-signal circuit design.

guidelines for robust mixed-signal circuit design and improving the yield prediction of models at an early design stage.

1.3 Thesis organisation

The present dissertation is organized as follows. In Chapter 2 we present a more in-depth look into the device variability by summarizing its different types and their characteristics. Additionally, we discuss in more depth the traditional statistical modeling techniques and classify more advanced methods of yield optimization to two categories; those that utilize circuit performance modeling and those that utilize device performance modeling techniques. Chapter 3 presents a methodology for customized, design-specific model generation that targets at improving existing models. We present both the mathematical background as well as simulation examples of the proposed methodology. Chapter 4 discusses the necessary test structures for data extraction that will enable the use of custom models. Finally, measurement data and corresponding customized models are presented in Chapter 5.

Chapter 2

Variability and statistical modeling

As outlined in Chapter 1, increasing variability is becoming a bottleneck for both digital and analog circuit design. In order to understand it better, it is necessary to take an analytical approach and understand the different types of variability as well as the existent solutions. This is the goal of the current chapter. In Section 2.1, variability is defined and categorized based on its sources and general characteristics. In Sections 2.2 and 2.3, traditional and state-of-the-art statistical modeling techniques are presented and categorized based on their modeling approach, before the chapter concludes with a brief summary in Section 2.4.

2.1 Variability sources and effects

In order to fully understand the need for yield optimization in deeply-scaled device modeling, it is crucial to know some basic information about variability and its effects on circuit design, performance and cost. Variability describes any deviation of device performance from its designed or typical value, which may be caused by known or unknown reasons. Such reasons may be environmental, such as supply and temperature variation, or physical, such as process variation or atomic-scale effects [8, 9].

Physical variations can be further classified in many different ways, depending on their nature, their sources or their spatial and temporal characteristics. In [10] two main categories are used; intrinsic and extrinsic. Intrinsic variations are those that originate from atomic-scale effects, like quantum-mechanical effects and statistical variation in dopant profiles and particles. Extrinsic variations are mainly attributed to any shift in the process conditions, which causes parameter fluctuation usually with some spatial correlation.

With respect to spatial characteristics, process variations can be further classified to:

1. Lot-to-lot
2. Wafer-to-wafer, within a lot
3. Die-to-die, within a wafer
4. Within-die

Within-die variations are defined by parameters that vary significantly over distances smaller than the dimension of a die. Parameters that vary gradually across a wafer cause die-to-die variations. Wafer-to-wafer variations cause different wafers to have different properties. In a typical design methodology, designs are made to satisfy the worst case corners which consist of the total within-die and die-to-die variations.

Finally, another useful classification divides variations to random and systematic. Random variations are a result of stochastic natural processes that are unpredictable and affect each device in an arbitrary way. Systematic variations are deterministic shifts in device parameters that are better understood and usually more easily modeled.

Systematic variations have their sources largely in the manufacturing process and its different steps [11]. The implant and annealing process cause a different number of dopants to be positioned in different parts of the wafer. Oxide thickness variations are caused by non uniformity in the process of oxide growth. Non-uniform annealing temperature can cause further variation in the threshold voltage, while strain and stress can affect carrier mobility in a systematic fashion. Finally, lithography and etching effects induce variation in critical device sizes, like channel length and width, which are much narrower than the light wavelength used to print them.

Random variation sources lie in atomic-level effects like random dopant fluctuation (RDF) and line-edge roughness (LER). These variations get significantly worse with scaling, as intrinsic parameter fluctuations introduced by the discreteness of charge start to dominate. As channel length scales, less dopants are deposited in the channel for a fixed doping density. With less averaging into play, these discrete dopants start causing potential fluctuations due to their random distribution. The effect is called RDF and manifests itself primarily as threshold voltage variability. At the same time, scaling of dimensions closer and closer to the size of an atom causes previously smooth, continuous and distinct interfaces to become granular and pebbled. Granularity that affects the channel length is called LER. The introduction of thin body devices like FDSOI devices and FinFETS has brought additional sources of variation; granularity that affects the channel thickness in thin body devices is called silicon thickness variation and granularity in the width of the fins in FinFETs is called fin width variation. Other random variability sources that affect all devices are gate work-function variation as well as random strain and stress effects.

Both systematic and random variation are of great interest since they strongly affect circuit performance, yield and cost. More specifically, variability affects the yield of integrated circuits, which is defined as the probability that a circuit will meet the required specifications for a given product. In turn, yield affects the cost in an inversely proportional manner; high circuit yield enables mass production and lowers production cost. It is evident that

properly defining the maximum performance margins for a circuit and a certain yield can help optimize performance and cost, while overestimation or underestimation of those margins can increase design complexity or compromise yield, respectively. Therefore, variability characterization and accurate yield modeling can enable lowering cost and making new technologies more accessible.

As mentioned in Chapter 1, in response to the increase in variability, statistical modeling methods have been developed to help designers improve circuit yield. Such methods include well-established techniques, such as generating worst-case corners for the model, introducing parameter variations through Monte-Carlo simulation or, in the case of digital design, modeling variations in gate delay and using statistical timing analysis (STA) to monitor how delay propagates through a circuit, as well as more sophisticated modeling techniques which generally seek to improve accuracy of corner modeling and Monte-Carlo simulation. Each method comes with advantages and disadvantages. An overview of the most well-established statistical modeling methods will be given in the following section.

2.2 Traditional statistical modeling techniques

A device model is basically the mapping of a set of input parameters to a set of output variables. For example, for a MOSFET device input parameters may be process parameters, like doping profile, carrier mobility and flatband voltage, and outputs include device threshold voltage and current, under certain operating conditions. In deeply scaled technologies models are typically a combination of equations that are derived through device physics and empirical equations; therefore, parameters can be physical as well as empirical. Adding variability to those parameters translates as device variability, and statistical modeling tackles the problem of determining what variability needs to be added and how, in order to properly represent the device.

Monte-Carlo simulation was introduced in the 1970s to integrated circuits for tolerance analysis [12]. To achieve that, random perturbations are added to model parameters, and the output variables are evaluated. By adding a distribution at the input and observing the output distribution, circuit yield can be calculated. Determining the right variance to add to each parameter is accomplished by measuring process variation data using a large set of devices. Since model parameters are generally correlated and large in quantity, principal component analysis (PCA) is typically used to reduce them to a smaller and more manageable set [13], while in newer models some parameter correlations may be preserved in order to improve accuracy. The disadvantages of Monte-Carlo simulation lay in the complexity and computational inefficiency. More specifically, since the nature of variability is very complex, accurately representing all different types of variability in a model quickly becomes an impossible task. Die-to-die and within-die are generally lumped together, and types of variation that are not well-modeled, like strain and time-dependent effects, are treated as random, which results to overly conservative design margins. Additionally, a large number of simula-

tions is required which makes the task of yield optimization very slow and computationally intense.

A faster and simpler way of accounting for systematic variability in models is corner modeling. The main idea here is to add deterministic shifts to process parameters, such that the output variables are evaluated not only for the typical case but also for cases where variation is present [14, 15]. By pushing the parameters n deviations away from their mean, the design performance can be evaluated at its worst-case corners. As variability rises, more corners become available in new models, including corners for analog design, based on device I-V curves, and corners for digital design, typically based on delay-chain measurements. Some of the corners for a 28nm technology are shown in Figure 2.1. Although corner modeling is fast and computationally efficient, in newer technologies it can result in overly optimistic or pessimistic results, as devices scale and variability becomes harder to track. Figure 2.1 compares the process corners to the results of a Monte-Carlo simulation; it is evident that a circuit that is designed to meet specification in the all corners spans a different design space than a circuit designed using Monte-Carlo, and therefore yields unreliable results. Additionally, precise yield prediction is hard to get, since yield represents individual device characteristics and therefore topology specific and layout-induced effects are ignored.

In order to tackle the insufficiencies of Monte-Carlo and corner modeling, design-specific or performance-aware modeling approaches have gained more popularity in the recent years. The main concept is illustrated in Figure 2.2 and involves identifying the actual design space

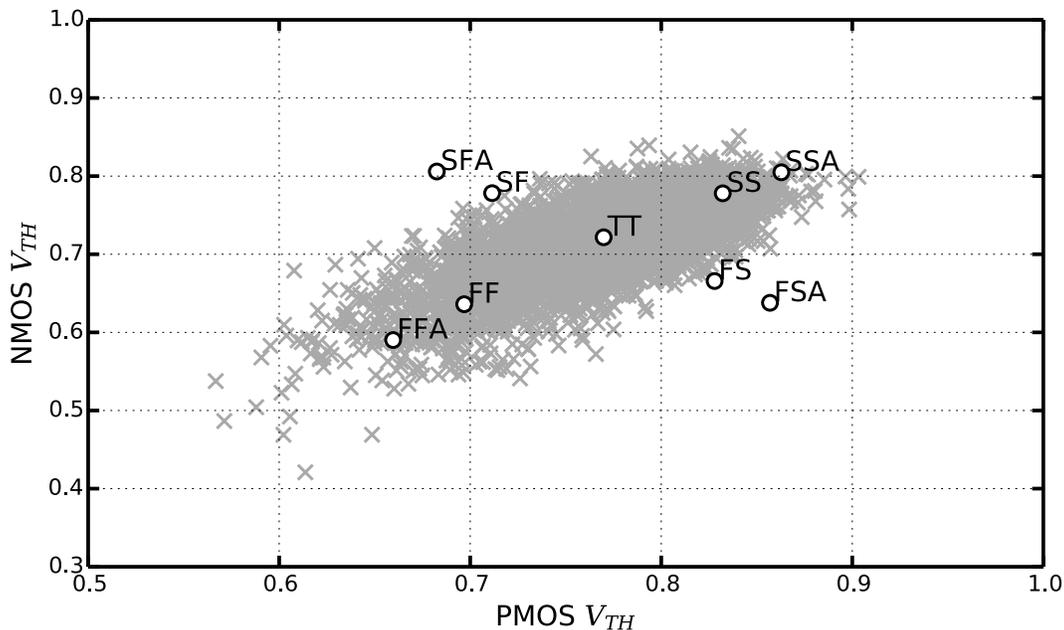


Figure 2.1: Corners and Monte-Carlo scatter plot of NMOS and PMOS V_{TH} (normalized).

and moving the design in order to increase the margin from failure. Some of these modeling techniques will be discussed in the following section.

2.3 Advances in statistical modeling

In the digital domain, variations play an important role since they are directly linked to timing and memory failures. However, digital CAD tools are fairly well developed and digital tool flows are automated, allowing for better design optimization and faster redesign. Along with corner modeling and Monte-Carlo, process variations have been traditionally modeled using static timing analysis (STA), which makes use of calibrated lookup tables for standard cells at multiple technology dependent corners. Traditional STA techniques are much faster than Monte-Carlo simulation, but due to the complex and interacting nature of various sources of variation they have become insufficient for highly scaled technology nodes, which led to the introduction of statistical static timing analysis (SSTA). SSTA computes an upper bound on the distribution of the exact circuit delay, accounting for die-to-die and within-die process variations and their spatial correlations [16–18]. Although more accurate, it has an exponential run time complexity. Methods and algorithms have been proposed in order to reduce runtime, however there remain large obstacles to the widespread use of SSTA in the industry [18, 19].

In analog and mixed signal integrated circuits, the design cycle remains long and error-prone [20]. As a result, although analog circuits are typically only a small fraction of a modern system-on-chip (SoC), their design and yield verification can often be the bottleneck for the

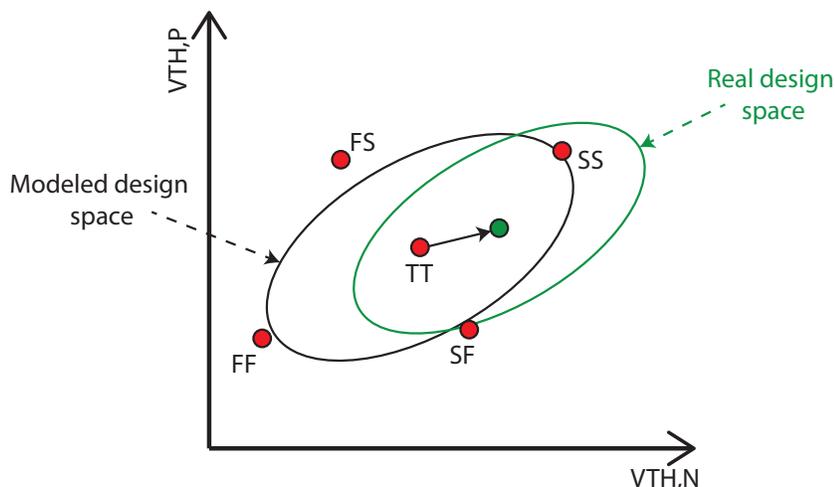


Figure 2.2: Illustration of design centering.

whole system. Recent advances in modeling try to address this problem by introducing yield optimization algorithms applied on circuit performance models or targeting at improving existing device models themselves. We classify these approaches into two broad categories:

1. Approaches that focus on circuit performance modeling
2. Approaches that focus on device performance modeling

These two types of modeling approaches are discussed in detail in the following paragraphs.

2.3.1 Yield optimization through circuit performance modeling

In the area of yield optimization through performance modeling, the main effort is to try to create an accurate parametric model of the circuit and then apply an efficient yield optimization algorithm in order to center the design. The general design procedure is illustrated in Figure 2.3. The first step of the optimization is to generate a parametric response surface model for the circuit that is a function of the design variables and the process variables, using some form of regression technique. This process is also called symbolic modeling. Then a yield optimization algorithm is applied that typically formulates the problem as an optimization problem with respect to the design variables, given the model process variation. Finally the optimization problem is solved by means of convex optimization or geometrical programming. Such performance models are used to speed up circuit sizing: in every iteration of the synthesis procedure, calls to the transistor-level simulator are replaced by evaluations of a suitably constructed parametric model. The model building process is a one-time up-front investment that has to be done only once for each circuit in each technology.

The simplest type of model that can be used in order to estimate yield is a linear approximation of circuit performance [21, 22]. A stochastic approximation algorithm is then applied to the linear model to determine the design parameters that optimize yield. Whereas a simple linear approximation may work well for some circuits, it is not adequate to describe analog or mixed-signal circuits that present non-linearities, especially in deeply scaled technologies. Therefore more efforts have been made over the year to create polynomial and posynomial models. In [23], quadratic-style posynomial performance models for analog circuits are created by fitting a preassumed posynomial equation template to simulation data created according to a design of experiments scheme. The problem with using higher order models is that they are more complex and therefore more challenging to solve for all the necessary fitting coefficients. In order to mitigate that, the authors in [24], who also use quadratic polynomial/posynomial models, suggest a methodology called ROBust Analog Design (ROAD). The methodology applies a projection operator with the goal of obtaining an optimal low-rank model by minimizing the approximation error, therefore achieving to simplify the problem and reduce modeling cost.

Another class of regression techniques borrows ideas from data mining, which focuses on extracting meaningful patterns to large amounts of high-dimensional data [25, 26]. The idea here is to exploit the large amount of simulation points one can get from the simulator in

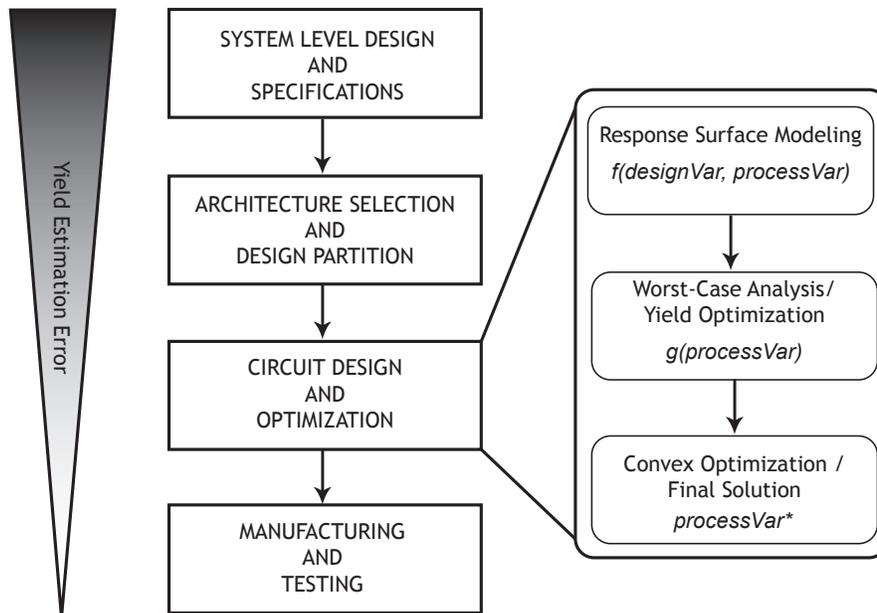


Figure 2.3: Block diagram of circuit design process and yield optimization through parametric performance modeling.

order to train a predictive model. Models including neural networks, boosted neural networks or support vector machines have been used for this purpose. As the aim of symbolic modeling is to use simulation data to generate interpretable mathematical expressions that relate the circuit performances to the design variables, these models generally fail to provide circuit insight.

All the aforementioned symbolic modeling techniques generally use some template for modeling, which means that model selection is restricted to a certain form, like a linear or quadratic form, for example. Another approach is presented in [27, 28]. The method is called Canonical Functional Form Expressions in Evolution (CAFFEINE) and it presents the first template-free model generation approach, i.e. the designer does not have to specify a priori a model template, but the model itself evolves as part of the optimization process.

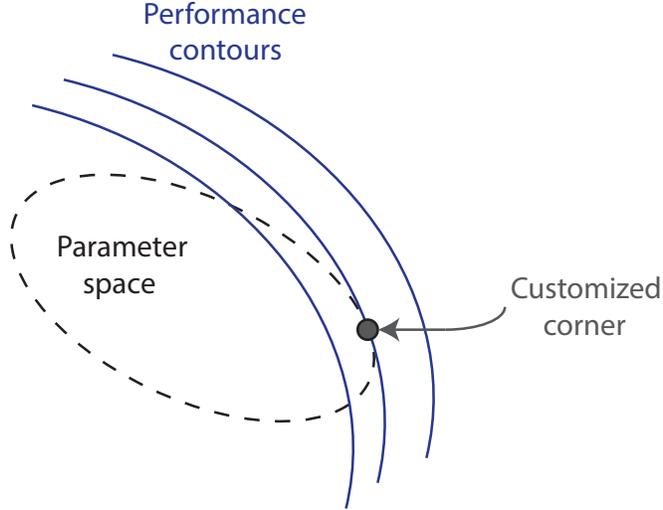


Figure 2.4: Simplified illustration of a customized corner.

2.3.2 Yield optimization through device performance modeling

All the techniques discussed in the previous section target at improving yield prediction by using existing simulation data. The accuracy of those techniques depends strongly on the accuracy of the variation data added to the device models. However, realistic worst case corners and variation data cannot be accurately predicted by IC foundries. The large number of different types of variability makes it impossible to capture, isolate and model all of them. Typically, different types of variation are lumped together in the models, sacrificing accuracy. Along with that, variation data is captured using individual device arrays, therefore ignoring topology-dependent and performance-dependent effects of more complex circuits. For this reason, it is beneficial to discover ways to improve the original device models to accurately represent variability for circuits of interest.

In [29, 30] the authors propose a hierarchical model for process variability. The model addresses both systematic and random variations at wafer, field, die, and device level, and spatial correlation artifacts are captured implicitly. Finally, layout dependent effects are incorporated as an additive component. In [31] the authors incorporate spatial correlation of model parameters into their models to further improve them. In both cases, models become significantly more complex.

Another approach is to create customized or performance aware corners for a given design. A customized corner is defined as the tangential point between performance function contours of the design and the parameter variability ellipsoid, as show in Figure 2.4 [32]. In [33, 34] the authors apply backpropagation of variance in order to evaluate model parameter variances based on measurements. By using the methodology with measurements from specific circuits, it is possible to create performance-driven device models for the circuits of interest. Though a generally efficient model for simple systems, it is dependent on

handpicking parameters and has increased numerical complexity for large systems. In [35] the authors propose a methodology to generate performance-aware corner models. Although promising, the methodology assumes prior knowledge of the variation of some of the physical parameters and is based on simplified model equations.

2.4 Summary

In this chapter, different types of classification of variations were presented; among them, variations were categorized to random and systematic, based on whether they are a result of unpredictable natural processes, or better-understood design conditions. This is one of the most important variation classifications, and it will be used throughout this thesis. Additionally, traditional and advanced modeling techniques were presented, and their advantages and disadvantages were discussed. It was shown that performance-aware device models, also described as custom corners, are possible.

Because of the curse of dimensionality, custom corners are very hard to analytically find and incorporate into the models. Technology scaling however imposes new restrictions to manufacturing resulting in more design rules - for example, gate orientation becomes fixed, poly and metal densities are carefully controlled, gate pitch gets confined to predefined values, transistor dimensions scale by fixed steps. Consequently, the number of practical circuit and layout topologies reduces, making design-specific modeling an increasingly attractive option. Containing the problem to a smaller set of designs and parameters, it is possible to create customized model cards to better predict design yield. In the following chapter, a methodology for creating customized models will be discussed.

Chapter 3

Customized model generation

In this chapter, a methodology for optimizing statistical models customized to a given class of designs is presented. The methodology builds on the existing body of modeling, employs the backpropagation of variance technique to improve the variation assigned to each model variable and is adapted to include existing model parameter correlations. We then formulate the problem as a constrained convex optimization problem, including parameter selection as well as the addition of physical, model-derived constraints by the designer. This enables the creation of customized model cards for robust design of high-performance circuit blocks.

Section 3.1 presents the theoretical background for base models and customized model creation, followed by a set of simulation examples for preliminary model validation, shown in Section 3.2. The chapter concludes with a brief summary of the main points in Section 3.3.

3.1 Customized models

3.1.1 Base models

The first step to model customization is understanding basic concepts of modeling and, especially, of how models treat variations. In general, models describe the device behavior to the circuit simulation program, and so they need to describe a variety of physical effects. For modern devices, a completely theoretical model based on the fundamentals of physics becomes practically intractable. On the other hand, use of a completely empirical model results in a loss of predictive capabilities. A compromise is usually made in developing models for circuit simulation. A combination of physics-based and empirical equations is used. The

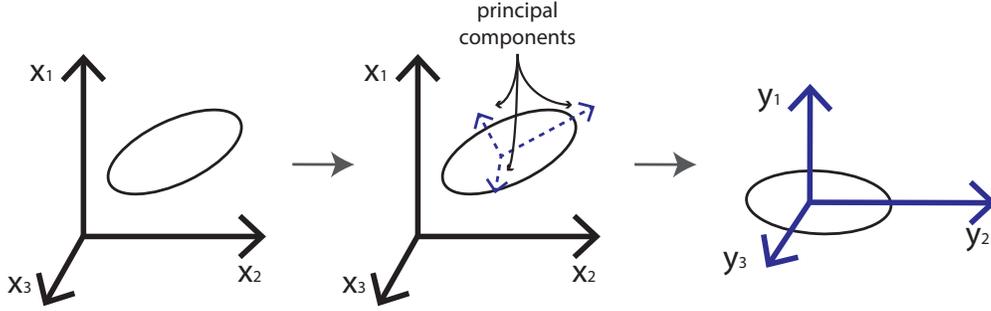


Figure 3.1: Illustration of principal component analysis using a 3-dimensional space as an example.

primary model parameters are closely linked to theory providing an engineering understanding of device physics and serving as a process control aid. The secondary model parameters are partly empirical and help keep the model equations simple. The extraction of optimum model parameter values is necessary to ensure that the device model equations represent the device characteristics closely. Although most models are based on physical theory, there are always some parameters which do not have physically well-defined values, and others for which the physical values do not give the best fit to actual device characteristics. Thus it is generally necessary to extract model parameters from transistor data, obtained from device characterization [36].

In order to include variations in a model, sets of device characterization test structures are designed, typically consisting of a large number of devices under test. Assuming there are M devices under test and N model parameters for each device, then the total number of parameters is a vector of length $d = M \times N$:

$$\mathbf{x} = [x_1 \ x_1 \ \dots \ x_d]^T \quad (3.1)$$

Now if the parameter vector is treated as a vector of random variables (RVs) \mathbf{X} and $\mathbf{X}_0 = E[\mathbf{X}]$ is the vector with the nominal values of the parameters, then $\Delta\mathbf{X} = \mathbf{X} - \mathbf{X}_0$ contains the variations from the nominal values with a zero-mean distribution. $\Delta\mathbf{X}$ is typically modeled as a set of zero-mean jointly normal RVs that are generally correlated. Let Σ be the covariance matrix of $\Delta\mathbf{X}$ and \mathbf{T} an orthogonal matrix such that:

$$\mathbf{T}'\Sigma\mathbf{T} = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d) \quad (3.2)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ are the eigenvalues of Σ . Let $\Delta\mathbf{Y} = \mathbf{T}'\Delta\mathbf{X}$, then it can be shown that the ΔY_j are uncorrelated and $\text{Var}[\Delta Y_j] = \lambda_j$. Then ΔY_j is called the j th principal component of $\Delta\mathbf{X}$ and the process is called principal component analysis (PCA). Intuitively,

starting from a complex set of variables, illustrated as a coordinate system in Figure 3.1, PCA finds a new set of variables orthogonal to each other and then rotates the coordinate system. It can also be shown that it is possible to reduce the dimension of $\Delta\mathbf{X}$ with small loss of information by essentially dropping the components that matter least. A more rigorous mathematical analysis of PCA and dimensionality reduction can be found in [37].

Principal component analysis is widely used in statistical models for two reasons. Firstly, it enables transforming a correlated set of normally distributed parameters to a set of mutually independent parameters. Secondly, it enables reducing the set of parameters to a smaller, more manageable set without significant loss of accuracy. From now on, when referring to base model parameters, we will refer to this reduced set of mutually independent parameters.

3.1.2 Base model tuning

A circuit simulator requires three types of information to specify a transistor model completely: fundamental constants, operating conditions and model parameters. We denote the fundamental physical constants as a vector \mathbf{c} , containing constants such as electronic charge. Those are defined inside the circuit simulation program. The operating conditions define the circumstances under which the model equations are to be evaluated, and we will denote them as a vector \mathbf{x} . The operating conditions are normally the transistor's bias voltages, temperature etc. Finally, the third set of information required is the set of model parameters for each device in the circuit, denoted here as \mathbf{p} . The model output vector \mathbf{y} is a function of \mathbf{x} , \mathbf{p} and \mathbf{c} .

$$\mathbf{x} = f(\mathbf{x}, \mathbf{p}, \mathbf{c}) \quad (3.3)$$

For a given design biased around a bias point $\mathbf{x} = \mathbf{x}^*$, Equation 3.3 can be re-written as $\mathbf{y} = f(\mathbf{x}^*, \mathbf{p}, \mathbf{c}) = f(\mathbf{p})$, where $\mathbf{y} = [y_1, \dots, y_m]^T$ is the output vector of the circuit and $\mathbf{p} = [p_1, \dots, p_n]^T$ is the parameter vector of the given base model. In order to incorporate variability to the models, each of the parameters in \mathbf{p} is treated as an independent random variable (RV) P_i and assigned a normal distribution with mean μ_i and standard deviation σ_i , $i = 1, \dots, n$. Let \mathbf{p}_0 denote the vector of mean values and $\mathbf{P} = \mathbf{p}_0 + \Delta\mathbf{P}$, we have:

$$\mathbf{Y} = f(\mathbf{P}) : \Re^n \rightarrow \Re^m \quad (3.4)$$

The Taylor expansion of (3.4) around the nominal point \mathbf{p}_0 is:

$$\mathbf{Y} = f(\mathbf{p}_0) + \mathbf{J}(\mathbf{P} - \mathbf{p}_0) + \frac{1}{2!}\mathbf{G}(\mathbf{P} - \mathbf{p}_0) \otimes (\mathbf{P} - \mathbf{p}_0) + h.o.t. \quad (3.5)$$

where $\mathbf{J} \in \Re^{m \times n}$ is the partial derivatives (Jacobian) matrix, defined in Equation 3.6, $\mathbf{G} \in \Re^{m \times n \times n}$ is the second-order partial derivatives matrix, defined in Equation 3.7, and \otimes symbolizes the Kronecker product.

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_{n-1}} & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad (3.6)$$

$$\mathbf{G} = \left[\frac{\partial \mathbf{J}}{\partial x_1} \quad \frac{\partial \mathbf{J}}{\partial x_2} \quad \cdots \quad \frac{\partial \mathbf{J}}{\partial x_n} \right] \text{ where } \frac{\partial \mathbf{J}}{\partial x_p} = \begin{bmatrix} \frac{\partial^2 f_1}{\partial x_p \partial x_1} & \frac{\partial^2 f_1}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f_1}{\partial x_p \partial x_n} \\ \frac{\partial^2 f_2}{\partial x_p \partial x_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f_m}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 f_m}{\partial x_p \partial x_{n-1}} & \frac{\partial^2 f_m}{\partial x_p \partial x_n} \end{bmatrix} \quad (3.7)$$

Assuming small perturbations we can remove the high-order terms from Equation 3.5 and get:

$$\begin{aligned} \mathbf{Y} &= f(\mathbf{p}_0) + \mathbf{J}(\mathbf{P} - \mathbf{p}_0) = f(\mathbf{P}_0) + \mathbf{J} \cdot \Delta \mathbf{P} \Rightarrow \\ \mathbf{Y} - f(\mathbf{P}_0) &= \mathbf{J} \cdot \Delta \mathbf{P} \Rightarrow \\ \Delta \mathbf{Y} &= \mathbf{J} \cdot \Delta \mathbf{P} \end{aligned} \quad (3.8)$$

Next, we transform Equation 3.8 as follows:

$$\begin{aligned} \Delta \mathbf{Y} \cdot \Delta \mathbf{Y}^T &= \mathbf{J} \cdot \Delta \mathbf{P} \cdot \Delta \mathbf{P}^T \cdot \mathbf{J}^T \Rightarrow \\ E[\Delta \mathbf{Y} \cdot \Delta \mathbf{Y}^T] &= E[\mathbf{J} \cdot \Delta \mathbf{P} \cdot \Delta \mathbf{P}^T \cdot \mathbf{J}^T] \Rightarrow \\ \mathbf{R}_{\Delta \mathbf{Y}} &= \mathbf{J} \cdot \mathbf{R}_{\Delta \mathbf{P}} \cdot \mathbf{J}^T \end{aligned} \quad (3.9)$$

where $\mathbf{R}_{\Delta \mathbf{Y}}$, $\mathbf{R}_{\Delta \mathbf{P}}$ the autocorrelation matrices of $\Delta \mathbf{Y}$ and $\Delta \mathbf{P}$, respectively. Since P_i are independent, the matrix $\mathbf{R}_{\Delta \mathbf{P}}$ is diagonal, and all the diagonal elements represent variances of the parameters, which are the unknowns. Similarly, in matrix $\mathbf{R}_{\Delta \mathbf{Y}}$, the diagonal elements are the variances of the outputs.

Denoting the k^{th} unit vector \mathbf{u}_k as a vector with all zeros except a one on the k^{th} row, we can derive an expression for the k^{th} diagonal element of $\mathbf{R}_{\Delta \mathbf{Y}}$ by using Equation 3.9:

$$\sigma_{\Delta Y_k}^2 = \mathbf{u}_k^T \cdot \mathbf{R}_{\Delta \mathbf{Y}} \cdot \mathbf{u}_k = \mathbf{u}_k^T \cdot \mathbf{J} \cdot \mathbf{R}_{\Delta \mathbf{P}} \cdot \mathbf{J}^T \cdot \mathbf{u}_k \quad (3.10)$$

Let $\mathbf{v}_k^T = \mathbf{u}_k^T \cdot \mathbf{J}$, where $k = 1, \dots, m$. It is evident that \mathbf{v}_k^T is the k^{th} row of the Jacobian matrix \mathbf{J} . From the above, we can now transform the problem to a linear optimization problem of the form $\mathbf{b} = \mathbf{A} \cdot \mathbf{x}$, relating the variances of the outputs on the left-hand side to the squares of the sensitivities and the input variances on the right-hand side.

$$\underbrace{\begin{pmatrix} \sigma_{\Delta y_1}^2 \\ \sigma_{\Delta y_2}^2 \\ \vdots \\ \sigma_{\Delta y_m}^2 \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} J_{11}^2 & J_{12}^2 & \cdots & J_{1n}^2 \\ J_{21}^2 & J_{22}^2 & \cdots & J_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ J_{m1}^2 & J_{m2}^2 & \cdots & J_{mn}^2 \end{pmatrix}}_{\mathbf{A}} \cdot \underbrace{\begin{pmatrix} \sigma_{\Delta p_1}^2 \\ \sigma_{\Delta p_2}^2 \\ \vdots \\ \sigma_{\Delta p_n}^2 \end{pmatrix}}_{\mathbf{x}} \quad (3.11)$$

Although typically PCA is used in order to get an independent model parameter set, in some cases parameter correlations may be preserved by the models, for example correlations between the NMOS and PMOS parameters. As long as the covariance of the correlated parameters can be calculated using the given model, the above function can be altered to incorporate correlations. If, for example, two parameters with indices u and v are correlated, a term $2 \cdot J_{ku} \cdot J_{kv} \cdot cov(\Delta p_u, \Delta p_v)$ is added to the right-hand side of the equation for the k^{th} output variance. From here, given a set of observations of the output vector extracted by test structures, we can calculate the variances of the parameters by solving Equation 3.11.

3.1.3 Parameter screening and system constraints

The resulting $m \times n$ system of equations has, in the general case, more inputs than outputs ($m < n$). Linear systems of equations are normally solved by minimizing the residual sum of squares as shown in (3.12).

$$\underset{x}{\text{minimize}} \|\mathbf{b} - \mathbf{A} \cdot \mathbf{x}\|_2^2 \quad (3.12)$$

In the case of overdetermined systems ($n < m$), the least-squares approach is guaranteed to find a unique closed-form solution. In underdetermined systems however, there is an infinite number of solutions, if any. In this case, it is common to formulate the problem as shown in (3.13) and select the minimum norm solution.

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_2^2 \\ &\text{subject to } \mathbf{b} = \mathbf{A} \cdot \mathbf{x} \end{aligned} \quad (3.13)$$

However, in this case the minimum norm solution will not generally be an applicable one. The reason for that is that there are constraints imposed by the models and the nature of the problem. For example, a solution that contains negative values is not acceptable, since the unknowns represent variances and are therefore always non-negative. In order to guarantee a physically acceptable solution, we perform parameter screening and then add physical explicit constraints to the problem.

Table 3.1: First step of parameter selection

Given $\mathbf{A}_{m \times n}, \mathbf{x}_{n \times 1}$	
1:	for $j = 1, 2, \dots, n$
2:	if $\ \mathbf{A}_{\bullet j}\ \leq \lambda$
3:	eliminate $\mathbf{A}_{\bullet j}$
4:	eliminate x_j
5:	return \mathbf{A}, \mathbf{x}

Parameter screening in the system of Equation 3.11 enables the reduction of the length of \mathbf{x} and is done in two steps. First, we observe that matrix \mathbf{A} will, in the general case, have some level of sparsity. This observation is motivated by the fact that only a certain number of all model parameters will play an important role for the chosen output in a given system. In other words, a given output may be insensitive to certain parameters. This observation allows the designer to eliminate columns of the sensitivity matrix using a simple algorithm shown in Table 3.1, where $\mathbf{A}_{\bullet j}$ denotes the j^{th} column of the matrix \mathbf{A} . Parameter λ provides a tradeoff between complexity (i.e. number of parameters) of the final system and accuracy, and should be selected by the designer to accommodate the given design. At this first step λ is kept very small, so that only column that have zero-norms or norms that are multiple orders of magnitude lower get eliminated. This conservative selection may be enough for some cases, depending on the given design and the given models. If it is not enough, we proceed with more sophisticated parameter selections methods, discussed below.

If after the first step of parameter selection, the system is still underdetermined, we introduce some form of regression in order to produce a final solution. For this, we use a combination of ridge regression [38] and the least absolute shrinkage and selection operator (LASSO) [39, 40]. Both are forms of regularization used in statistics and machine learning in order to find a solution to ill-formed problems by adding an additional constraint to the system. Ridge regression constraints the l_2 -norm of the solution, as shown in (3.14), which helps shrink large coefficients to reduce overfitting, as illustrated in Figure 3.2 using example data.

$$\begin{aligned} & \text{minimize } \|\mathbf{b} - \mathbf{A} \cdot \mathbf{x}\|_2^2 \\ & \text{subject to } \|\mathbf{x}\|_2 < t \end{aligned} \tag{3.14}$$

The LASSO, shown in (3.15), constraints the l_1 -norm of the solution. This not only helps shrink large coefficients, like ridge regression, but also forces some coefficients to zero, therefore achieving parameter selection. In Figures 3.3 and 3.5, it is shown that as t increases, more and more coefficients get non-zero values and prediction error is reduced, and therefore t provides a trade-off between complexity and accuracy. LASSO has been previously used as

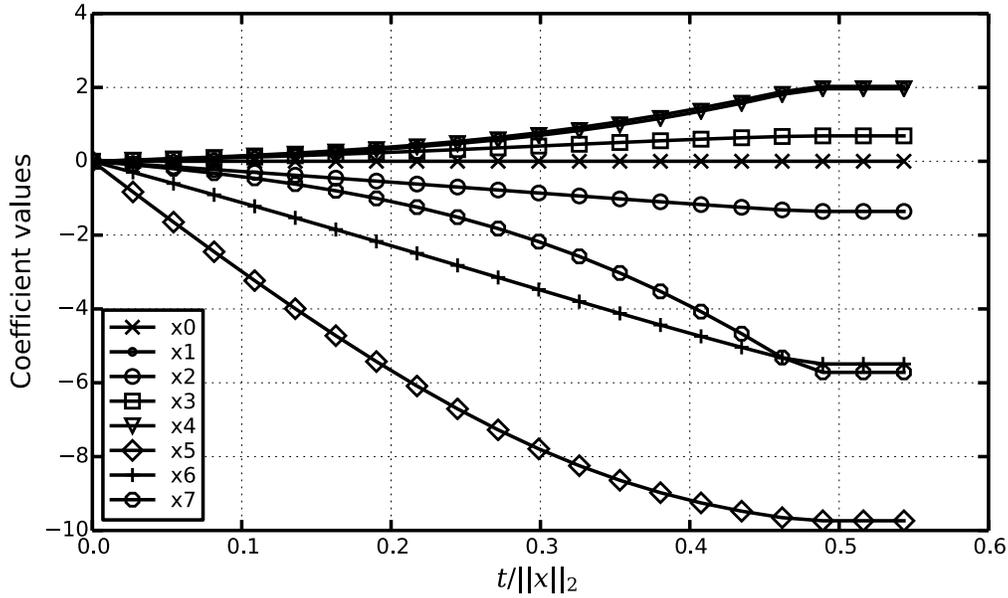


Figure 3.2: Ridge regression, produced with artificially generated data. Coefficients are kept small and converge to their final values.

an improvement of response surface modeling for large-scale performance modeling problems, demonstrating dimensionality reduction without overfitting [41].

$$\begin{aligned} & \text{minimize } \|\mathbf{b} - \mathbf{A} \cdot \mathbf{x}\|_2^2 \\ & \text{subject to } \|\mathbf{x}\|_1 < t \end{aligned} \quad (3.15)$$

Stable convergence, shrinkage and parameter selection can be achieved by combining the two regularization methods to an elastic net formulation, shown in (3.16) [42]. Figure 3.6 shows an illustration of the penalty terms for each of the aforementioned regularization techniques, for a 2-parameter system. The elastic net penalty term is a convex combination of ridge regression, for $\rho = 1$, and the LASSO, for $\rho = 0$. Figure 3.4 shows the effect on the coefficients when applying his method to the same example data as before.

$$\begin{aligned} & \text{minimize } \|\mathbf{b} - \mathbf{A} \cdot \mathbf{x}\|_2^2 \\ & \text{subject to } (1 - \rho)\|\mathbf{x}\|_1 + \rho\|\mathbf{x}\|_2 < t \end{aligned} \quad (3.16)$$

So far we have exploited several statistical concepts in order to reduce the parameter space and produce a system solution while avoiding overfitting. However, in device modeling there is only a limited set of solutions that are physically acceptable, and the solution produced

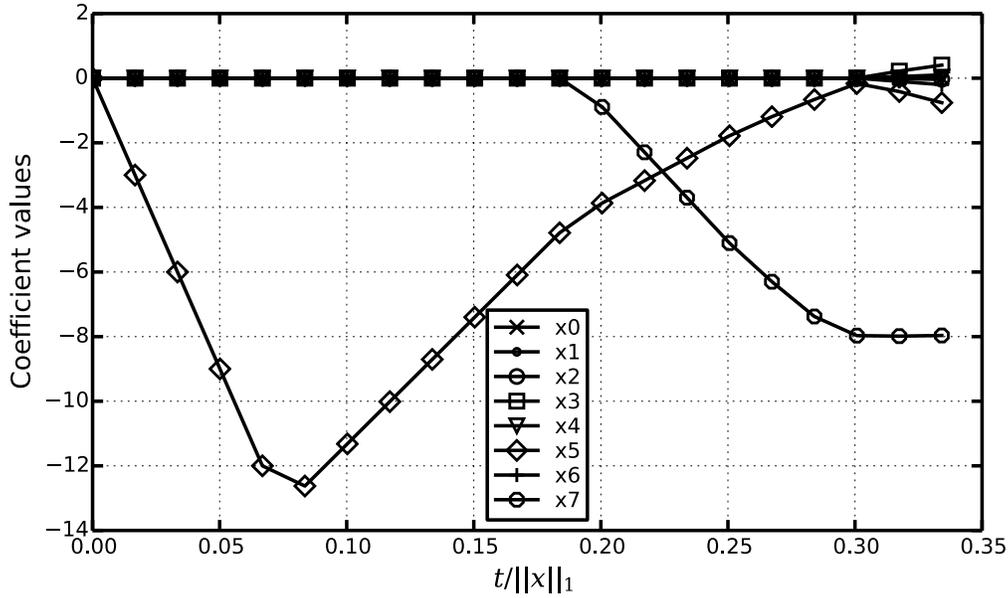


Figure 3.3: LASSO solution, produced with artificially generated data. For decreasing t , more and more coefficients are forced to zero.

from (3.16) may not be one of them. In order to limit the solution space into that physically acceptable set, we introduce a number of constraints based on knowledge of the system and models. Such constraints are derived directly by the nature of the solution, for example $\mathbf{x} \geq 0$ since \mathbf{x} consists of variances, and from the given model documentation. Therefore, we formulate the problem as a constrained convex optimization problem.

$$\begin{aligned}
 & \text{minimize } \|\mathbf{b} - \mathbf{A} \cdot \mathbf{x}\|_2^2 \\
 & \text{subject to } \begin{cases} (1 - \rho)\|\mathbf{x}\|_1 + \rho\|\mathbf{x}\|_2 < t \\ \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{cases} \quad (3.17)
 \end{aligned}$$

In Equation 3.17, \mathbf{l} and \mathbf{u} indicate the lower and upper boundaries of \mathbf{x} , respectively. Note that not all parameters need to be constrained, in which case the corresponding elements of vectors \mathbf{l} and \mathbf{u} can be set to 0 and a very large value, respectively.

The final solution, \mathbf{x}^* , contains the variances of each parameter, which are then added back into the model, therefore generating a customized model card, tied to the specific design.

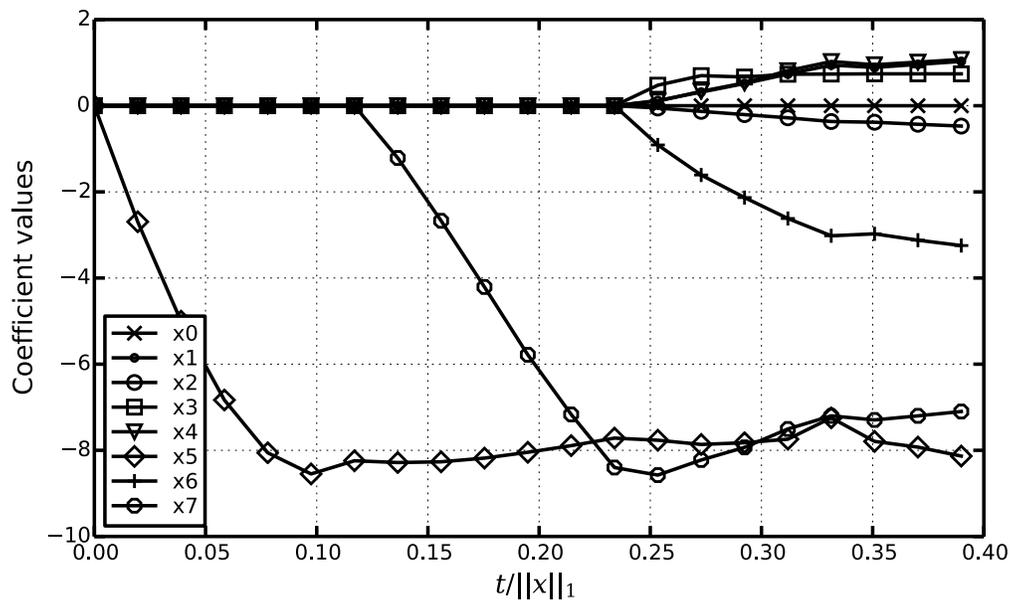


Figure 3.4: Elastic net solution with $\rho = 0.5$, produced with artificially generated data. The solution combines the properties of ridge regression and LASSO, depending upon the value of ρ .

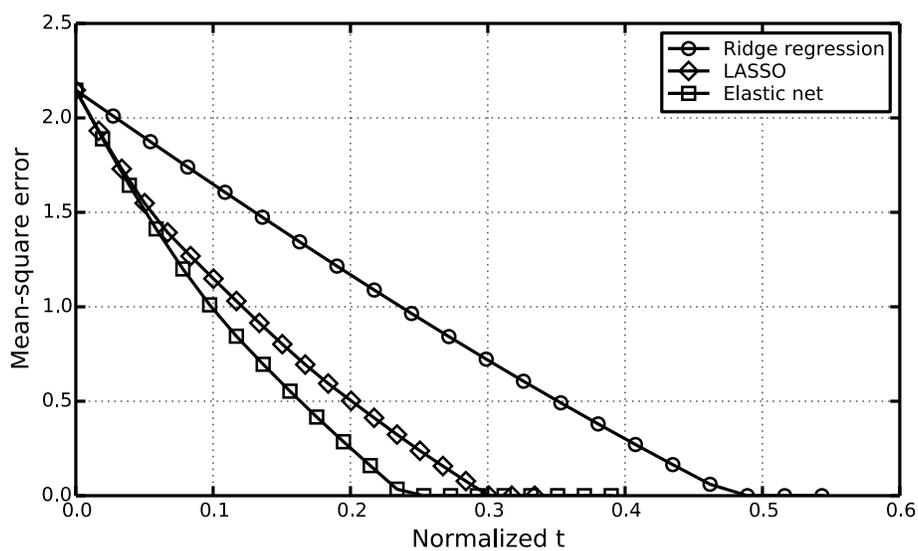


Figure 3.5: Root mean square error for ridge, LASSO and elastic net regression.

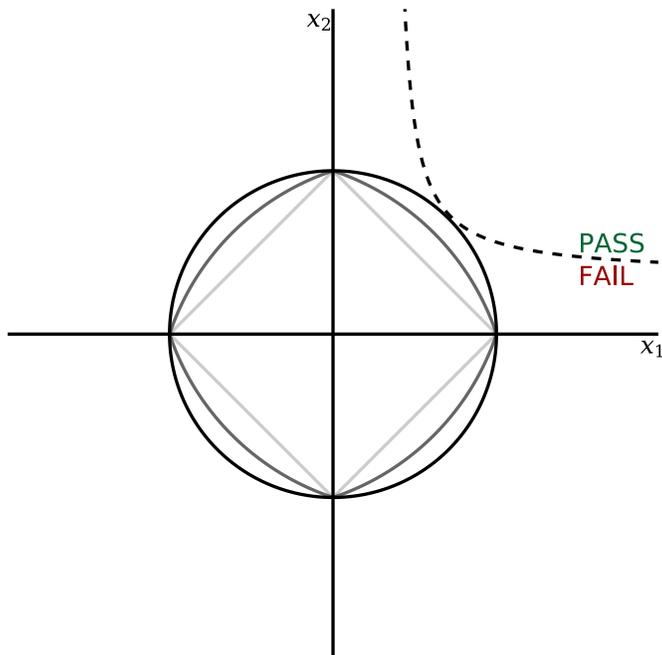


Figure 3.6: Two-dimensional contour plots of the ridge (black), LASSO (light grey) and elastic net (dark grey) penalty terms for a hypothetical 2-parameter system. The dotted line indicates the valid solution space.

3.2 Application examples

3.2.1 Customized models for device characterization

For a first-order validation of this methodology through simulation, we extract voltage-current characteristics of NMOS and PMOS devices in a 28nm FDSOI technology, given a PSP technology model. The circuit used for extraction is shown in Figure 3.7. The selection of this test structure allows us to compare the customized model cards directly to the original cards.

Methodology

In order to set up the system of Equations 3.11, we begin with identifying the complete statistical parameter set used in the given models. Figure 3.8 shows scatter plots of some of the parameters extracted from MC simulation revealing full correlation in two pairs of parameters that are physically linked, which are incorporated into Equation 3.11. Only a

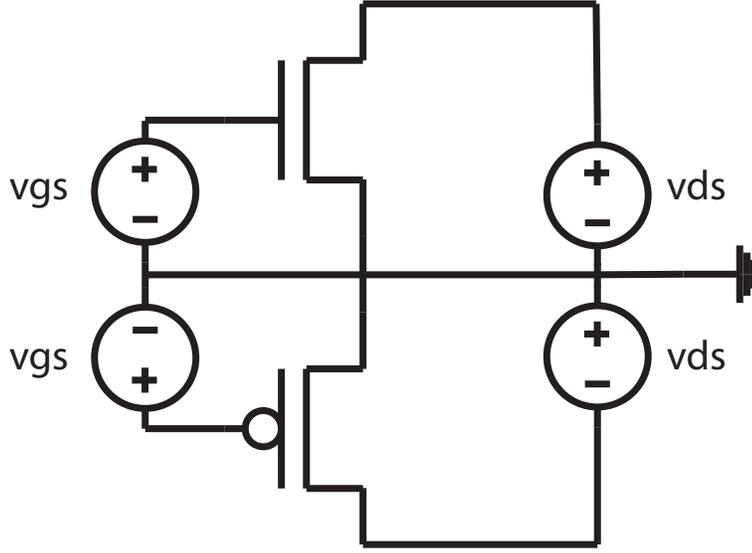


Figure 3.7: Simulation circuit for IV curve extraction of NMOS and PMOS.

subset of the parameters used is shown in order to simplify plots. The full set of statistical parameters in this particular model is shown in Table 3.2. The fourth column of the Table shows the model used for certain parameter correlations in the model.

The next step is the selection of the outputs of interest for the specific design. Here, we select a set of 5 outputs that are of interest for digital and/or analog design:

$$\mathbf{y} = [I_{on}, V_{th}, \log I_{off}, g_m, g_o]$$

Next, we use finite differences to extract the Jacobian matrix from the given simulation model. The top part of Table 3.3 shows the extracted percent sensitivities of each output with respect to the parameter subset and therefore corresponds to matrix \mathbf{J} , and the bottom part shows the calculated norms of each column of matrix \mathbf{A} . This format shows the tradeoff between choosing a smaller or larger λ for parameter selection. Zero columns are automatically removed from the system, and the choice of λ in the order of $1e^{-8}$ further reduces the size of the final system to 8 parameters without significant error.

After parameter selection, physical constraints are added in order to limit the solution space. The first and most obvious constraint that needs to be added to the system is $\mathbf{x} \geq 0$, since \mathbf{x} is a vector of variances. From there, minimum and maximum values for some parameters are derived from the model documentation, whereas other parameters remain unconstrained. The resulting optimization problem is then solved using a commercially available convex programming package, like CVX [43, 44].

Figure 3.9 shows a comparison of the histograms extracted by MC simulation before and after adding these constraints. It is evident that in the second case the system solu-

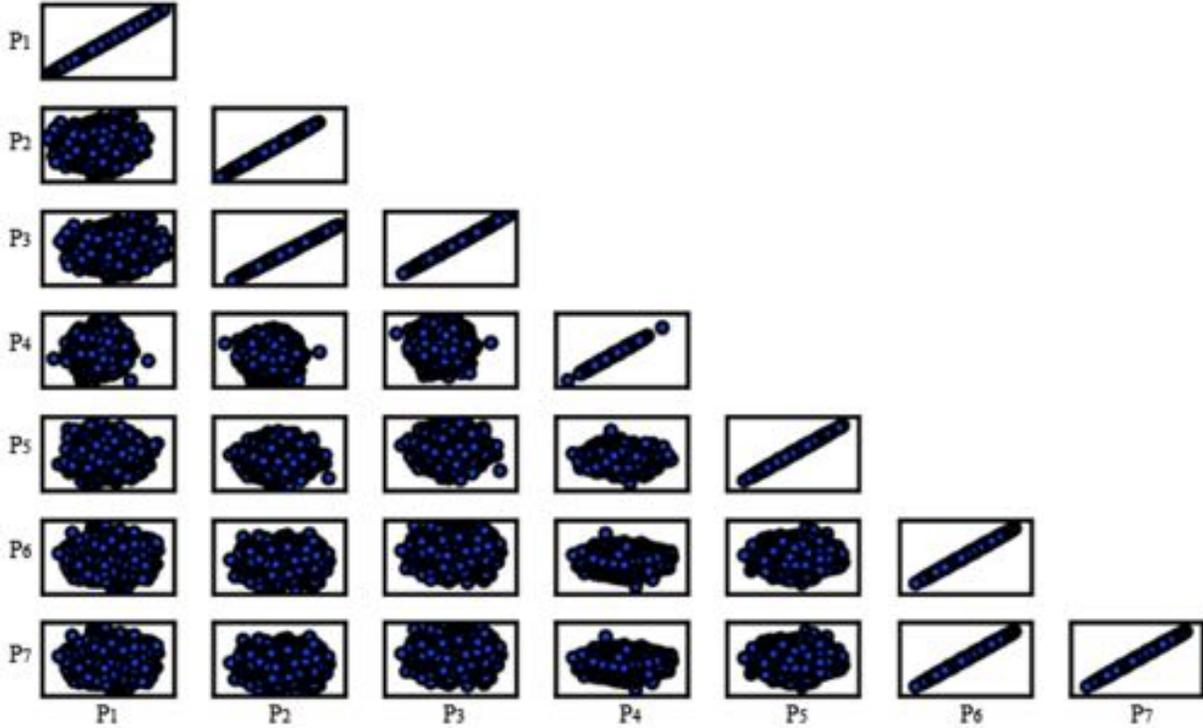


Figure 3.8: Scatter plots of a subset of the parameters used.

tion matches more accurately the body of the original distribution. Figure 3.10 shows the corresponding quantile-quantile plots. We observe a better match in the tails of the distribution as well. Finally, Figure 3.11 compares the resulting distributions for all the selected outputs, after customizing the model cards with a reduced statistical parameter set, each one of which is assigned a normal distribution with standard deviation calculated by solving Equation 3.11. The calculated standard deviations for a subset of the parameters of the customized model (CM) are shown in Table 3.4, compared to the original model (OM). Although these results are based only in simulation, the fact that they are very close shows that the simplified linear model with reduced parameter set that was used is capable of predicting device performance. Customized models using measured data will be presented in Chapter 5.

Importance sampling

The methodology so far relies on Monte-Carlo simulation using customized model cards in order to predict design yield. One major limitation of Monte-Carlo simulation is the large number of simulation runs needed in order to find a rare event. Although a DC analysis on a simple circuit like the one of Figure 3.7 can simulate relatively fast, anything more complicated can result in excessive runtimes. Importance sampling has been proposed as a

method to reduce runtime for SRAM failure prediction [45, 46], and is leveraged in order to achieve simulation speedups in this work.

Importance sampling is based on the introduction of a proposal distribution for the model parameters, such that the rare event f of interest is converted to an event with higher probability (close to 0.5), therefore enabling the reduction of the number of simulation runs. From there, the original probability of the rare event can be calculated through a series of algebraic manipulations. The success of the importance sampling approach depends crucially on how well the proposal distribution matches the desired distribution. In this case, the proposal distribution is gaussian, just like the original one, but the mean is shifted such that $p(f) \approx 0.5$. This choice provides good matching with the original distribution as well as simplified calculations.

A key problem with this approach is the selection of the shift vector to be applied to the model parameters. The selection of the shift vector is done using some sort of search algorithm, however it is necessary to ensure that the algorithm converges fast in order to actually achieve simulation speedup. For this reason, we propose a simple and efficient spherical search algorithm, which exhibits both a variable radius (similar to the one in [46]) in a divide-and-conquer manner as well as a shift of the center of the search. The algorithm

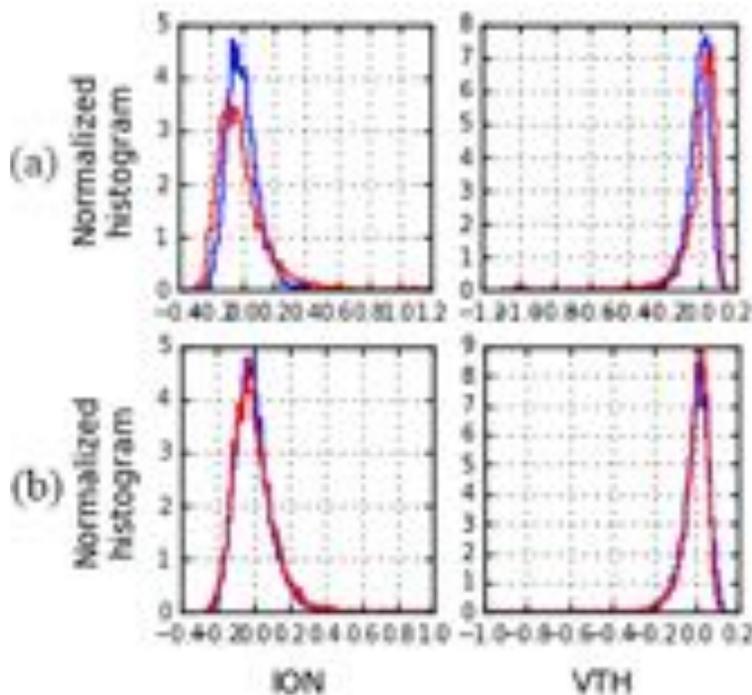


Figure 3.9: Comparison of extracted histograms (a) without adding physical constraints and (b) with physical constraints. In both cases blue denotes the original model and red the customized model. All data are normalized.

is shown in Table 3.5. For Step 1 of the algorithm a maximum number of iterations is set. If Step 1 does not conclude in that time, the initial radius R of the search needs become larger. Step 2 may be repeated with a shifted search center and reduced radius for increased accuracy.

Figure 3.12 shows a comparison of the estimated probability of failure between Monte-Carlo simulation with and without using importance sampling. The target probability of failure is 0.001. It is shown that importance sampling can produce results with fewer samples ($N \leq 1000$), while regular Monte-Carlo fails. For $N > 1000$ the relative absolute error of the importance sampling method is consistently less than the regular Monte-Carlo. For large number of simulation runs the results of the two approaches converge. In order to produce results with 90% confidence and 90% accuracy, regular Monte-Carlo requires 100.000 simulation runs, while importance sampling requires only 3000 runs.

This algorithm uses $N_{Step1} + N_{Step2}$ simulations to find the right shift vector and N simulations to estimate the probability of interest. Heuristically, it is found that N_{Step1} and N_{Step2} are in the order of ~ 50 simulation runs, therefore the total number of simulations still remains less than the number required for Monte-Carlo. N_{Step1} can be further reduced

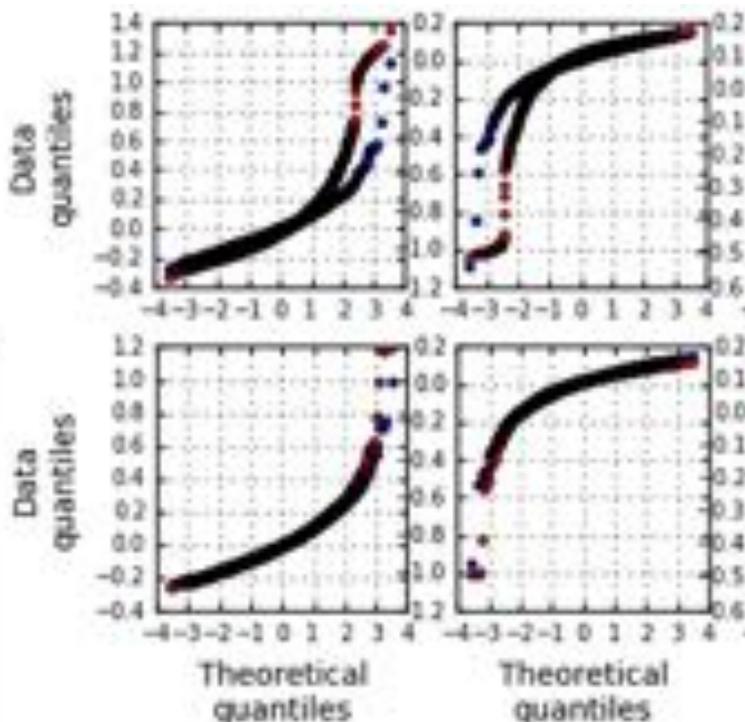


Figure 3.10: Comparison of extracted histograms (a) without adding physical constraints and (b) with physical constraints. In both cases blue denotes the original model and red the customized model. All data are normalized.

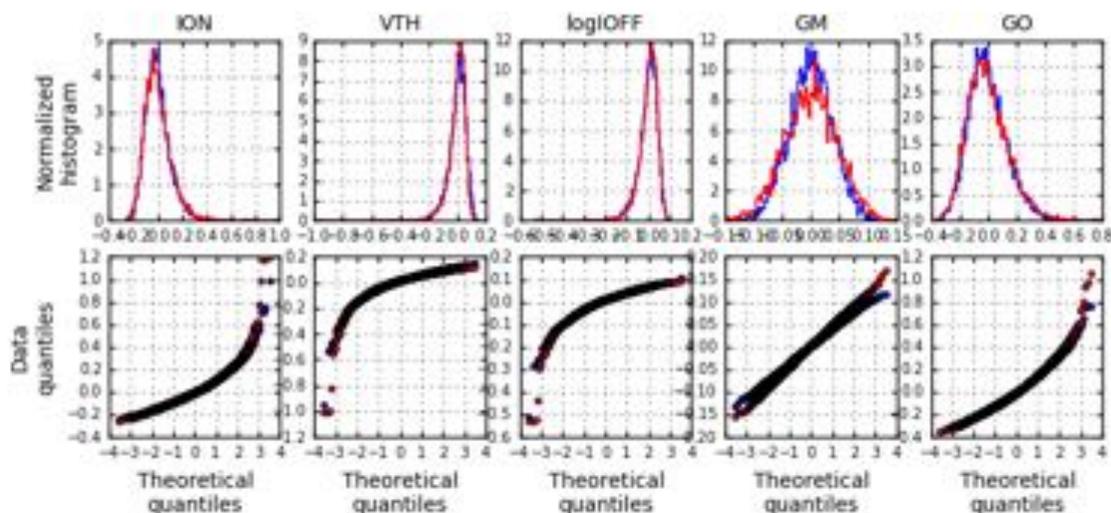


Figure 3.11: Output histograms (top) and quantile-quantile plot (bottom) comparison. In both cases blue denotes the original model and red the customized model. All data are normalized.

by performing numerical optimization on the target function $p(f) - 0.5$. Although more robust, this comes at an increased complexity which limits its usability.

3.2.2 Customized models for comparator characterization

Next we characterize by simulating one of the most useful and most used mixed-signal circuits, the StrongARM comparator [47, 48]. The circuit schematic of a StrongARM comparator is shown in Figure 3.13. This topology finds wide usage as a sense amplifier, a comparator, or simply a robust latch with high sensitivity. Its zero static power consumption, rail-to-rail swing and manageable input-referred offset make this latch very popular in all types of circuit design, ranging from analog front-ends and analog-to-digital converters to memory.

Impulse Sensitivity Function

In order to characterize the comparator we will treat it as a linear time-invariant (LTI) system, within one clock period, and extract its Impulse Sensitivity Function (ISF). ISF system theory has been used in the past to describe sampled and periodic systems [49–51].

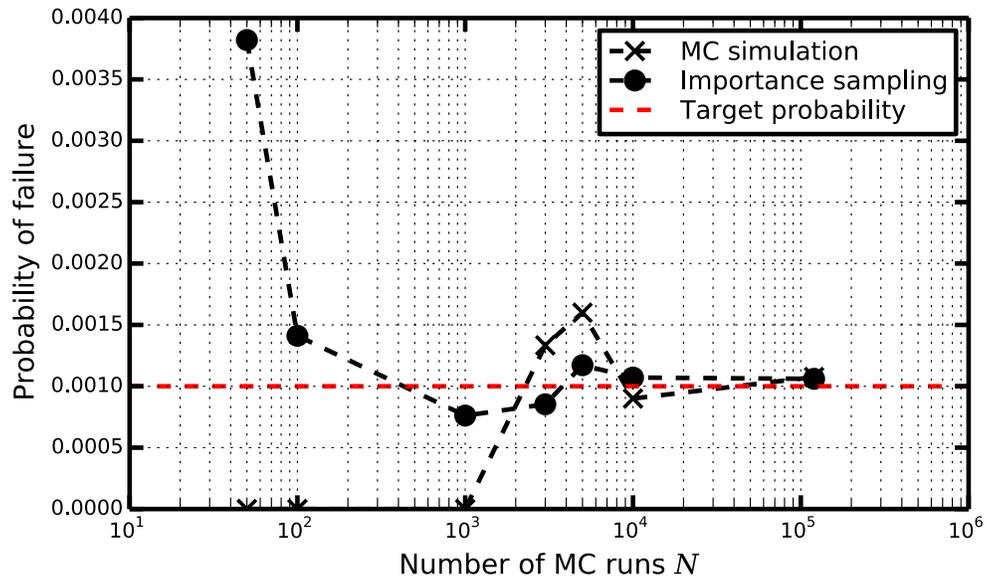


Figure 3.12: Simulated probability of failure using Monte-Carlo and importance sampling.

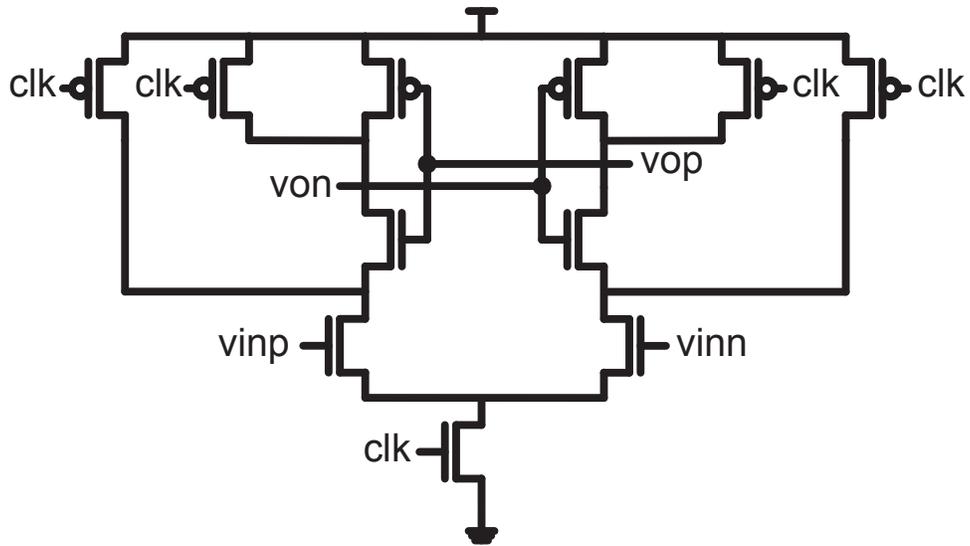


Figure 3.13: Circuit of a StrongARM latch including pre-charge devices.

Although these systems are generally non-linear, if we assume that the input exhibits small perturbations around a certain DC bias, we can approximate the system as linear around that bias, within a clock period.

In order to calculate the impulse response $h(t)$ of a continuous-time, LTI system, lets assume a step input $x(t) = u(t)$. Then we have:

$$\begin{aligned}
y(t) &= \int_{-\infty}^t h(t - \tau)u(\tau)d\tau \\
&= \int_{-\infty}^0 h(t - \tau)u(\tau)d\tau + \int_0^t h(t - \tau)u(\tau)d\tau \\
&= \int_{-\infty}^0 h(t - \tau) \cdot 0 \cdot d\tau + \int_0^t h(t - \tau)d\tau \\
&= C + \int_0^t h(t - \tau)d\tau
\end{aligned} \tag{3.18}$$

In Equation 3.18 we substitute $u = t - \tau$ and get:

$$y(t) = C + \int_0^t h(u)du \tag{3.19}$$

From here if we take derivatives in both parts of (3.19):

$$\begin{aligned}
\frac{d}{dt}y(t) &= \frac{d}{dt} \int_0^t h(u)du \\
&= \int_0^t \frac{d}{dt}h(u)du
\end{aligned} \tag{3.20}$$

Therefore, we have an expression of the impulse response of the system:

$$h(t) = y'(t) \tag{3.21}$$

To further understand the concept we will consider the example of a simple NMOS sampler. We can treat the NMOS sampler as an LTI system, described by Equation 3.18, followed by an ideal sampler which samples the value of the output of the system at time t_s . We will denote the input of the LTI system as $v_i(t)$, the output as $v_o(t)$, and the output of the ideal sampler as $v_s(t_s) = v_o(t_s)$, where t_s is the sampling instant. Therefore:

$$v_s(t_s) = v_o(t_s) = \int_{-\infty}^{t_s} h(t_s - \tau)v_i(\tau)d\tau \tag{3.22}$$

and assuming the input is a small step of amplitude A , $v_i(t) = Au(t)$, from Equation 3.21 we have:

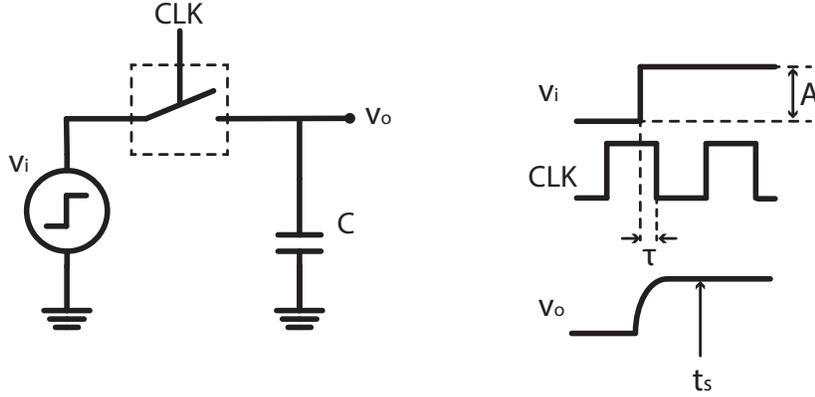


Figure 3.14: Simulation setup for ISF extraction of a sampler.

$$h(t_s) = \frac{1}{A} v'_s(t_s) \quad (3.23)$$

From the above derivation, which is based on the one presented in [49], we see that in order to characterize the impulse response of a sampled system in measurement, we can take the derivative of the step response and normalize it by the amplitude of the input step. This mitigates the need for an impulse at the input. More importantly, this technique characterizes two different effects:

1. The sampled system bandwidth, as that is determined by the finite on-resistance, the load capacitance and the parasitic capacitances.
2. The sampled system aperture width, as that is determined by the clock finite rise and fall times.
3. The sampled system gain, as that is determined by the integral of the ISF.

Once the impulse response $h(t)$ has been determined, its Fourier transform can reveal information about the switch bandwidth, since applying the convolution theorem in (3.22) we have:

$$V_s(j\omega) = H(j\omega) \cdot V_i(j\omega) \quad (3.24)$$

In simulation, the ISF of a sampler can be extracted with the simple setup shown in Figure 3.14, i.e. by moving the clock edge progressively closer to the input step edge and monitoring the sampled output at time t_s . Figures 3.15 and 3.16 show the simulated ISF and spectrum for various clock rise times, respectively, for a semi-ideal switch. For shorter clock

rise/fall times the switch ISF is closer to an ideal impulse response and its bandwidth closely matches that of an ideal RC circuit. For larger clock rise/fall times the ISF methodology captures the bandwidth degradation due to increased aperture width.

Impulse Sensitivity Function for comparators

A comparator is essentially a combination of an amplifier and a sampler - it samples a small voltage swing in the input and amplifies it with a very high gain. As such, it can also be characterized using an ISF extraction setup similar to the one described previously. Note that any comparator during its "sampling" operation can be broken down to an amplification stage, typically analog in nature, and a regeneration stage, typically digital. The amplification stage samples the voltage difference at the input, and provides an amplified voltage difference to the regeneration stage. When that amplified voltage difference exceeds a threshold a positive feedback loop is enabled, which provides further amplification at a very high speed through positive feedback.

Due to the more complex nature of the comparator the ISF characterization setup has to be modified, but the main operating principle remains the same. A step is applied at the input and the clock edge is swept relative to the input edge. The sampled voltage of the amplification stage needs to be read at the output at each step. However, access to the internal nodes of the comparator is not available, at least not without significantly altering the circuit and its characteristics. To solve this problem, an offset voltage is applied at the input and swept until the comparator reaches the metastable point [52]. The metastable point can be determined either by an ideal negative feedback loop that forces the amplification stage outputs to be equal, without loading the comparator, or by statistical analysis of the digital output.

Figure 3.17 shows the simulation setup for measuring the offset of a comparator by means of statistical analysis. The offset of the comparator is the voltage that needs to be applied to the input so that the comparator reaches its metastable point. A DC voltage is applied to one input and slow ramp around that DC voltage is applied to the other input. Then the comparator is clocked N times. If the comparator is at its metastable point when the two inputs are equal, then the number of ones read equals the number of zeros. If not, the ratio of the number of zeros to the N reveals the offset of the comparator.

The same concept is used to extract the ISF of the comparator. One of the inputs is held at the common mode while the other is fed with a square pulse of small swing, around a DC level V_{OS} . V_{OS} is swept slowly, just like the slow ramp in the offset characterization setup. The point at which the number of ones equals the number of zeros at the output is the metastable point of the comparator. The key distinction here is to find the metastable point as a function of τ , where τ is the time difference between the clock edge and the input edge. In simulation, that can be done with just a sweep statement. However, simulation speedup can be achieved by eliminating one of the sweeps, but implementing the setup shown in Figure 3.18. The circuitry remains the same as before, but the timing of the inputs changes; the input is at a reference frequency f_{ref} , but the clock of the comparator is at a frequency

slightly offset from the reference. Assuming that $T_{clk} = T_{ref} + \Delta\tau$, the first sample of the comparator will be $\Delta\tau$ away from the input edge, the second will be $2\Delta\tau$ away, and so on. The total number of cycles needed in order to span one whole input period is:

$$N = \frac{T_{ref}}{\Delta\tau} \quad (3.25)$$

In a simulation setup due to absence of noise, selecting a high ($\sim 1GHz$) reference frequency can speed up simulations. The speed and design limitations in a practical design will be discussed in the Chapter 4.

Application of centering methodology

As a comparator application example, we apply the centering methodology described in Section 3.1 to the setup shown in Figure 3.17. The model used is a PSP 28nm FDSOI model provided by ST Microelectronics. In this base model, variation is being added to two parameters, VFBO and UO, for each transistor instance. Since the output of interest is the comparator offset, we assume variation only in the matched pairs of the StrongARM comparator (shown in Figure 3.13), which results in a total of 12 device parameters, ignoring the digital reset devices.

Table 3.6 shows the simulated sensitivities of the StrongARM comparator offset to the model parameters of its devices. As expected, variation in the flat-band voltage of the input devices is proportional to the offset with a 1 : 1 ratio, approximately. Variation in the NMOS devices of the cross-coupled pair affects the offset less due to the gain of the amplification stage of the comparator. Finally, variation in the PMOS devices affects the offset the least, since by the time the PMOS devices come into play the positive feedback has already been enabled, which means an even higher gain.

Using the simulated sensitivities we set up Equation 3.11 and use elastic net regression with $\rho = 0.5$. The solution converges for approximately $t > 10$, producing a root-mean square error of less than 10^{-19} (Figure 3.19). The elastic net solution allows us to customize the models across different biases. For example, for various values of the supply voltage the comparator offset voltage extracted from the customized models closely tracks the offset predicted by regular Monte-Carlo on the original model cards (Figure 3.20). Finally, Figure 3.21 shows the comparison between the produced distributions and QQ plots, showing that both the body and tails of the distribution match closely the Monte-Carlo data. Therefore, the plots demonstrate the ability of the methodology to predict correct variances for the model, and make it a promising solution for reducing the prediction error of the models.

3.3 Summary

In this chapter some basic modeling concepts were presented, as a preface to the customized modeling methodology introduced by this work. The core of the customized modeling methodology is the combination of backward propagation of variance technique, parameter screening, and statistical regression via the elastic the net, which produces a set of parameter variances that can significantly improve the original models. The methodology was tested using simulation data of device IV curves and comparator offsets as a preliminary validation before the experimental validation.

Table 3.2: FDSOI 28nm PSP model statistical parameters for Monte-Carlo (MC) and Fixed-Corner (FC) simulation

Parameter	MC	FC	Correlation	Description
CFL	✓	✓	-	DIBL parameter (length dependence)
CFLEXP	✗	✓	-	DIBL parameter (exponent for length dependence)
CJORBOT	✓	✓	-	Zero-bias capacitance per unit area of bottom for SB junction
CJORGAT	✓	✓	-	Zero-bias capacitance per unit length of gate-edge for SB junction
CJORSTI	✓	✓	-	Zero-bias capacitance per unit length of STI-edge for SB junction
CTL	✗	✓	-	Interface states factor (length dependence)
CTLEXP	✗	✓	-	Interface states factor (exponent for length dependence)
IGINVLW	✗	✓	-	Gate channel current pre-factor for a channel area of $W_{EN} \cdot L_{EN}$
IGOVW	✗	✓	-	Gate overlap current pre-factor for a channel width of W_{EN}
LAP	✓	✓	A+B(LOV+C)	Effective channel length reduction per side due to lateral diffusion of S/D dopant ions
LOV	✓	✓	-	Overlap length for overlap capacitance
LVARO	✓	✓	-	Geometry independent difference between actual and programmed poly-silicon gate length
RSW1	✓	✓	-	Source/drain series resistance for a channel width W_{EN}
TOXO	✓	✓	-	Gate oxide thickness
TOXOVO	✓	✓	ATOXO/B	Overlap oxide thickness (geometry independent part)
UO	✓	✓	-	Zero-field mobility
VFBL	✗	✓	-	Flat-band voltage (length dependence)
VFBLW	✗	✓	-	Flat-band voltage (area dependence)
VFBO	✗	✓	-	Flat-band voltage (geometry independent)
WVARO	✓	✓	-	Geometry independent difference between actual and programmed field-oxide opening

Table 3.3: Percent sensitivities to parameters

J	p_1	p_2	p_3	p_4	p_5	p_6	p_7
y_1	0.26	0.0	0.0	0.22	1.10	0.0	0.86
y_2	0.18	0.0	0.0	0.13	0.30	0.0	0.03
y_3	0.11	0.0	0.004	0.10	0.12	0.04	0.04
y_4	0.04	0.0	0.0	0.08	0.76	0.0	0.82
y_5	0.88	0.0	0.0	0.33	0.90	0.0	0.88
$\ \mathbf{A}_{\bullet j}\ $	$9.4e^{-3}$	0.0	$4.0e^{-5}$	$4.4e^{-3}$	$1.6e^{-2}$	$0.4e^{-3}$	$1.5e^{-2}$

Table 3.4: Parameter standard deviations in original model (OM) and customized model (CM)

	p_1	p_3	p_4	p_5	p_6	p_7	p_8
<i>OM</i>	$8.54e^{-8}$	$7.50e^{-10}$	$2.42e^{-9}$	0.011	$6.43e^{-11}$	$1.38e^{-6}$	$4.85e^{-11}$
<i>CM</i>	$8.74e^{-8}$	$7.63e^{-10}$	$2.39e^{-9}$	0.012	$6.43e^{-11}$	$1.39e^{-6}$	$4.94e^{-11}$

Table 3.5: Proposed search algorithm

Step 1. Initial search
1: Set center of search to $\mathbf{s} = \mathbf{s}_{\text{nom}}$
2: Set initial search radius to $R = [-5\sigma, 5\sigma]$
3: Uniformly sample shift vector $\Delta\mathbf{s}$ from R
4: Calculate $p(f)$
5: Repeat 2-3, stop when $p(f) \approx 0.5$ and save vector $\Delta\mathbf{s}^*$
Step 2. Variable radius search
1: Move center of search to $\mathbf{s} = \mathbf{s}_{\text{nom}} + \Delta\mathbf{s}^*$
2: Set search radius to $R = R/2$
3: Perform 3-4 of Step 1 M times and save $\Delta\mathbf{s} _{p(f)\approx 0.5}$ in a collection L
4: If no $\Delta\mathbf{s} _{p(f)\approx 0.5}$ were found, set $R = R+R/2$ and repeat, else select $\Delta\mathbf{s}^* = \Delta\mathbf{s} _{p(f)\approx 0.5}$ of minimum norm

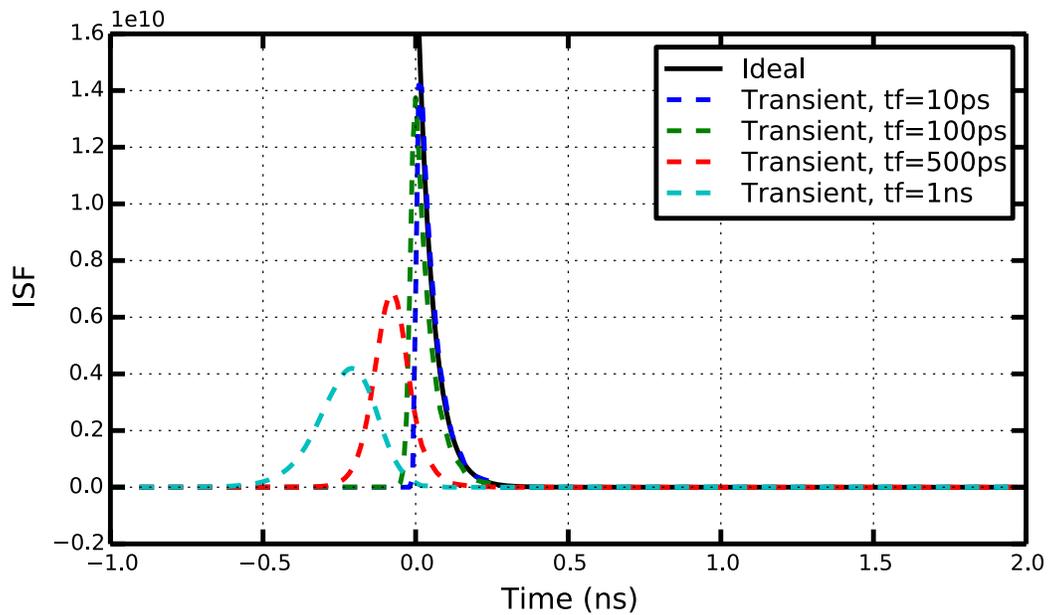


Figure 3.15: Impulse response of the sampler for various clock fall times, compared to an ideal response. Larger clock fall times increase the aperture width.

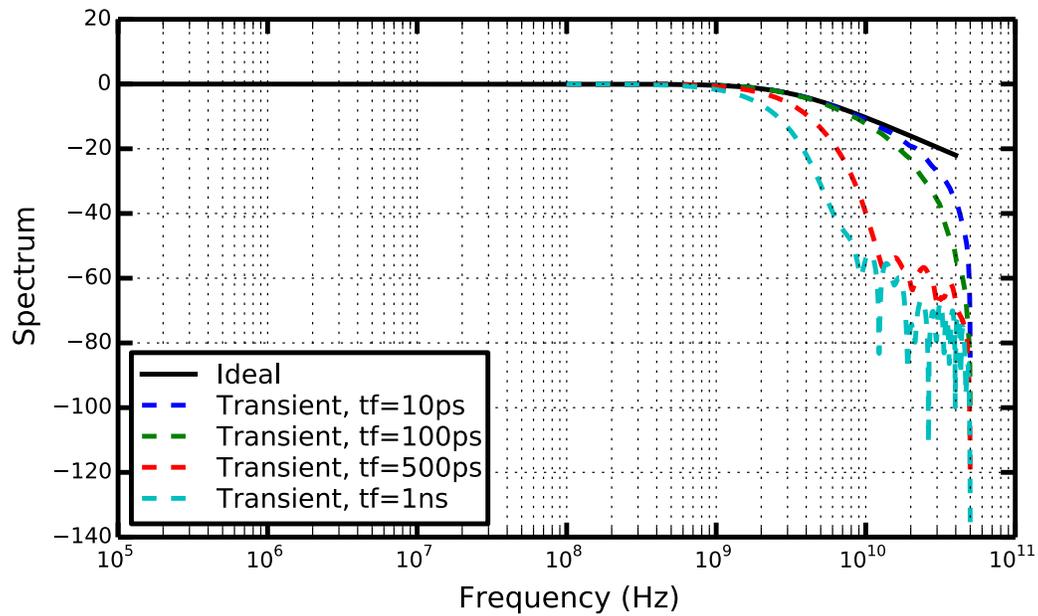


Figure 3.16: Extracted spectrum for various clock fall times, compared to an ideal response. Increased aperture width limits the bandwidth of the sampler.

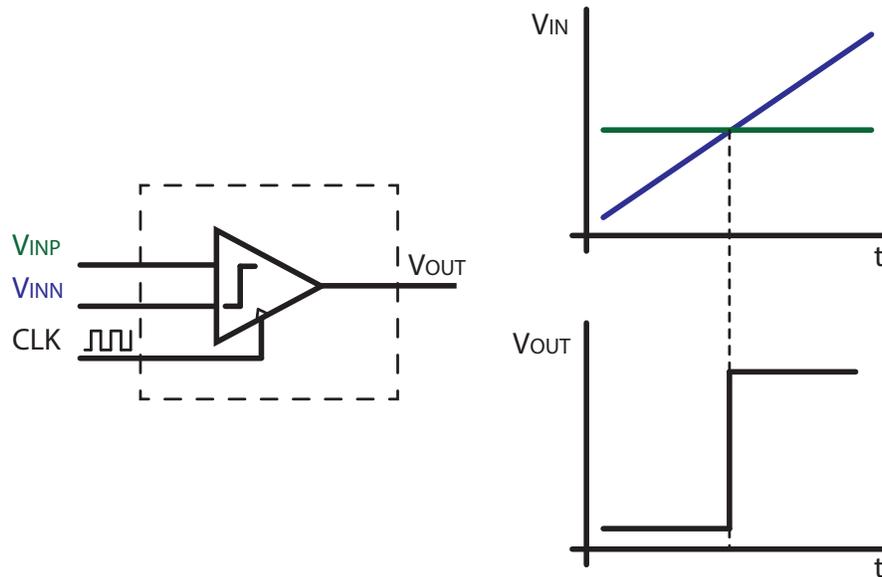


Figure 3.17: Simulation setup for offset measurement of a comparator.

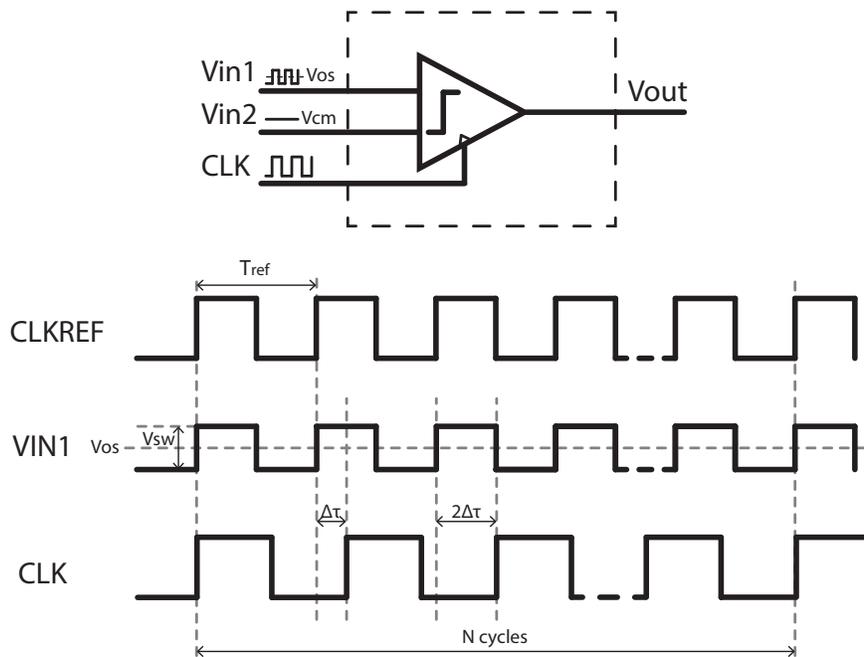


Figure 3.18: Simulation setup for impulse sensitivity function measurement of a comparator.

Table 3.6: Comparator offset sensitivity to statistical parameters for input devices and cross-coupled pair devices at nominal supply voltage.

Instance:	N_{in1}	N_{in2}	N_{cc1}	N_{cc2}	P_{cc1}	P_{cc2}
VFBO	-1.029	1.029	-0.498	0.495	0.107	-0.099
UO	1.882	-1.899	0.793	-0.802	-0.864	0.897

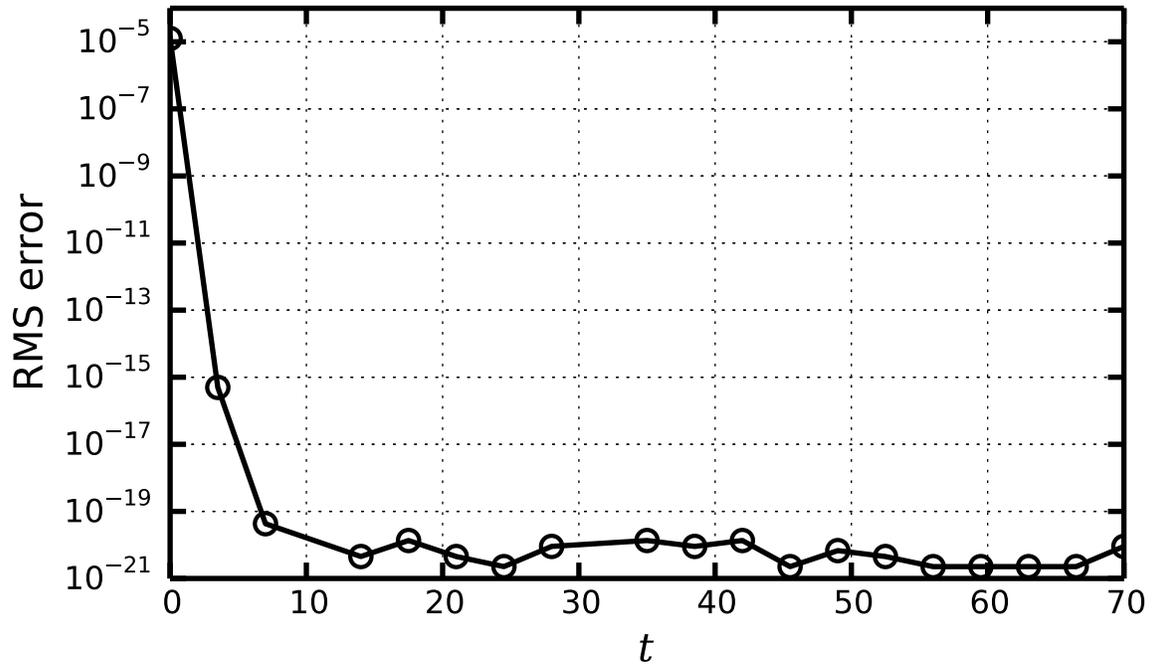


Figure 3.19: Root-mean square error of elastic net regression for comparator offset.

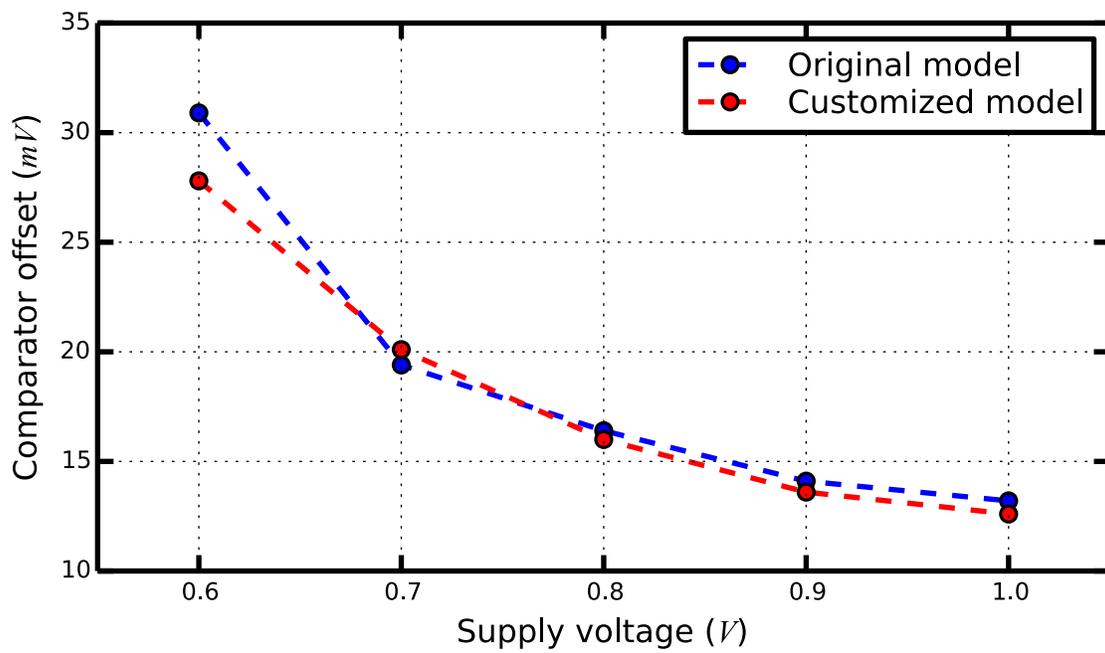


Figure 3.20: Comparison of offset prediction between original model and customized model.

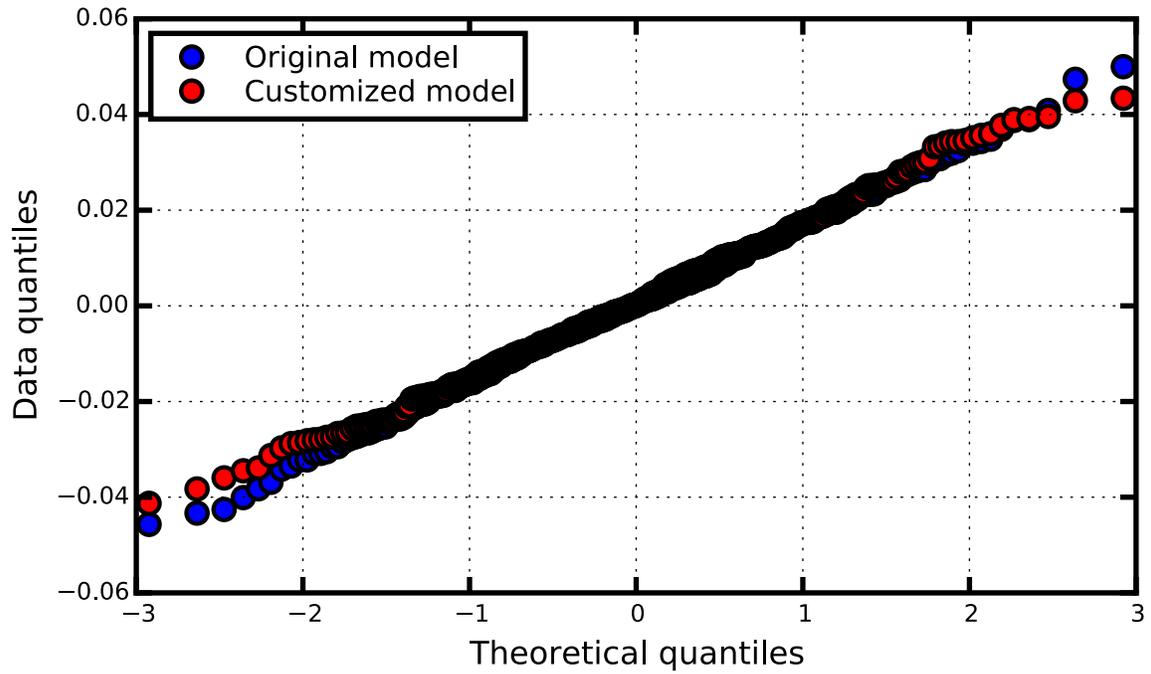
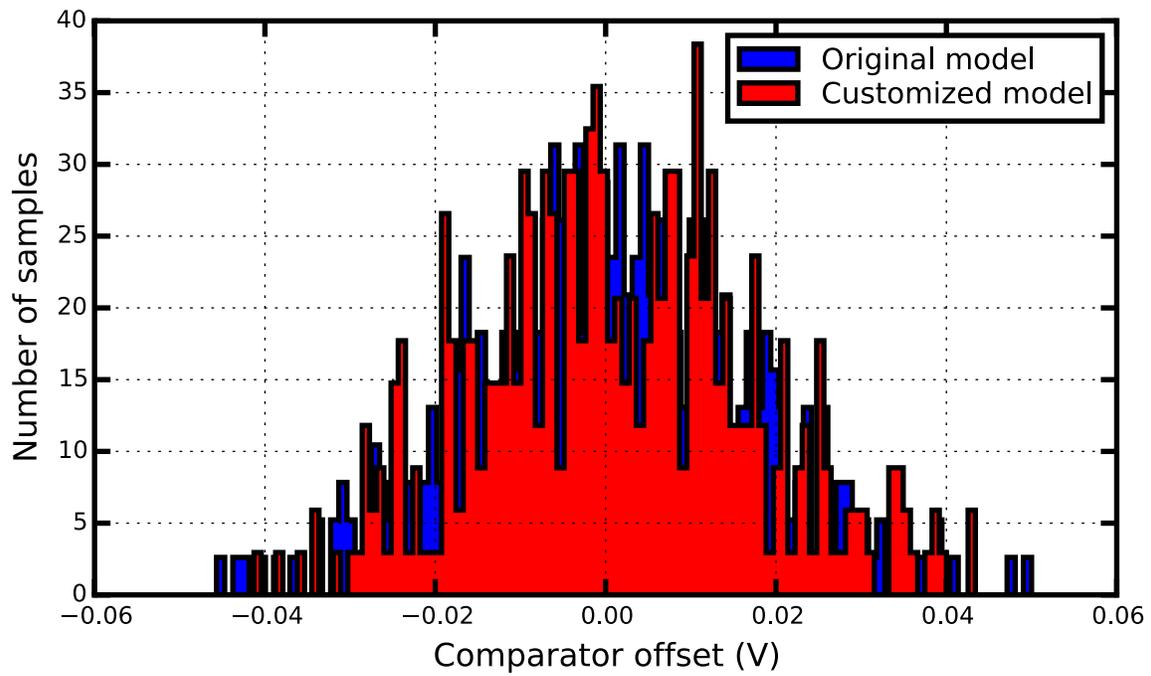


Figure 3.21: Comparison of the distributions and QQ plots between original model and customized model.

Chapter 4

Test structure design for data extraction

In the previous chapter, a customized modeling methodology was presented, which has potential of producing high-yielding models tied to a specific design. In this chapter we begin the discussion on experimental validation, by proposing a set of test structures for variability characterization and model tuning.

Section 4.1 presents a high-level overview of all test structures, while the following sections dive into detail on the test structure design, topology and layout selection, expected results, testability features and measurement process. The chapter concludes with a brief summary.

4.1 Overview of test structures and goals

As discussed in Chapter 1, the goal of this work consists of two discrete parts. The first goal is to demonstrate design-specific yield optimization for high-speed comparators, using the methodology presented in Chapter 3. The selected outputs to be optimized are the comparator offset and the comparator bandwidth. For this purpose, we design arrays of test structures targeting at measuring both the offset and the impulse sensitivity function of a large number of comparators.

The second goal is to investigate variability in high-speed comparators and characterize design-dependent, layout-dependent and topology-dependent sources of variation. For this reason we incorporate various comparator topologies, as well as several variations in the layout in the test structures, targeting at characterizing specific variation effects. Along with the comparators, we include sets of individual devices that can reveal information about technology variability, accuracy of the original variability models as well as correlations between device performance and comparator performance.

Figure 4.1 shows a high-level block diagram of the test-chip, designed in a 28nm FDSOI

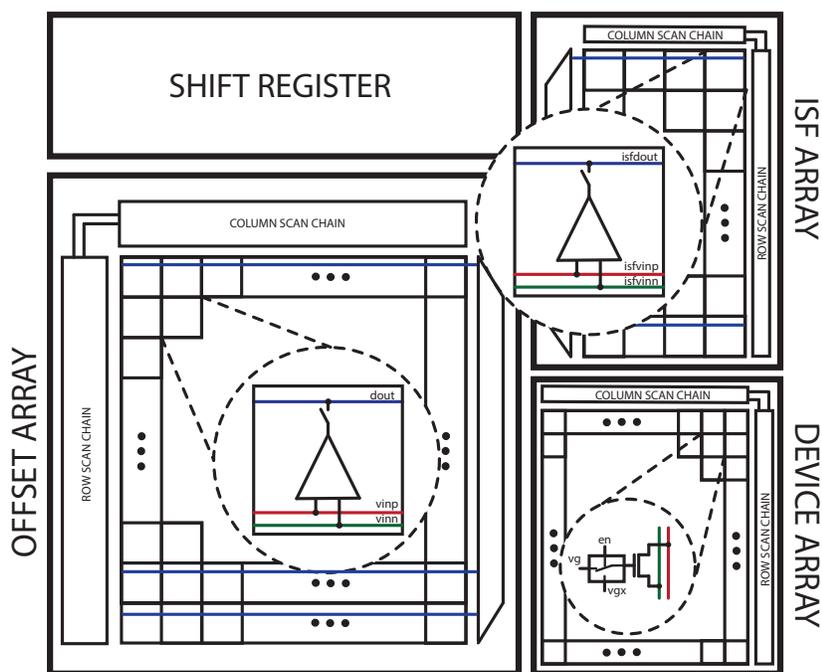


Figure 4.1: High-level chip block diagram, featuring the offset, ISF and comparator arrays and the on-chip memory.

technology. The chip consists of three discrete test arrays; a comparator offset characterization array, a comparator impulse sensitivity function characterization array and a device current-voltage characteristic extraction array. An on-chip memory in the form of a shift-register is also included. A more detailed high-level overview of the chip circuitry is shown in Figure 4.2. All structures will be discussed in detail in the following sections.

4.2 Test structure design

In order to be able to account for variation, variability test structure design is necessary. Variability test structures are structures that allow the measurement and statistical analysis of specific characteristics of a technology or design. Device arrays can provide device variability data within a die and reveal their spatial characteristics, with low or high spatial resolution. When matched pairs of devices placed in close proximity to each other are incorporated in such arrays, systematic effects can be removed by subtraction and random mismatch can be characterized. Averaging data over multiple dies exposes die-to-die

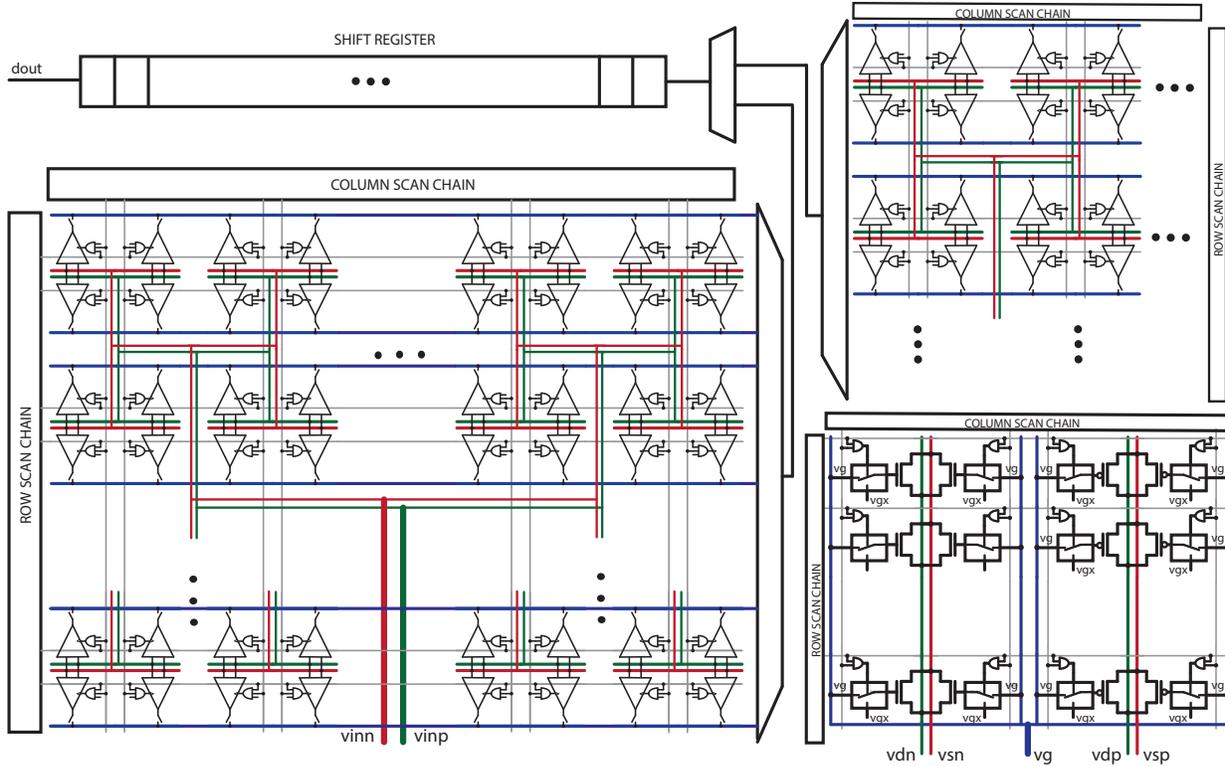


Figure 4.2: Simplified schematic of the complete chip.

(D2D) variation, averaging data over a wafer exposes wafer-to-wafer (W2W) variation and knowledge of die location enables spatial analysis of D2D effects.

Over the years, various types of test structures have been proposed and used. Arrays of padded-out active or passive devices can provide very accurate direct current measurements, at the cost of very high area and long measurement times. To mitigate that, devices can be organized in device matrix arrays (DMA) [53]. DMAs contain devices that are all connected to common measurement buses, but are individually selectable, therefore allowing a much more compact design and enabling measurement automation. A limitation of this structure is lack of accuracy due to the increased leakage floor on the measurement buses and non-idealities introduced by the selection logic.

In this work we adopt arrays similar to those in [53]. We expand them to include test circuits besides only devices, and add leakage calibration and accuracy improvement features when necessary. The test structure arrays are discussed in detail in the following subsections.

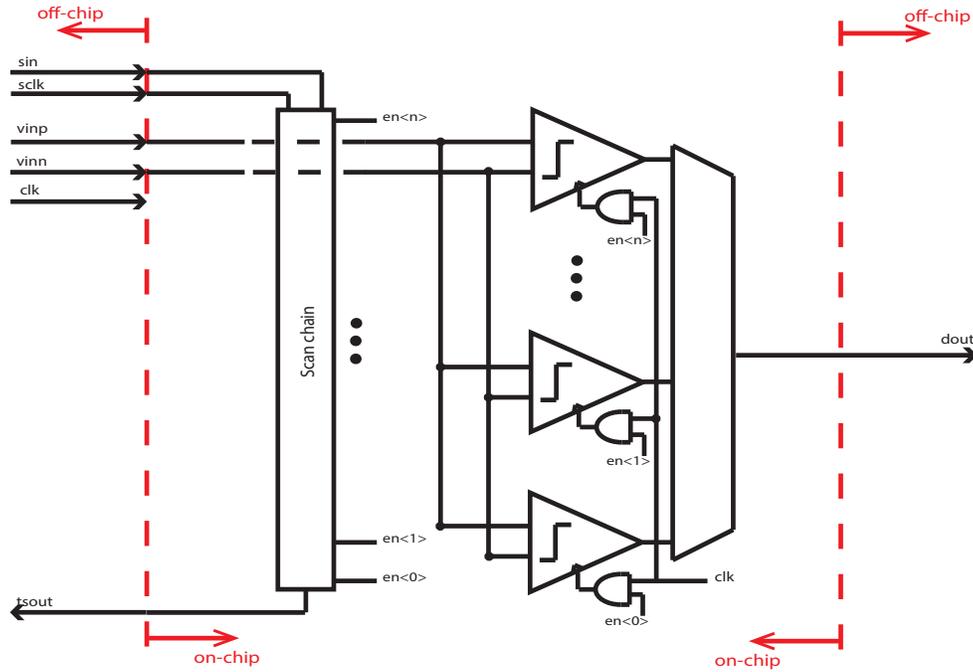


Figure 4.3: One column of the offset measurement array.

4.2.1 Offset characterization array

Offset characterization is a low-speed measurement performed with a simple test structure like the one shown in Figure 3.17. The challenges here are acquiring a large sample size and dealing with noise in measurements.

In order to acquire measurement of multiple comparators we implement a test structure array, similar in concept to the one presented in [11]. The primary reason for employing such an array is that it provides a large amount of comparators, thus making accurate statistical analysis possible. Moreover, the array, in combination with row/column decoders, helps overcome the problem of limited pad number. Finally, it enables statistical analysis with both low and high spatial resolution and reduces design time due to its repetitive nature.

Figure 4.3 shows one column of the array. Each column is repeated multiple times and a column scan chain is added which, in combination with the row scan chain, produce the en signal that selects or de-selects a comparator. The column and row scan chains are connected in one long chain, therefore sharing the input and output scan signals and minimizing pad overhead. All comparators share the same input signals $vinn$ and $vinn$, clock clk and a single output signal $dout$. The clock and input wires are laid out in an H-tree form in order to equalize the wire lengths and avoid systematic variation and skew. Each comparator output is connected to the output wire $dout$ through a tri-state buffer, so as to reduce area

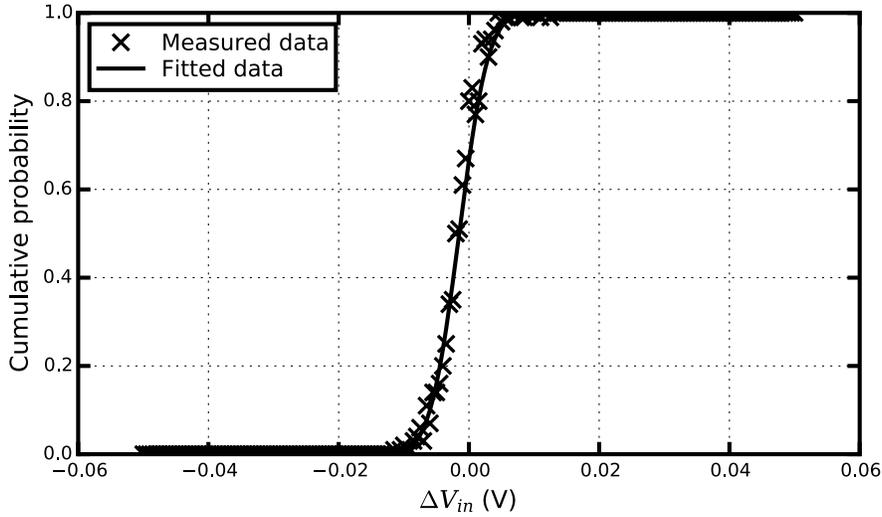


Figure 4.4: Example of a cumulative probability function of a single comparator, after noise averaging.

overhead and wiring complexity of the multiplexer. The final output and the clock output of the offset array are driven to a 30kbit shift-register, and also directly to two digital output pads. This built-in redundancy ensures better testability and facilitates debugging during the measurement process.

The complete offset array, including the comparators and the peripheral circuitry, occupies $1.25mm \times 1.35mm$ of space and consists of 56 rows and 64 columns, giving a total of 3,584 comparators.

To measure the offset, one of the comparators is selected and the input is swept slowly. For each differential ΔV_{in} , N measurements of the output are taken, and the probability $P(dout = 1)$ is calculated. N can be 30000, if the on-chip memory is used, or can be set to any number if the $dout$ output pad is used directly, bypassing the memory. The resulting plot of $P(dout = 1)$ against ΔV_{in} is fitted to a cumulative distribution function. An example cumulative distribution function for a single comparator is shown in Figure 4.4, after noise averaging over 30000 samples. The mean of the noise distribution is the comparator offset.

4.2.2 Impulse sensitivity function characterization array

The basic simulation setup for ISF measurement was shown in Figure 3.18. To translate that design to hardware, we have to consider many practical issues, like testability, bandwidth and accuracy limitations, and noise.

In order to allow for high-speed testing, we save the output in an on-chip memory, the

size of which is determined by N . As a reminder, N is the number of output samples needed when the input has a period T_{ref} and the clock has a period $T_{ref} + \Delta\tau$ (Equation 3.25). In order to get samples with better resolution i.e. lower $\Delta\tau$, more outputs need to be saved, and therefore the memory has to be bigger. The choice of $\Delta\tau$ depends on the frequency of the comparator that is being characterized. According to the Nyquist-Shannon sampling theorem:

$$\Delta\tau \leq \frac{1}{2f_{comp}} \quad (4.1)$$

Table 4.1 shows various combinations of f_{ref} and N , and their corresponding time step. In order to maintain a 1ps resolution, many different combinations can be used. Although in simulation high frequencies are preferable because they shorten the runtime, in a practical design a lower frequency should be selected in order to avoid the difficulties of providing a low-jitter high-speed clock. The tradeoff, again, is a on-chip larger memory. To allow for simplicity and flexibility at a reasonable area, we design a 30kbit shift register which is connected to the output of the comparator array, as shown in Figure 4.5. The shift-register is shared with the offset characterization array through a multiplexer. The ISF measurement array is otherwise designed in the same way as the offset measurement array, and signal redundancy is also implemented for testing purposes..

In this setup, each ISF measurement will characterize not only the comparator itself, but also the input channel before each comparator. This input channel consists of the board input and trace, the bondwires, the loading of the chip I/O pad as well as the loading of the termination resistor and the comparator array itself. All of these components significantly degrade the measured bandwidth. For this reason we add two probe-pads on-chip, placed at the input of the ISF array. The resulting signal path is shown in Figure 4.6. The ISF measurement reveals $h_{meas}(t)$. Although the addition of the probepad further loads the inputs, it also enables us to measure $h_{chan}(t)$. The deconvolution of the two can give a good approximation of the ISF. In order to reduce wire resistance after the probepad and comparator loading, we keep this array smaller, with a total of 160 comparators.

Table 4.1: Time step $\Delta\tau$ for various combinations of f_{ref} and N for ISF characterization.

$f_{ref} \setminus N$	1Kb	10Kb	100Kb	1Mb
1MHz	1ns	100ps	10ps	1ps
10MHz	100ps	10ps	1ps	0.1ps
100MHz	10ps	1ps	0.1ps	0.01ps
1GHz	1ps	0.1ps	0.01ps	0.001ps

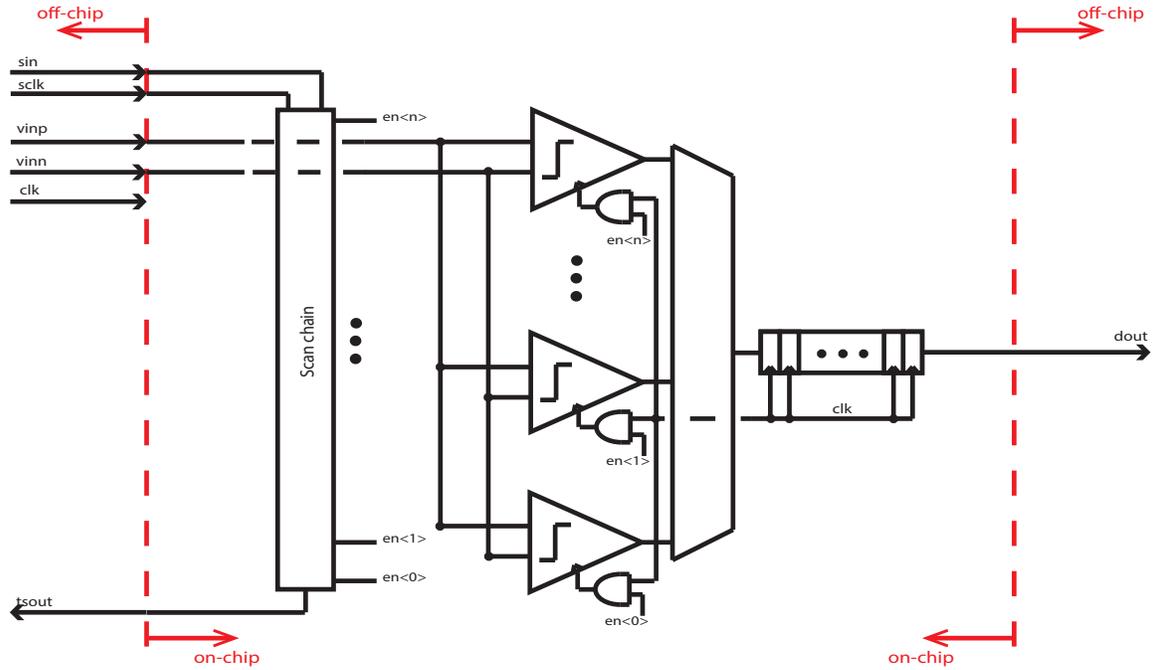


Figure 4.5: One column of the ISF measurement array.

4.2.3 Device characterization structures

Along with offset and ISF structures, transistor arrays are used in order to explore variation and characterize technology. Two arrays were designed, one for NMOS and one for PMOS testing. In each array, all devices under test (DUTs) were connected in parallel to each other, sharing the same gate, source, drain and body buses, as shown in Figure 4.7. Two buses are used per source/drain terminal, in order to enable four-terminal sensing. During measurement, only one of the devices is enabled through the enable signal en , which activates a pass-gate switch that connects the gate bus V_G to the actual DUT gate, and also connects the drain and source to the sense buses. All other devices are disconnected from the sense buses and their gates are driven to a voltage V_{GX} .

Configuring the DUTs in such an array has multiple advantages. Firstly, it allows for a large number of devices to be tested. The shared buses reduce complexity and eliminate the need for additional multiplexing, while a single shared scan chain can be used to generate the enable signal for both arrays, therefore relaxing pad number limitations. Secondly, it allows for leakage control in two different ways. When one device is selected, the gates for all other devices are driven to V_{GX} . V_{GX} can be set slightly below ground for NMOS, or slightly above the supply voltage for PMOS, providing a reverse bias to the gate that reduces leakage. Figure 4.8 shows simulated I-V curves of the a single NMOS device which is part of

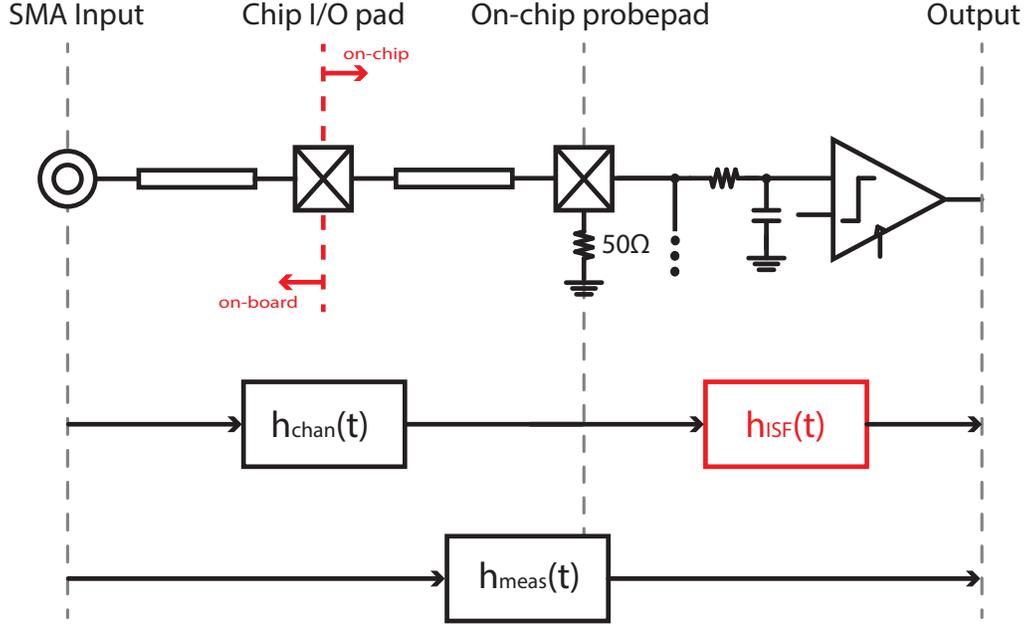


Figure 4.6: Input signal path consisting of SMA input, PCB trace, bondwire, chip input pad, probepad, wires and the comparator.

a larger array of devices. When no reverse gate bias is used there is a leakage floor at $100\mu A$, whereas when a reverse gate bias of $-100mV$ is used it is possible to measure currents in the nA range. Additionally, when all devices are deselected a leakage measurement can be used to calibrate leakage current out of the actual DUT measurements. Finally, the existence of separate pairs of current-carrying and voltage-sensing electrodes enables more accurate measurements as the wire and contact resistance is removed from the measurement.

4.3 Characterization and comparison of comparator topologies

Along with design centering the chip targets at characterizing design variability among different topologies and effects of different layouts, to allow for design-specific and layout-specific effects to be studied.

Regarding topology selection, we design a wide variety of the most common clocked

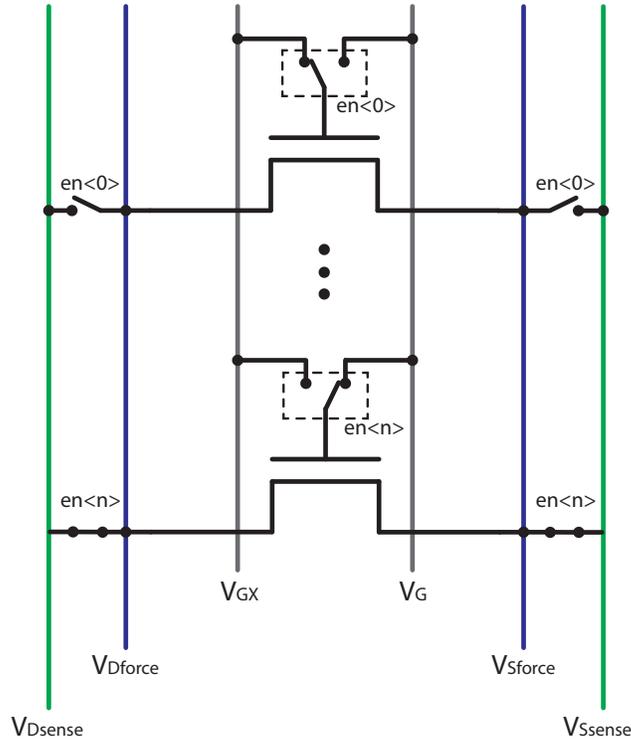


Figure 4.7: Schematic design of the transistor array for Kelvin-sensing measurement of a large number of DUTs.

comparator designs. All included topologies are shown in Table 4.2 and are assigned the abbreviated names shown. This includes the strong-arm latch comparator [47], which is broadly used as a low-power high-speed comparator, and was discussed in detail in Chapter 3. The double-tail latch (INT2) [54] was developed for low-power deep-submicron applications, as it features reduced device stacking by breaking up the latch into two stages. Another variation that uses a single clock phase (INT1) was presented in [55] as a low-voltage low-noise alternative. Finally, two new variations of the double-tail latch were developed, which use two (DSA2)¹ and one (DSA1) clock phase respectively, and combine the advantages from both comparator types. The dual-SA latch is appropriate for ultra high-speed operation as it uses two stages of positive feedback, while the dual design allows for lower offset as offset is now the result of a joint distribution, while power-wise it is comparable or better than other two-stage designs. All comparators are designed to have the same input and output capacitance, and sized for $\sim 10GHz$ speed under nominal conditions.

¹designed by Vladimir Milovanović in Berkeley Wireless Research Center, University of California at Berkeley

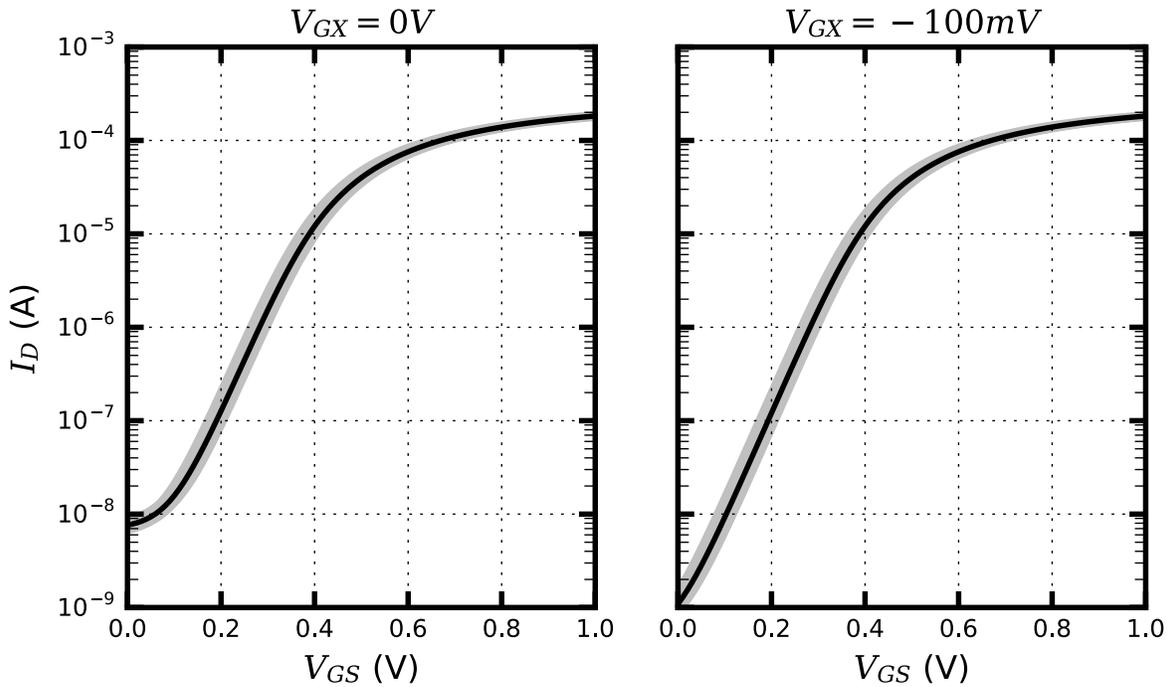


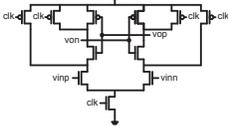
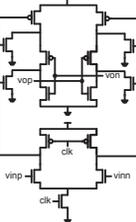
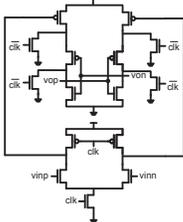
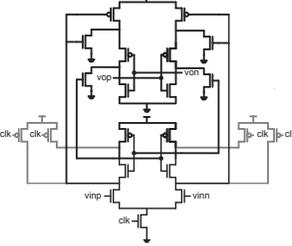
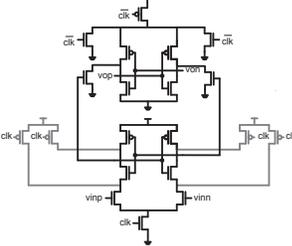
Figure 4.8: Monte-Carlo simulation showing the effect of reverse gate bias for leakage reduction.

4.4 Characterization of random and systematic variability

In the designed test-chip, several variation effects have been targeted for characterization both in the context of comparators as well as individual devices. In Chapter 2, many of the most well-known sources of systematic and random variation were discussed, and variations were classified to within-die (WID) and die-to-die (D2D) variations. In this work, we target at characterizing random variation as well as some of the lesser-known and less explored sources of systematic variation in order to determine their importance, their effect on circuit design locally and across different dies, and for customized model building.

Table 4.3 outlines all the layout variations and device geometries included in the test-chip, and shows the abbreviated name assignment that they were given. Those layout variations are applied on the input devices and/or the clock device of a comparator design, and are also included in the test-chip as individual devices in the device characterization array. All devices are of minimum length. The last column of the table shows the targeted variation

Table 4.2: Comparator topologies included in the test-chip

Comparator type	One clock phase	Two clock phases
SA		N/A
INT1, INT2		
DSA1, DSA2		

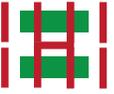
effect that the corresponding device will be used to characterize, systematic or random. Those targeted effects are discussed in detail in the following subsections.

4.4.1 Random variability test structures

Random variability sources such as random dopant fluctuations, gate work function variation, and line-edge roughness can contribute to variations in V_{TH} , I_{OFF} , and I_{ON} between devices with identical layouts. In order to isolate the impact of random variability from

that of systematic variability, transistor pairs (i.e. mismatch test structures) are often used. These test transistors are identically drawn structures that are placed in close proximity to one another on the chip. If there is a systematic source of variability, its impact would be the same for both devices. As a result, when the difference (as opposed to the absolute value) of the performance parameter between the two transistors in a pair is analyzed, the impact due to systematic variability is canceled out, i.e. the difference is due entirely to random variability. To ensure that the transistors in a pair are identical in every possible aspect, it is important to make sure that the surrounding area is the same for both transistors. For this reason, transistors in the device characterization array are layed-out in columns; this allows for transistors of the same flavor to be in close proximity to each other, while the guard rings and the peripheral circuitry are kept identical, as shown in Figure 4.9.

Table 4.3: Layouts and geometries of devices under test

Name	W (nm)	Layout	Targeted effect
N0	1240	Area = $0.0372 \mu m^2$	Random variation, device geometry
N1	1240		Number of fingers, corner rounding
N2	1240		Segmented channel mobility variation
N3	1240		Mobility variation due to gate proximity
N4	2480	Area = $0.0744 \mu m^2$	Random variation, device geometry
N5	2480		Number of fingers, corner rounding
N6	310	Area = $0.0093 \mu m^2$	Random variation, device geometry
N7	620	Area = $0.0186 \mu m^2$	Random variation, device geometry
N8	4960	Area = $0.1488 \mu m^2$	Random variation, device geometry

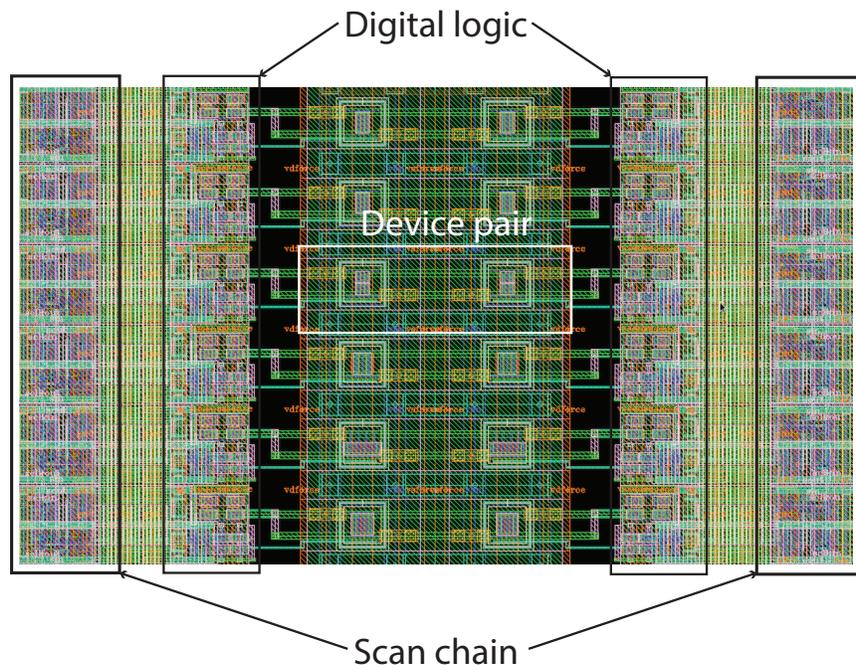


Figure 4.9: Part of the device characterization array layout

It is known that device variability is inversely proportional to device area [56]. However this approximation is based on the square law transistor current model and becomes less and less valid at smaller device geometries. In order to assess the effect and characterize the Pelgrom coefficient for this technology, devices of various geometries are included in the device array. The various channel areas that were characterized are shown in Table 4.3. The selected sizes and layouts match those of the devices used in the comparators in order to explore correlations between I-V curves and comparator performance.

4.4.2 Systematic variability test structures

In addition to device structures used to study random variability, several device structures are included to assist with the study of a few different effects of systematic variability. From Table 4.3, device N0 is used as the reference, designed with 2 fingers, a gate pitch of 106nm and a fixed STI distance.

Firstly, we explore the effect of number of device finger selection. It is expected that the variation of multi-fingered devices of a fixed area will exhibit systematic shifts across different dies due to effects like poly corner rounding. Carefully designed structures can help identify and model the effect.

Another targeted effect comes from the selection of gate pitch. As gate pitch is scaling,

it is becoming harder to accurately control the length of devices, which is why design rules get more restrictive. Design rules now require the addition of dummy gates, but allow a few different options on the gate pitch used. In sub-wavelength lithography narrow poly lines with varying pitch will have different channel lengths, while dense lines have higher depth of focus, and are more immune to defocusing of the optical system [57]. These dummy poly structures can influence the stress within the channel region of the device-under-test. In [58] it is shown that different gate pitches can add deterministic shifts in device electrical performance due to mobility difference. In this test-chip, transistors of two different gate pitches, 106nm and 222nm, are used in order to assess any associated systematic effects.

Shallow trench isolation (STI) is used to electrically isolate adjacent transistors. Traditional methods use S_iO_2 in the STI trenches, which create compressive strain on the channel substrate that varies with distance from the edge of the STI/diffusion interface to the channel region [59]. This can affect device carrier mobility and therefore the drive strength. To quantify the impact of STI-induced stress, dummy active regions are drawn at different distances away from the device under test.

The final targeted effect is that of a segmented transistor channel. Instead of a transistor having a continuous width, the channel region can be segmented into multiple stripes of equal width. From an electrostatic control standpoint, a segmented channel transistor can offer improved short-channel effect due to the slight wraparound of the gate over the channel and the gate fringing electric field coupling to the channel region through the STI, if the stripe width is comparable to the channel length. Thus, even though the segmented channel design takes up more layout area as compared to a conventional channel design, the improvement in device performance can provide a net benefit when normalized to the same layout area [60]. Additionally, larger and more uniform mechanical stress can be induced within narrow channel segments.

4.5 Chip overview and testing

The test-chip was fabricated in a 28nm FDSOI technology provided by ST Microelectronics. The final layout is shown in Figure 4.10. The chip uses approximately a $2mm \times 2mm$ area with 31 IO pads per side. The nominal supply voltage for this technology is $V_{DD} = 1V$, and the IO supply is $V_{DDE} = 1.8V$.

From the included test structures the IV array is independent, running from a separate supply V_{DDX} . The comparators in the comparator arrays also run off a separate analog supply V_{DDA} , while all other logic on chip uses the nominal supply. The clock and signal outputs of the two comparator arrays are routed to the on-chip memory through a multiplexer, while also buffered versions of them are routed directly to the pads for added testability and easier debugging. For the ISF array, the clock input can be supplied either through a digital pad or an LVDS pad, which allows for higher-frequency testing.

All chip voltage supplies come from programmable regulators on a motherboard. The motherboard interfaces through FMC connectors with an Opal Kelly Shuttle LX1 board

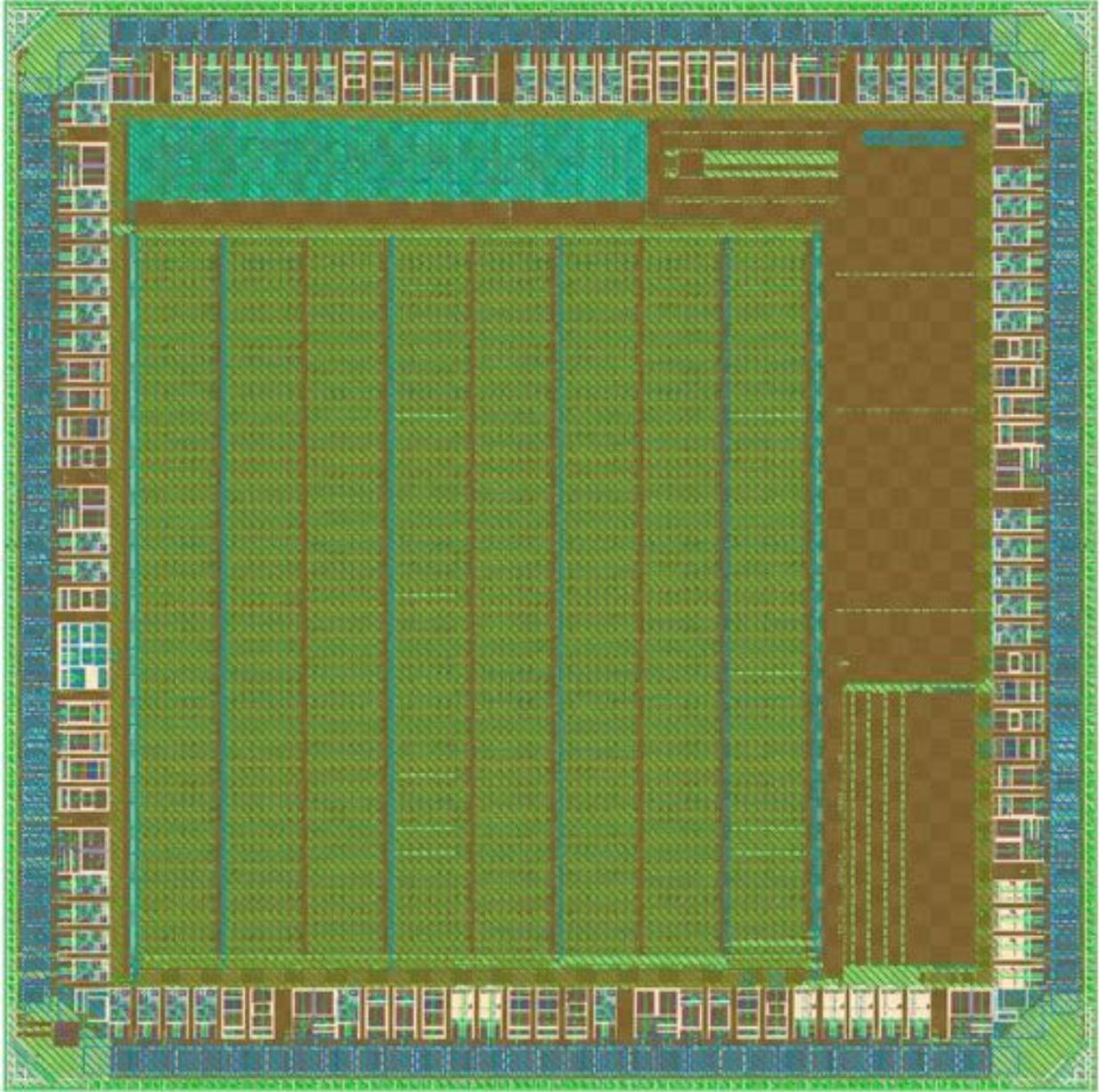


Figure 4.10: Layout picture of the complete chip.

and with a $4' \times 4'$ daughterboard. The complete board setup is shown in Figure 4.11. All chip digital signals are controlled through the Opal Kelly FPGA. All chip analog signals are supplied by lab instruments using SMA connectors on the daughterboard. The FPGA and the lab instruments are all controlled by a single python script in order to automate the testing process.

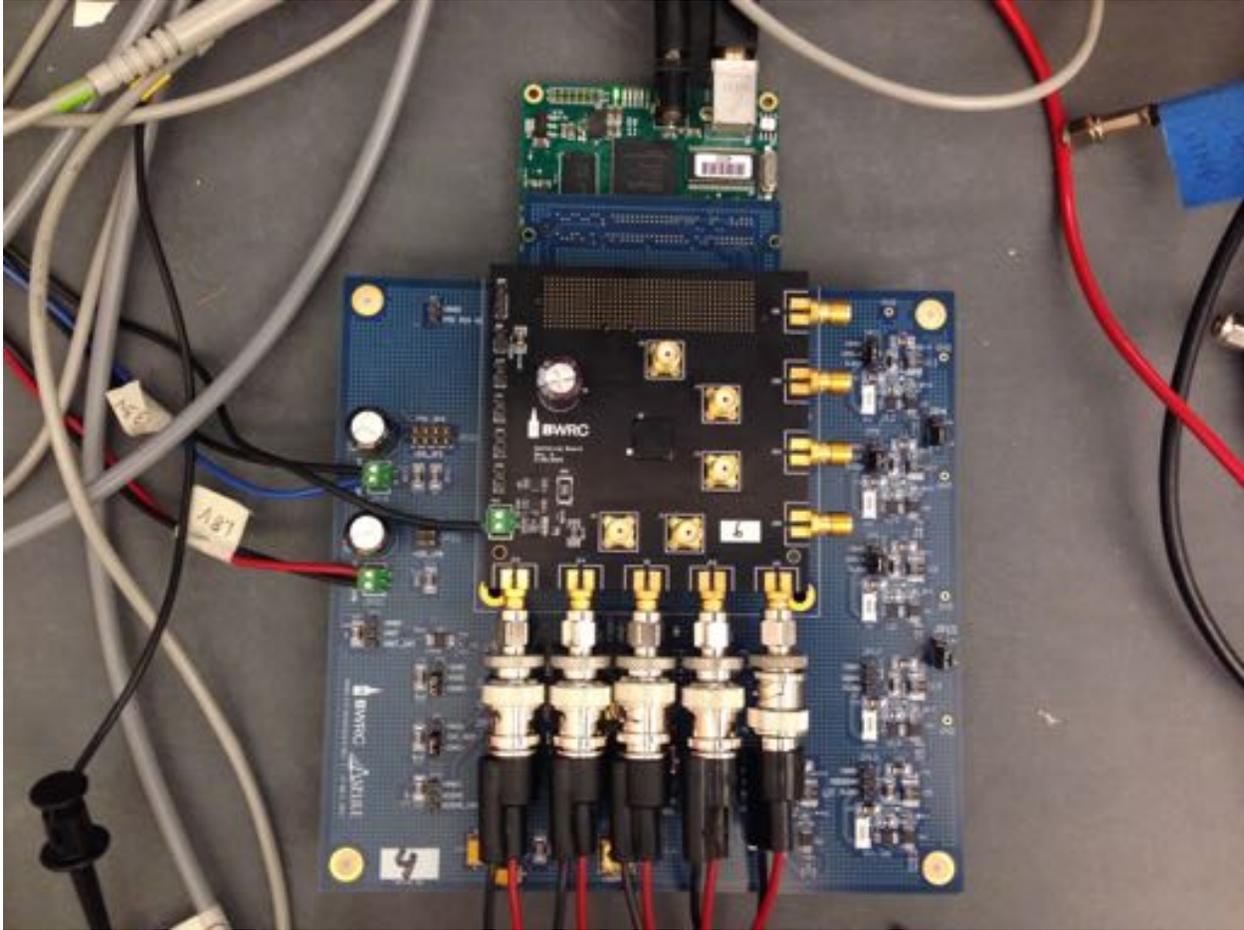


Figure 4.11: Test board setup, including motherboard, daughterboard and Opal Kelly Shuttle LX1 board.

4.6 Summary

A test chip vehicle is designed in order to study the impact of device variability. Device characterization of different geometries and layouts can reveal information about both random and systematic variability. In addition to that, sets of comparator arrays of multiple topologies and layouts are used to assess systematic effects. Along with variability characterization, the dataset can then be used for customized model building, resulting to design-specific, high-yield models.

Chapter 5

Experimental evaluation of customized models

A testchip has been designed and fabricated in a 28nm FDSOI technology and measured. The chip is shown in Figure 5.1, and contains the offset and ISF characterization arrays as well as the device characterization array described in Chapter 5.

Measured data reveal information about variability in thin body devices, including the impact of layout in D2D and WID variation, as well as the impact of layout and topology in comparator variation, and are presented and analyzed in Sections 5.1 and 5.2. In Section 5.3, the variability data extracted from comparator measurements are used to create customized models to improve yield prediction error. The chapter concludes with a summary of the results.

5.1 Characterization of device variability in 28nm FDSOI

Transistor current-voltage characteristics were extracted from the designed arrays. The scan chain signals were software-generated and passed to the chip through the FPGA. All data post-processing was done in Python.

For device I-V curve extraction, a reverse gate bias of -100mV was applied to the gates of the NMOS devices that were not under test in order to reduce leakage. Additionally, an I-V curve measurement when all devices are off was taken and was subtracted from each measured I-V curve, in order to reduce the leakage floor. The voltage threshold was then extracted using the constant current method on $I_D(V_{GS})$ curves at the linear region [61],

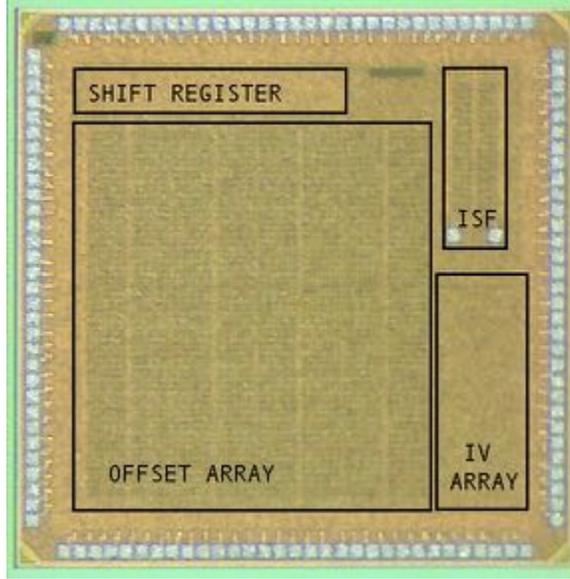


Figure 5.1: Die photo of the chip, showing the various blocks described in Section 4.2.

according to which the estimated threshold voltage is the value of V_{GS} when $I_D = 10^{-7} \frac{W}{L} A$ and $V_{DS} = 50mV$, as shown for an example device in Figure 5.2.

A total of 13 dies were measured. Statistics of all raw data measured are shown in Figure 5.3, including both systematic and random effects of variation. In the following sections we will differentiate and analyze effects of systematic and random variability and investigate random mismatch, random within-die variability and systematic effects of variability across different dies.

5.1.1 Matching performance

In order to analyze random mismatch, device pairs are measured across all 13 dies and the voltage threshold is extracted from their corresponding $I_D(V_{GS})$ curves. Then the variance of ΔV_{TH} is calculated as:

$$\sigma_{\Delta V_{TH}}^2 = E\{(V_{TH,1} - V_{TH,2})^2\}$$

where 1 and 2 are the indices of the two matched devices. Figure 5.4 shows the measured and simulated Pelgrom plot for this technology. The measured data come from 160 NMOS devices per chip per plot point, and the devices used are N0, N4, N6, N7 and N8 (Table 4.3). In both the simulated and the measured case, a straight line was fit through the data using least-squares optimization. From the slope of each line we extract the Pelgrom coefficient, which is estimated at $1.93mV\mu m$ for simulation and $1.51mV\mu m$ for measurement. The Pelgrom coefficient is a measure of transistor matching, and the lower it is the better the matching [56].

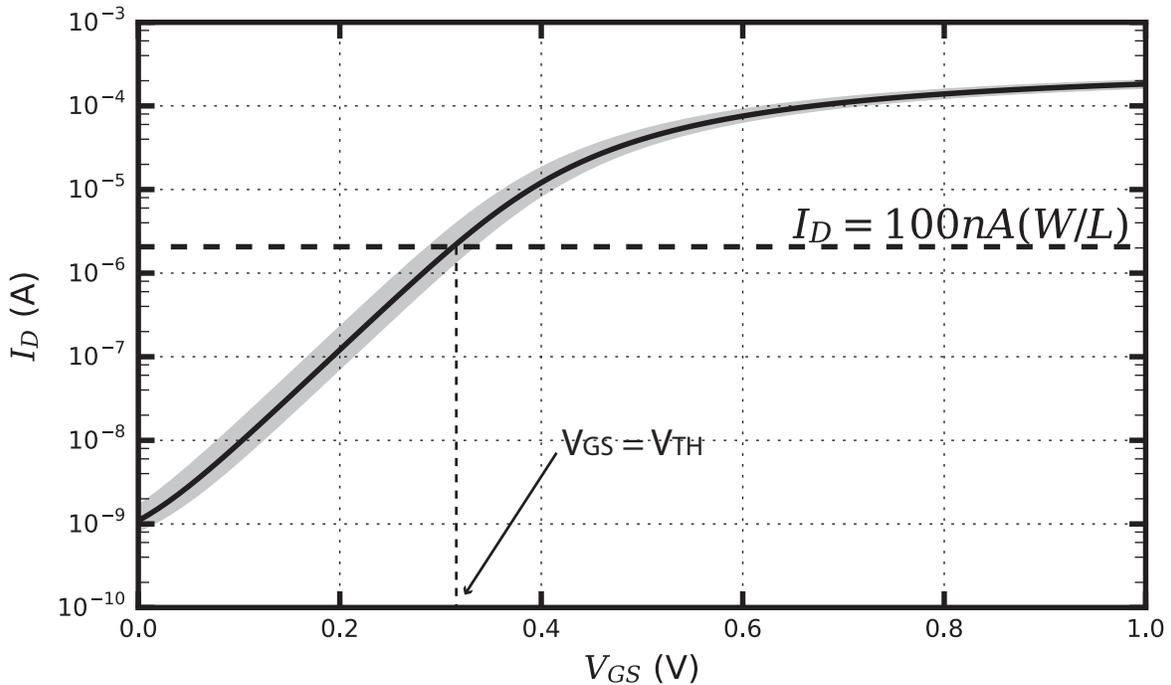


Figure 5.2: The constant-current voltage threshold extraction method.

From the plot, firstly we observe that the models overestimate the Pelgrom coefficient by approximately 30%. This is likely due to the fact that statistical parameters for variability models are typically extracted using variation data with different sources of variability lumped together.

Secondly, we observe that the Pelgrom coefficient for FDSOI technology is very low, which is consistent with previously published data for this technology [62–64]. This is due to the fact that planar FDSOI does not use dopants to set the voltage threshold, making the contribution of random dopant fluctuation to variability virtually zero. The threshold voltage is set by gate work-function engineering [62], while excellent electrostatic control of the channel is achieved due to its thin body.

5.1.2 Within-die variability

Process variations that vary rapidly within the dimension of a die cause within-die (WID) variability [57]. WID variability can have systematic sources, resulting from the processing and mask imperfections, in two cases: 1) when the die is very large relevant to the wafer and 2) when the process variation has a strong layout dependence [9, 65]. Due to device scaling

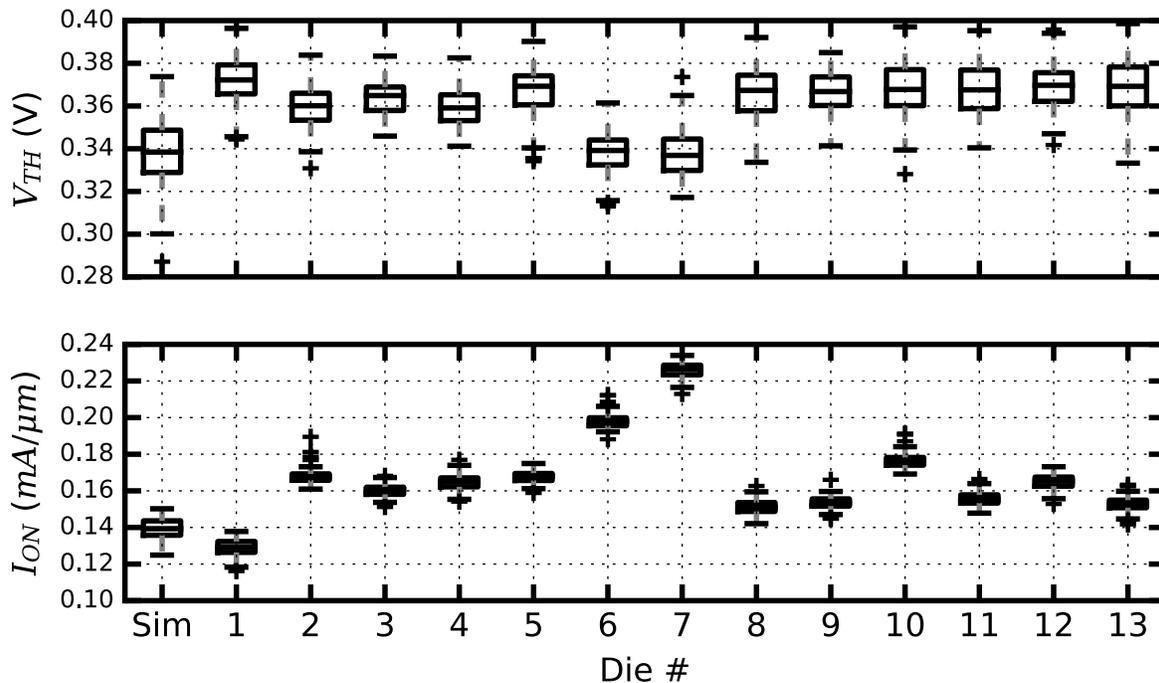


Figure 5.3: Distributions of measured V_{TH} and I_{ON} for a 310nm/30nm device across different dies, compared to the simulated distributions.

and shrinking of dimensions, currently most WID variations are treated as random, either because they are due to random sources or because their sources are unknown.

Analyzing the performance distributions of identical structures within a die reveals the WID variation for that die. Figure 5.5 shows the $3\sigma/\mu$ percent variation of the measured V_{TH} and I_{ON} of the smallest available NMOS device for each measured die. The average WID variation measured is 8.9% for V_{TH} and 6.5% for I_{ON} , while their worst-case is 11.1% and 9.4% respectively. WID variation scaling is inversely proportional to area and consistent with previously published data in larger technology nodes [4, 57].

In order to evaluate the WID variation component of different layout configurations of the measured data we plot the $3\sigma/\mu$ variation measured over different dies for each device layout. Table 4.3 shows all layout variations included and their corresponding abbreviations. Figure 5.6 shows the mean as well as minimum and maximum values of the measured variation.

We observe that layout variations have limited impact on WID device variability. The effect of number of fingers is negligible. A segmented channel achieves a slight decrease of the voltage threshold variation on average, with tighter distributions measured in all dies, while I_{ON} variation remains comparable to other layouts. This likely means that segmented devices exhibit slightly less gate work-function variation within a die, while any mobility

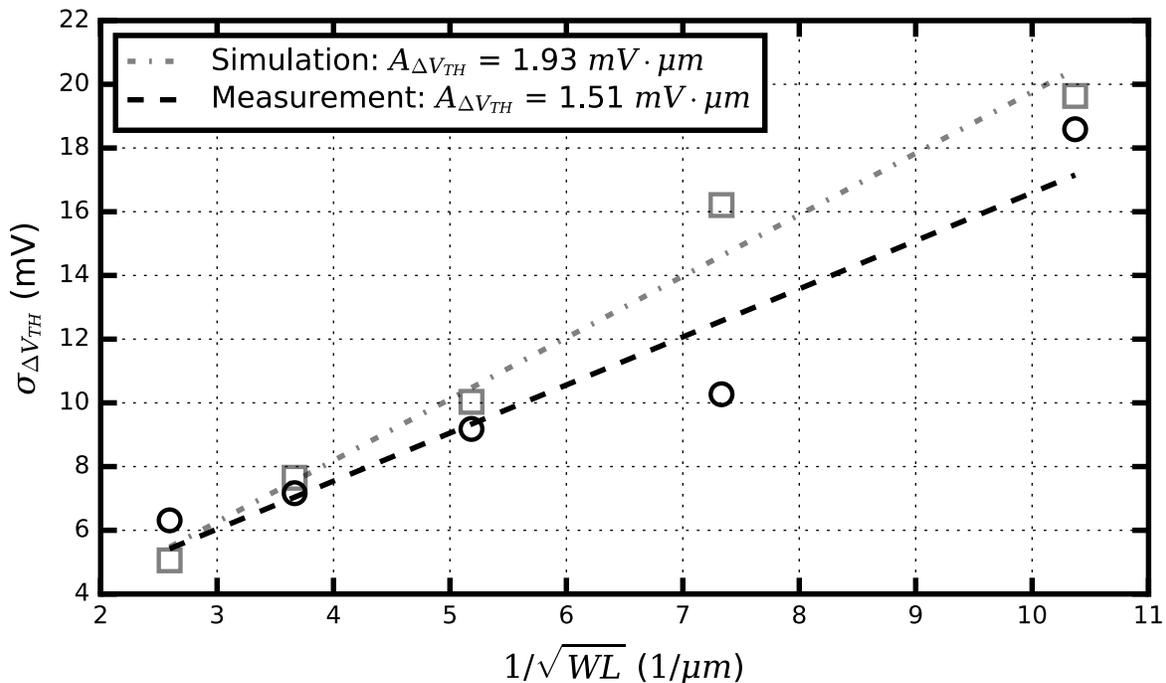


Figure 5.4: Pelgrom plot using simulated and measured data.

variation is kept low due to better electrostatic control of the channel. Finally, the larger average variation in both voltage threshold and I_{ON} is measured in the device that has increased gate pitch. The effective variability differences however are very small. Overall, systematic layout-dependent effects are not very significant, which means that the measured WID variability is mostly due to random components of variation.

5.1.3 Die-to-die variability

The sources of die-to-die (D2D) variability are generally systematic due to the manufacturing process. Photolithography and etching, oxide deposition and growth, lens imperfections and thermal annealing can all contribute to these types of variations.

Die-to-die systematic effects manifest themselves as a gradual mean shift across the wafer. The spatial characteristics of D2D variation cannot be assessed in this case as the die location on the wafer is not known. Die-to-die systematic effects related to device layout are assessed by averaging each die and examining the distribution of the means for each device layout. For the smallest device available, an 10% $3\sigma/\mu$ variation was measured in V_{TH} and a 46% shift in I_{ON} . The accuracy of these estimations is limited by the small sample size, they are

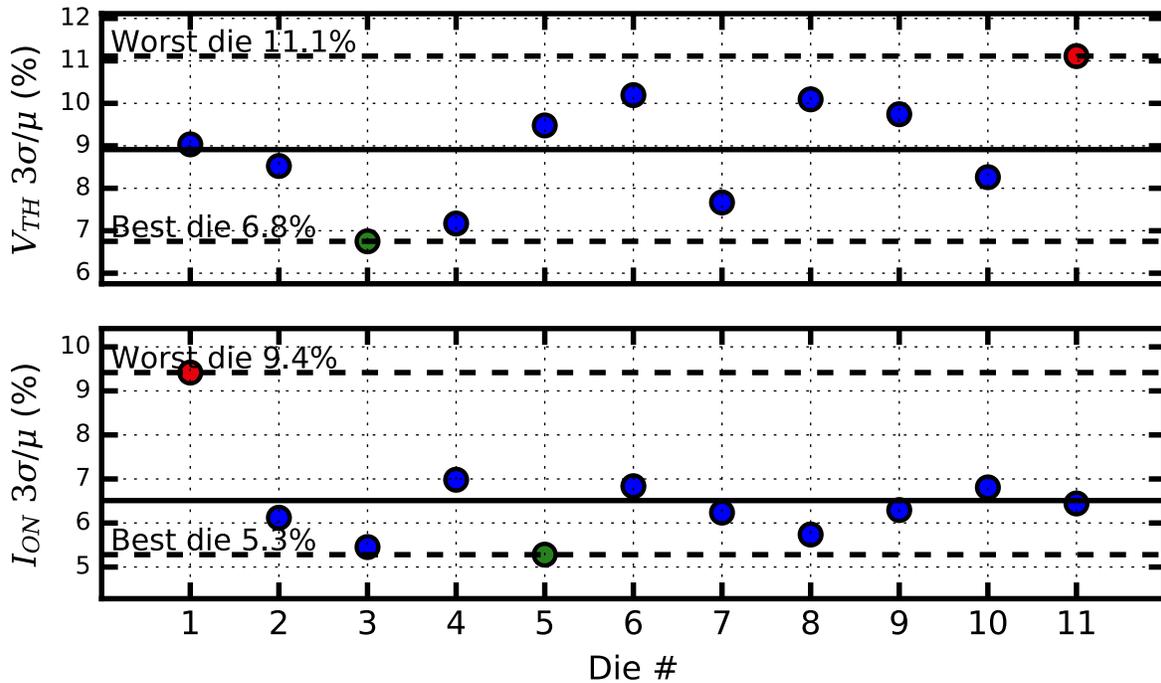


Figure 5.5: Measured WID $3\sigma/\mu$ variation for a 310nm/30nm device across different dies. The solid line shows the average WID variation, while the dashed lines mark the best and worst WID variation measured.

however in agreement with previously published data on scaled technology nodes [4, 57]. In this node, current variations appear to be dominant.

In order to investigate the systematic layout effects, Figures 5.7 and 5.8 compare the distributions of the fastest and slowest dies for V_{TH} and I_{ON} , respectively, for all different layout configurations. Device N0 in the slowest chip is used as a reference. No significant shifts in the measured $3\sigma/\mu$ variation were found in V_{TH} variation, with the larger shift occurring for the segmented device N2 and being approximately 1%.

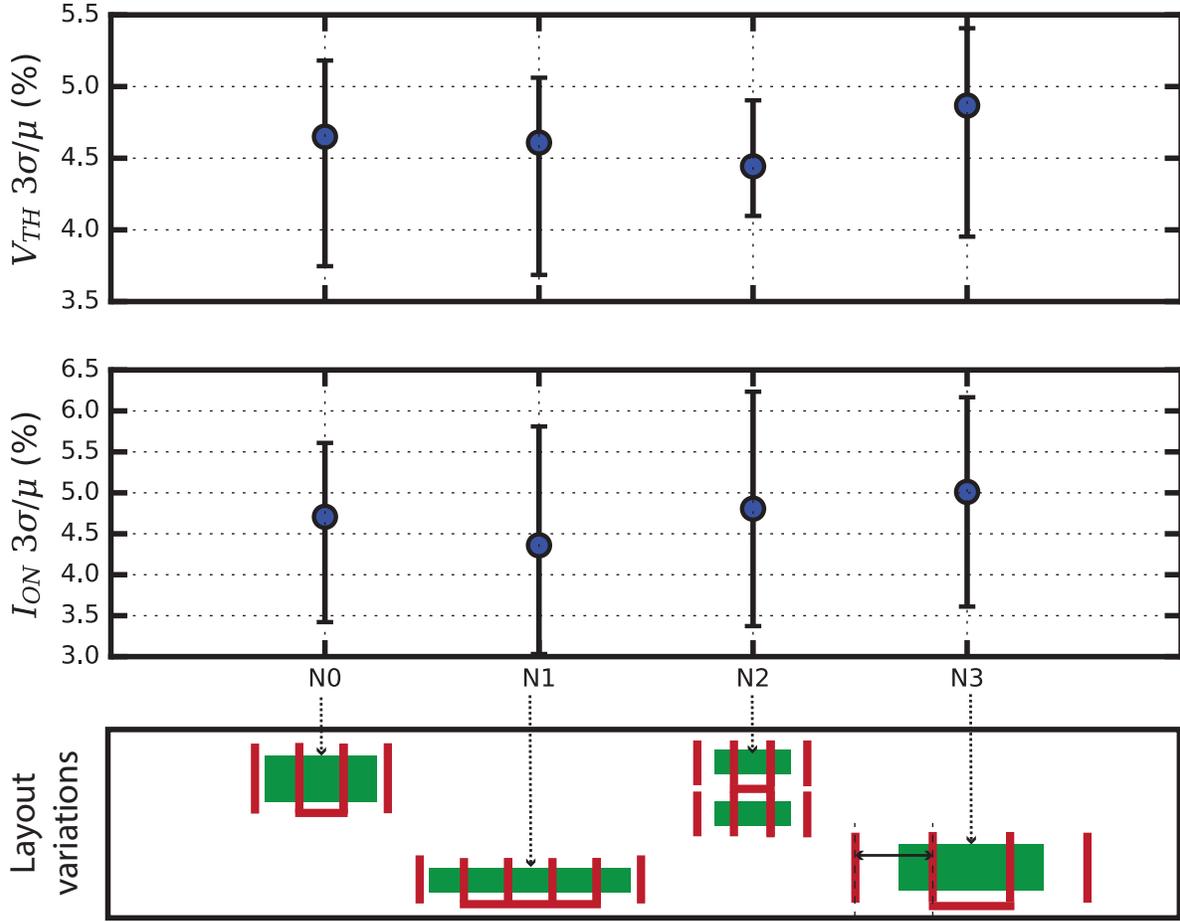


Figure 5.6: Measured WID $3\sigma/\mu$ variation from all dies across different layouts. The blue dots correspond to the mean value.

5.2 Characterization of comparator variability in 28nm FDSOI

Comparator offsets were also extracted from the designed arrays, in order to assess the effects of variability and perform design centering. The scan chain signals were software-generated and passed to the chip through the FPGA. Comparator offset measurements were taken using an FPGA generated clock of 12MHz. A slow voltage ramp swinging from -100mV to +100mV around the common-mode voltage was applied at the input, generated by a high precision source-measure unit. Supply voltage was controlled by on-board regulators. For each voltage step, 30k output bits stored on the on-chip memory were measured. The extracted points of the cumulative distribution function were fit to the cumulative distribution function of a Gaussian (Equation 5.1) using a least-squares fit, in order to extract the mean

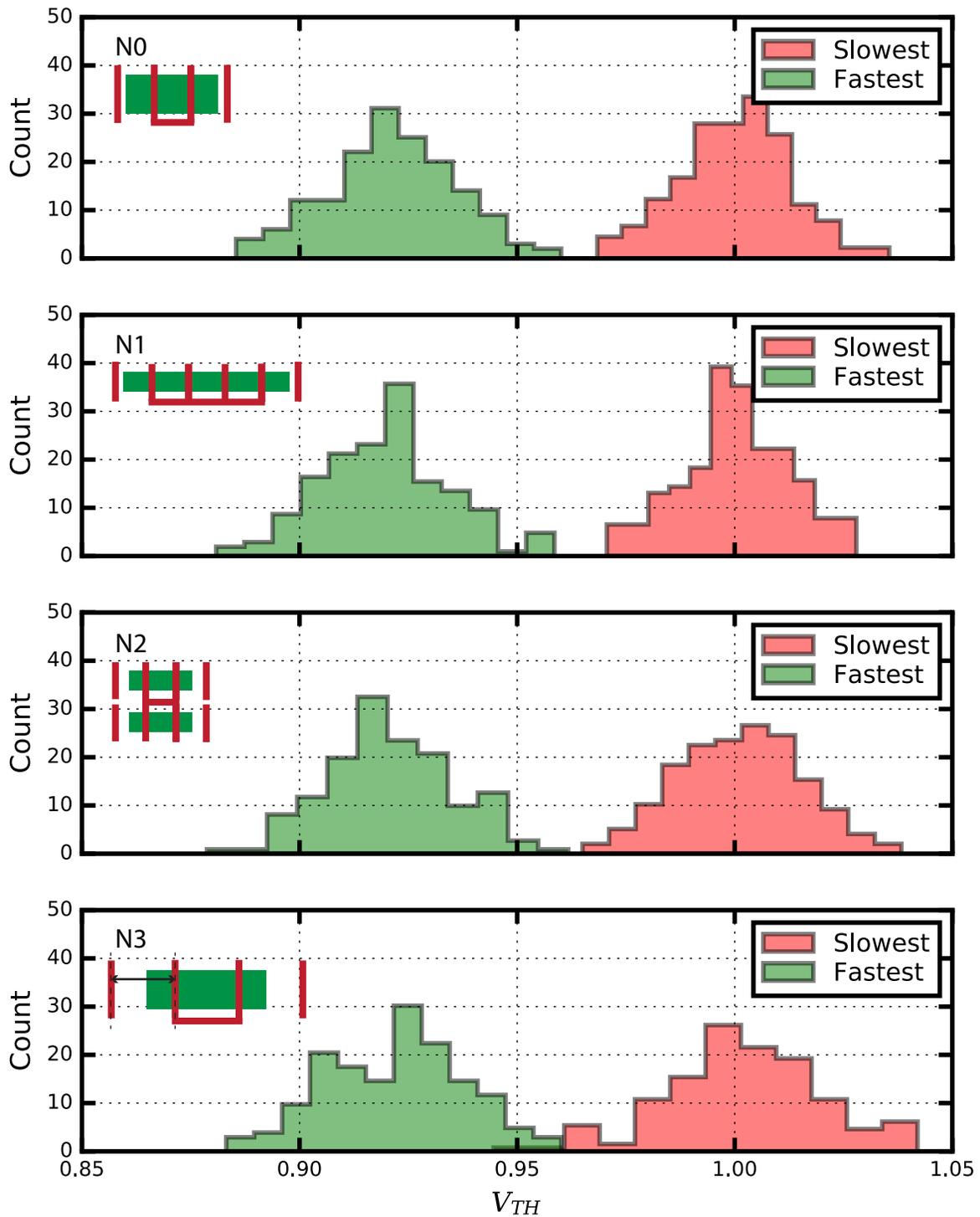


Figure 5.7: Comparison of measured V_{TH} distributions of fasted and slowest die for different layouts. Data are normalized to the mean V_{TH} value of the reference device N0 in the slowest die.

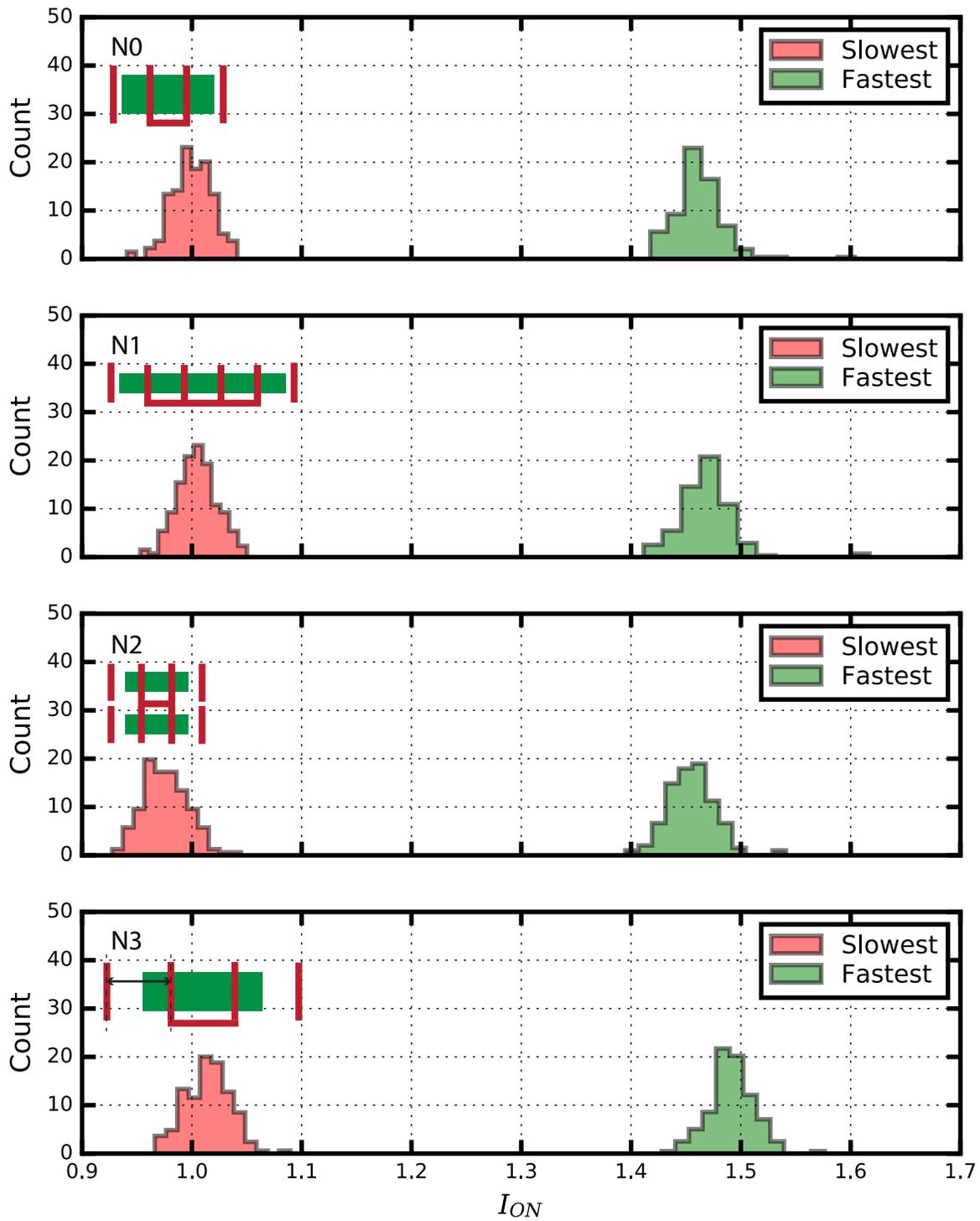


Figure 5.8: Comparison of measured I_{ON} distributions of fasted and slowest die for different layouts. Data are normalized to the mean I_{ON} value of the reference device N0 in the slowest die.

and standard deviation of the noise for each comparator, as shown in Figure 5.9(a). The extracted mean values correspond to the comparator offset.

$$\Phi(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sigma\sqrt{2}} \right) \right] \quad (5.1)$$

For each comparator type, 223 instances were measured from each die in order to evaluate the offset distribution within the die (Figure 5.9(b)). A limited number of outliers found outside the range of the input voltage step were eliminated in some cases, in order to avoid repeated time-consuming measurements. If too many samples were outside the input range, for example in the case of very low supply voltages, the measurement was marked as failed.

5.2.1 Layout-related effects of variability

We will first discuss some of the effects of comparator layout in variation and overall performance. The strong-arm comparator was used as a reference for these measurements, and layout variations were added either in the input pair or the clock device of the first stage. Table 5.1 shows abbreviated names for all measured strong-arm comparators of different layouts, and includes the type of device used for the input pair and the tail of the first stage for each one. All other devices were kept the same. Compared to Table 4.3, one additional layout variation is used in the comparator array, shown in the last column of the table. In SA7, the input device is identical to N0, but a dummy active region is placed close to the device in order to examine any stress-induced effects.

Figure 5.10 shows colormaps of the measured offsets within a die for all comparators with input device layout variations. Comparators of the same type are spaced $160\mu\text{m}$ apart in

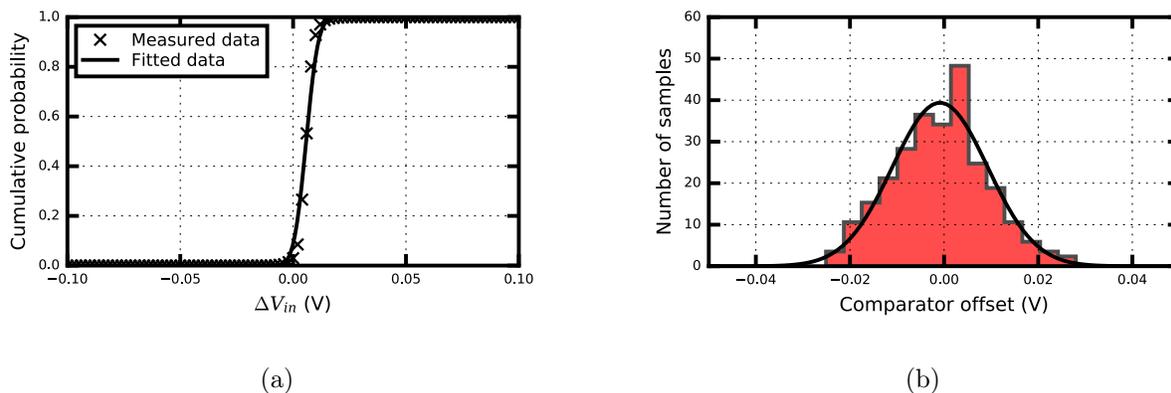


Figure 5.9: (a) Noise cumulative distribution function of a single SA comparator instance and (b) offset distribution of SA comparator within a die

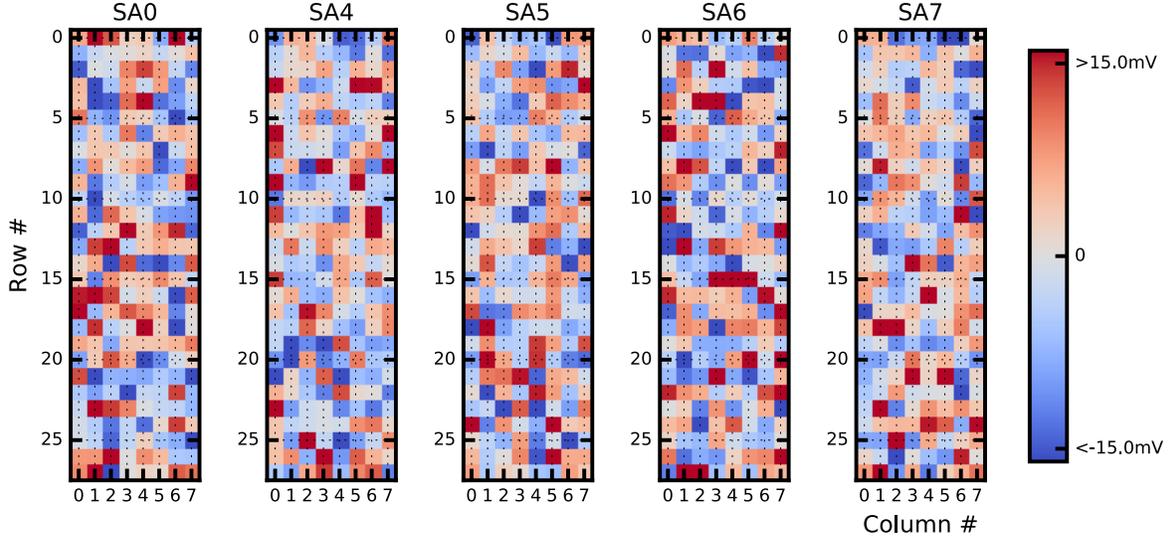


Figure 5.10: Colormaps of measured WID comparator offsets for different comparator layouts.

the horizontal direction and $45\mu m$ apart in the vertical direction. We observe no significant spatial correlation at these distances within the die.

Table 5.2 shows the percent offset shifts measured, averaging results from 6 measured dies. The offset shifts are compared to the reference comparator SA0, which has a pair of reference devices N0 at its input. The second row of the shows the layout variations used at the input devices of each one of the comparators under test. From the measured data we observe that layout variations in the input device generally do not affect device mismatch, except in the case of the device with increased gate pitch, where a $\sim 6\%$ increase

Table 5.1: Comprator layout configurations

Comparator	Input devices	Clock device	Targeted effect
SA0	N0	N8	Reference
SA1	N0	N4	Number of fingers of clock device
SA2	N0	N5	Number of fingers of clock device
SA4	N1	N8	Number of fingers
SA5	N2	N8	Segmented channel
SA6	N3	N8	Gate pitch
SA7	N0'	N8	STI effect

in the measured offset is observed. Increased gate pitch increases the distance between the channel region of the two matched devices, resulting in larger mismatch, proportional to that distance.

5.3 Variability-aware comparator design in deeply-scaled technologies

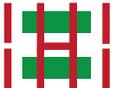
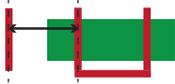
As comparators are a critical component of all mixed-signal systems, their design and optimization has always been of interest for circuit designers. While the strong-arm latch has traditionally been a great low-power solution, two-stage designs have been introduced offering potential for high-speed and low-voltage operation. Several of these designs were measured and analyzed, in order to understand how process variation affects their performance. The measured comparator topologies were are shown in Table 4.2.

Figure 5.11 shows the colormaps of measured comparator offsets within a die, comparing all topologies. As expected, there is no perceivable spatial correlation for WID variability as variation is primarily due to random mismatch. It is evident however that offset depends on topology and design decisions.

5.3.1 Variation-driven comparator topology selection

Figure 5.12 compares the measured offsets of all designed comparators that use a single clock-phase. The double-tail latch appears to have the highest offset, and even fails at very low supply voltages. The dual-strong-arm latch performs the best in terms of offset.

Table 5.2: Percent offset shifts of comparators with respect to SA0

Reference	SA0			
				
Comparators	SA4	SA5	SA6	SA7
				
Offset shift	0.9%	0.1%	5.9%	-0.3%

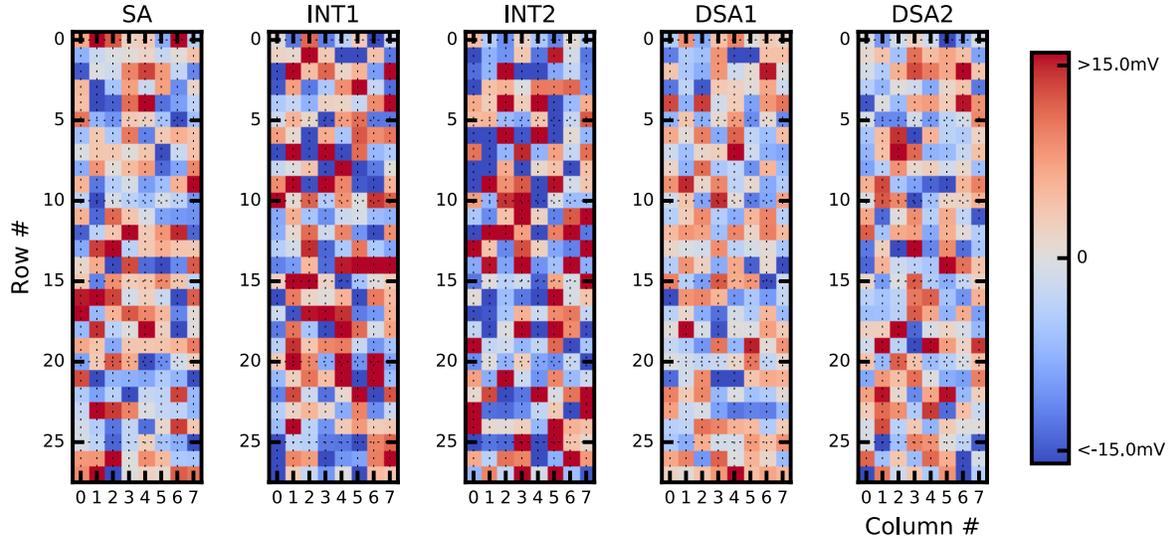


Figure 5.11: Colormaps of measured comparator offsets for different comparator topologies within a die.

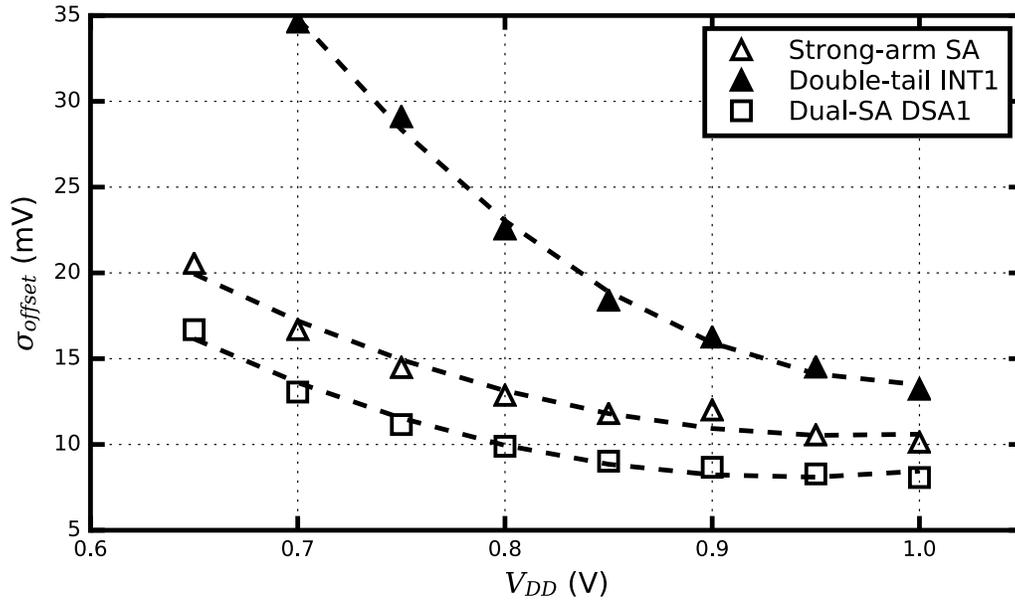


Figure 5.12: Comparison of all single clock-phase topologies.

In order to better understand this behavior, it is useful to take a look at the sensitivities of each comparator device to the device parameters. We note here that there are two device parameters in the given models, VFBO and UO. VFBO is the flat-band voltage

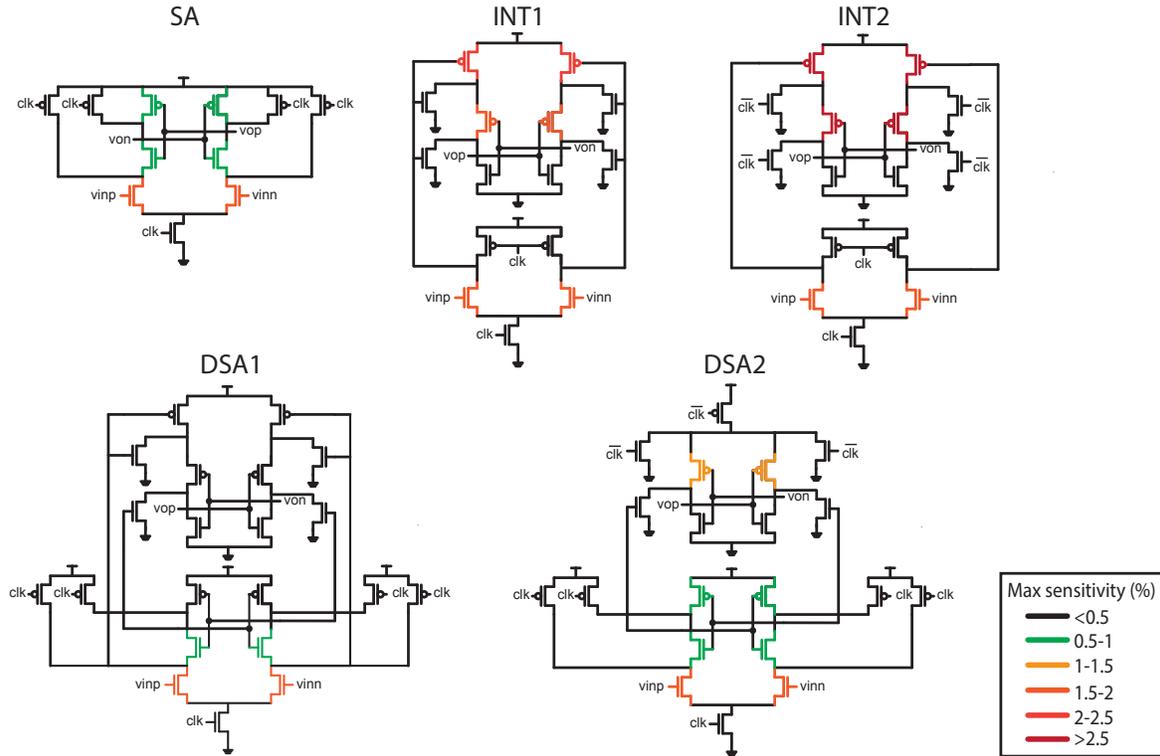


Figure 5.13: Illustration of sensitivities for each device of each comparator type. Colors correspond to the maximum sensitivity of the device.

parameter which represents variations in the voltage threshold, and U_0 is the zero-field mobility parameter which is directly related to current variations. This is illustrated in Figure 5.14 which shows scatter plots of simulated data from a $1\mu\text{m}$ -wide NMOS device.

Comparators are mixed-signal, non-linear circuits that are notoriously hard to analyze theoretically. Although several sources have attempted deriving analytical equations [66–68], these are typically based on simplifications and, especially given the increased complexity of deep-submicron device models, they are rarely accurate enough to predict design performance, let alone design sensitivity. In order to facilitate quick, flexible and effective sensitivity analysis for the circuit, we automate the process using Python in collaboration with the circuit simulator. The Python script has the ability to modify any set of desired model parameters by shifting them around their mean - or by any other desired amount - as well run simulations and gather the simulated circuit performances. This allows the script to calculate the sensitivity matrix of the circuit.

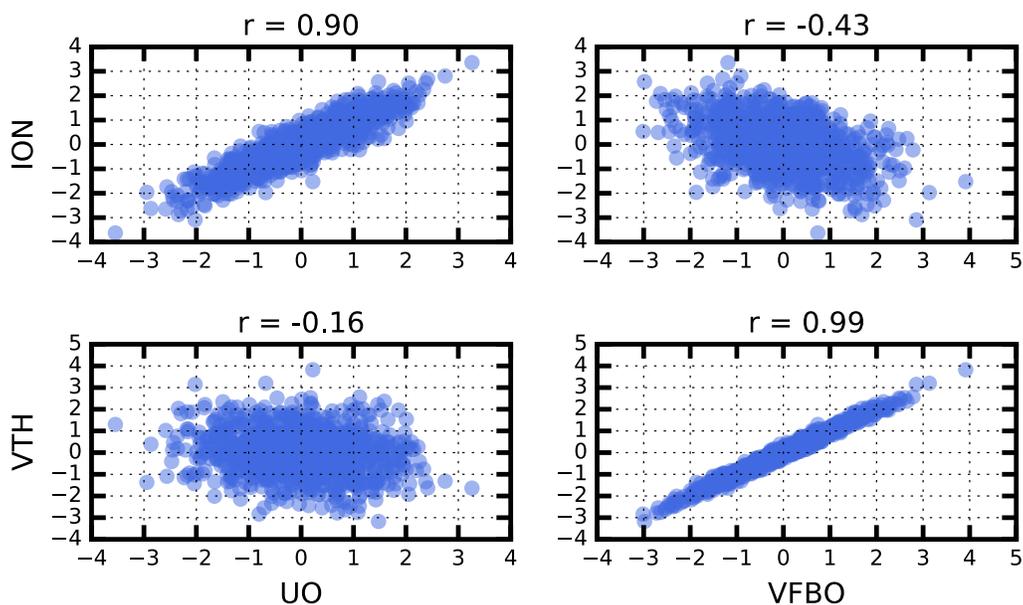


Figure 5.14: Scatter plots of simulated I_{ON} and V_{TH} with respect to model parameters. Data are normalized to zero mean and unit variance. Each plot title shows the corresponding Pearson correlation coefficient.

Figure 5.13 shows the schematics of each comparator topology, with colors assigned to the each device depending on how sensitive it is to parameter variations. The color assigned to each device is determined by the maximum offset sensitivity measured. We note that the maximum sensitivity measured was always with respect to parameter UO ; therefore sensitivity to current variations is larger than sensitivity to voltage threshold variations for all comparators.

Comparing the colored schematics of the strong-arm (SA), double-tail (INT1) and dual-strong-arm (DSA1) comparators, we observe that the double-tail topology is most sensitive to variations. Specifically, the lack of a current-setting tail device in the second stage, makes devices on the second stage very sensitive to current variations, which is why this topology has the highest measured offset. This problem is mitigated for DSA1, as in this case the first stage of the comparator is a high-gain full-swing latch. This means that the input to the second stage gets quickly amplified and brings devices to operate in the linear region for longer time, which makes them more insensitive to variations.

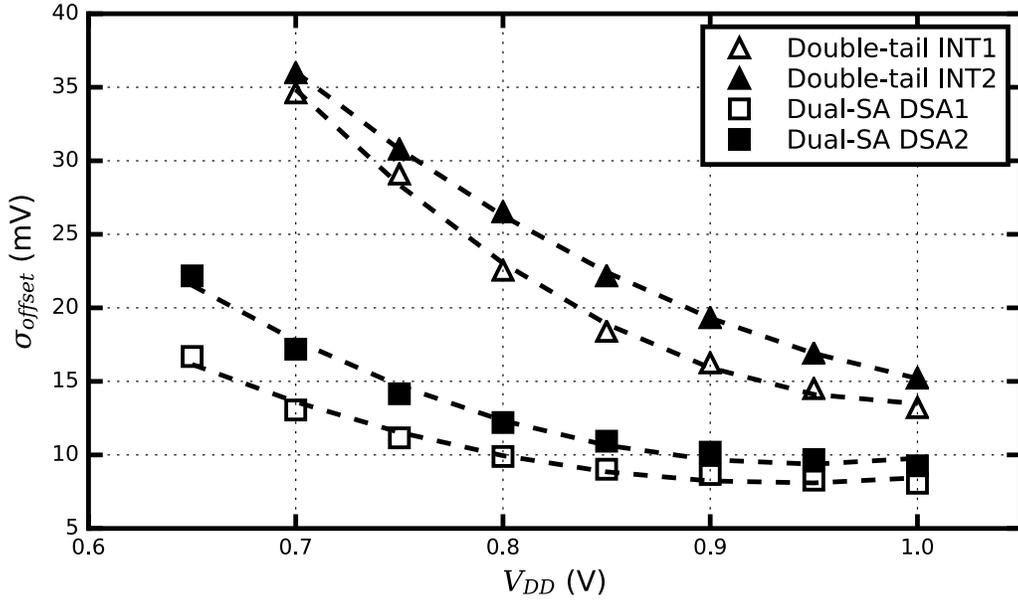


Figure 5.15: Comparison of comparator offset in all two-stage topologies.

5.3.2 Variation-driven comparator clocking scheme selection

Figure 5.15 compares the measured offsets of all two-stage comparators. We observe that a single clock-edge helps reduce offset. Continuing with our sensitivity analysis for variations, we observe in Figure 5.13 that all two clock-phase topologies exhibit much higher offset sensitivities. Like before, both INT1 and INT2 topologies lack a current-setting tail device and are by default sensitive to current variations, but the single clock-phase topology has additional pull-down devices that help the comparator evaluate faster. Once again, the dual-strong-arm is more insensitive to variations than the double-tail latch. For the two clock-phase, the output of the first stage is applied directly to nodes v_{op} and v_{on} (Figure 5.13), making the output sensitive to the devices of the cross-coupled pair in the second stage, and more specifically the PMOS devices, as it is a PMOS device that needs to start charging the output node in order for the comparator to evaluate.

Overall, when it comes to comparator design, a sensitivity analysis is necessary in order to be able to identify the optimal design. From the above analysis and keeping in mind Figure 5.14, we can conclude that current variations are the main cause of comparator offset. This fact in combination with the fact that the measured current variation within a die is generally larger than the voltage threshold variation (Figure 5.6) illustrate the importance of addressing variation in the topology selection and design stage.

Table 5.3: Comparison of simulated and measured variation for a 310nm/30nm device

	V_{TH}	I_{ON}
Simulated $3\sigma/\mu$	13%	11.1%
Measured $3\sigma/\mu$	9%	9.4%

5.4 Design-specific model customization

In the previous sections we have seen how variability affects device and circuit performance. Design yield is ultimately determined by the process variability, therefore it is critical to be able to predict it and assess its effects early at the design stage. For this reason, statistical models are incorporated to existent deterministic device performance models. However, those statistical models prove insufficient for predicting yield in larger designs.

5.4.1 Base model performance

Statistical models tend to give overly pessimistic or optimistic performance predictions, as models fail to capture the dynamic and wide-varying nature of deep-submicron processes [69, 70]. In our given technology, a wide range of within-die fluctuations has been measured, shown in Figures 5.3 and 5.5. Table 5.3 compares the measured variation in V_{TH} and I_{ON} .

Except for the deviation observed between simulation and measurement in device IV curves, we observe a significant deviation in model predictions for comparator offsets. To quantify the effect we calculate the mean absolute percent error as shown in Equation 5.2.

$$MAPE = \left| \frac{V_{off,sim} - V_{off,meas}}{V_{off,meas}} \right| \quad (5.2)$$

Figure 5.16 shows the mean absolute percent error measured for all comparator types. The bars indicate the minimum and maximum error across all supply voltages in the range of [0.7, 1.0] V, and the dots indicate the mean value.

Firstly, we observe that the error in some cases reaches almost 50%, which means that the models indeed fail to accurately predict comparator offset. Secondly, we observe that the measured error is not consistent across different topologies, which reinforces the statement that variability in deeply-scaled nodes strongly depends on the specifics of the design and layout. Our final observation is that the prediction error is also not consistent across different supply voltages, as we observe a maximum range of error from 10% to 50% for the same comparator topology. This is due to the fact that statistical models are optimized for nominal supply voltage. However, as voltage-scaling techniques for low-power electronics become more popular, designers need models that can accurately represent variability in their designed supply voltage.

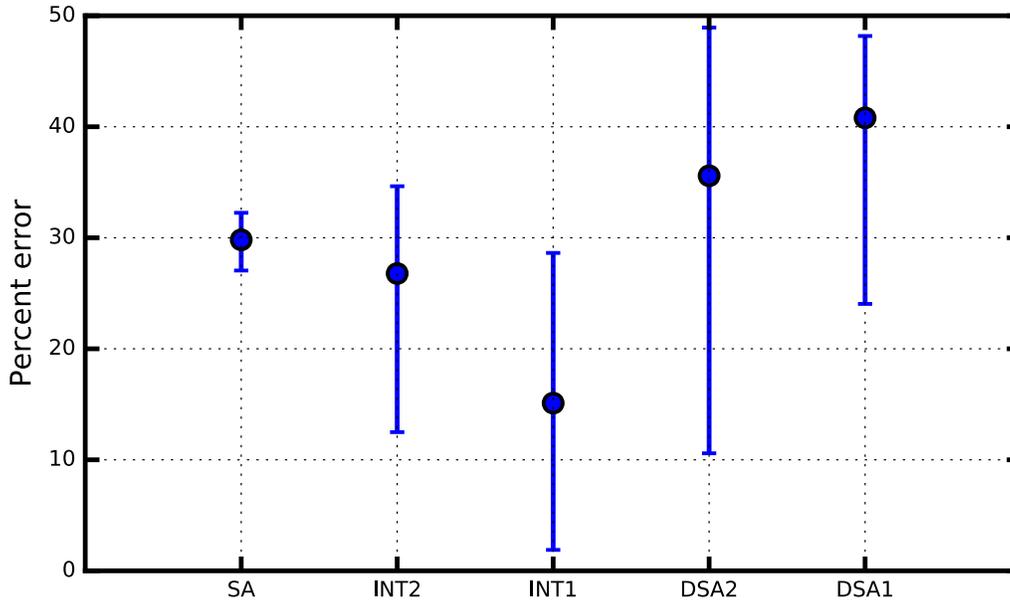


Figure 5.16: Mean absolute percent prediction error across different supply voltages for all topologies.

5.4.2 Model customization overview

Our proposed way for dealing with erroneous performance predictions of existent models is model customization. Model customization is a methodology that can be used to improve the base models using measured data from dedicated test structures. The mathematical background for model customization was presented in Chapter 3. In this section we will discuss all practical aspects of the methodology and demonstrate the effectiveness of customized models using data from the designed 28nm FDSOI testchip.

Firstly, we need to address the need for model customization, by examining the current circuit design process steps, their shortcomings and the contribution of the recommended methodology. Figure 5.17(a) shows a typical design flow with an illustration of its corresponding yield prediction error. At the circuit design stage, the yield prediction heavily depends on the given technology models. However, as it was shown in the previous section, model yield prediction is often not accurate. Models are calibrated to measurements from specific variability structures, typically voltage-current characteristics of devices or delays of ring oscillators, and as designs and design parameters deviate from those structures model statistical accuracy is lost. As a result, designers are forced to either sacrifice design performance by heavily over-designing or perform multiple expensive design iterations.

Our proposed design flow is shown in Figure 5.17(b). The goal is to improve existent device models at the first iteration, in order to achieve both high performance and high yield

in the final design. This is both possible and necessary, especially in deeply-scaled nodes, for a number of reasons. It is necessary because it eliminates multiple design iterations, therefore decreasing the overall cost of the design process. Also, the methodology addresses the issue of increased, design and layout-specific variability that is not captured by the original device models. The restricted design rules that come with deep-submicron technologies facilitate the use of post-manufacturing design centering methodologies, as now there are less layout combinations possible; therefore building test structures for characterization can be easier. This is even more pronounced as we move towards a future of pre-defined and carefully-controlled layout, in order to contain manufacturing variations. Finally, model customization provides a systematic and concise way to utilize data from the first design iteration, in order to ensure better model accuracy for the next step.

In practice, the methodology consists of two steps, before the final design is possible:

1. Test structure selection and design
2. Measurement and model customization

Test structure selection strongly depends on the desired system. For a large system, it is important to identify its most sensitive components. This can be done in the architecture selection stage (Figure 5.17(b)), where typically a high-level model of the design is developed in order to analyze the effects of different design decisions to the system. In this case, in order to validate the methodology we have selected the clocked comparator as a representative circuit as it is one of the most widely used components in mixed-signal circuits. The details of test structure design are discussed in Chapter 4.

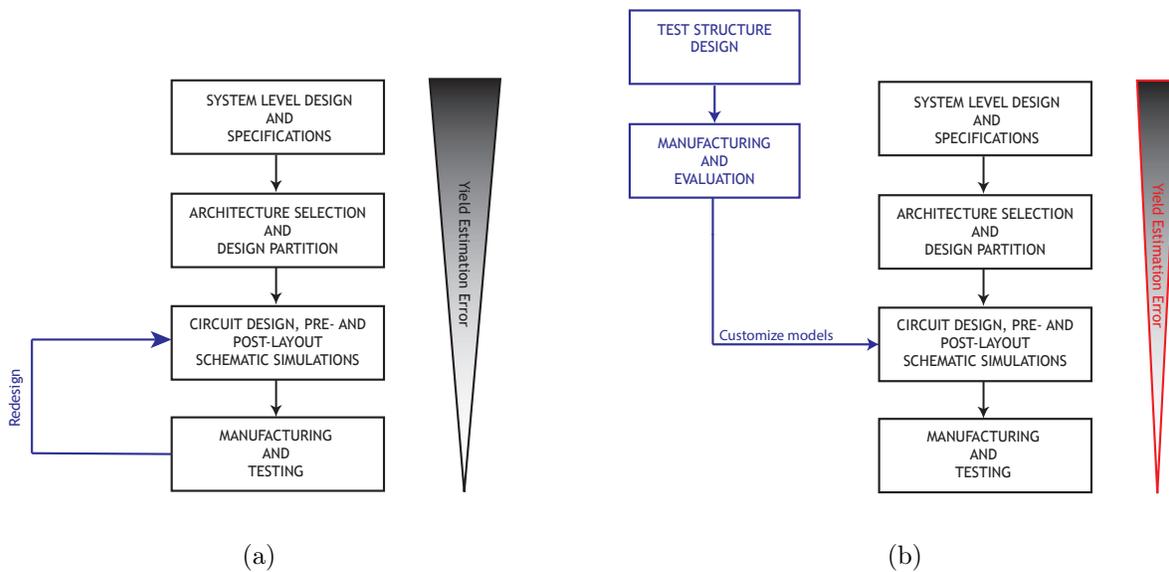


Figure 5.17: (a) Conventional design process steps and (b) proposed design process steps.

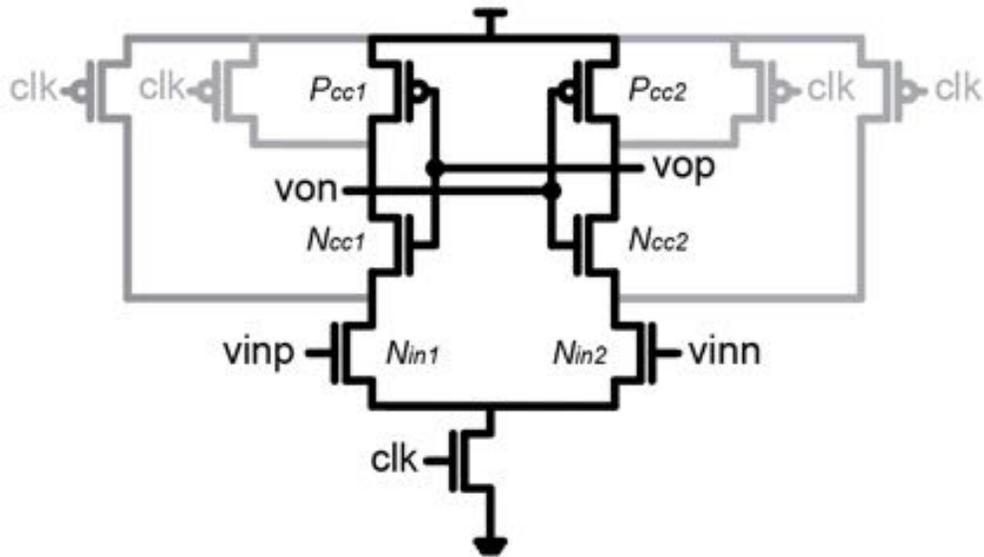


Figure 5.18: Strong-arm latch used for design centering. The greyed-out devices were not assigned statistical parameters.

Measurement results from the designed test structured are used for model customization. All different components of model customization are implemented entirely in Python. This includes netlist and library model file manipulation in order to access model parameters, sensitivity analysis and calculations, as well as setting up the system and and solution to enable the final model tuning. The software can automatically run SPICE simulations, acquire measurements and gather results to be used for customizing the models. Combining all steps of the methodology in a single piece of software not only increases speed but also adds scalability, as the software can be easily expanded to include more models types and simulators, and enables future integration with commercial design tools.

5.4.3 Customized model performance

In order to demonstrate model customization, the performance chosen was comparator offset, which was measured after noise averaging from 224 comparators per die. Statistical parameters were assigned to all devices except the pre-charge devices for simulation speedup, as they do not significantly affect offset. The excluded devices are shown in the example of a strong-arm comparator in Figure 5.18.

In order to set up the system of equations shown in Equation 3.11, the sensitivity matrix was extracted from the given simulation model using finite differences. Table 5.4 shows the extracted percent sensitivities of the output with respect to various model parameters

Table 5.4: Comparator percent offset sensitivity to statistical parameters

Instance:	N_{in1}	N_{in2}	N_{cc1}	N_{cc2}	P_{cc1}	P_{cc2}
VFBO	-1.018	1.018	-0.491	0.491	0.107	-0.107
UO	1.890	-1.890	0.810	-0.777	-0.864	0.799

assigned to each device, and therefore contains the elements of matrix \mathbf{J} . The resulting optimization problem of Equation 3.17 was then solved using commercially available convex programming tools. Parameter ρ was set to 0.5 and parameter t was selected by minimizing the root-mean-square error. An explicit constraint of $\mathbf{x} \geq 0$ was added, since the unknowns represent parameter variances.

Figure 5.19 compares the measured offset distribution of a strong-arm comparator with the distributions produced by the original and customized models, at nominal supply voltage. From the quantile-quantile plot it is evident that the models fail to accurately predict offset. The deviation at the tails of the distribution becomes even larger as the supply voltage is scaled and model accuracy is lost, as shown in Figure 5.20. Using the customized model, however, the predicted distribution matches more accurately both the body and the tails of the measured data, comparing to the original models.

The measured standard deviation of the offset across different supply and bias points is shown in Fig. 5.21, and compared to model predictions. For the customized model, design centering has been applied at each voltage step. Customized models show superior performance across the voltage range.

The presented methodology allows design centering at various supplies and biases, how-

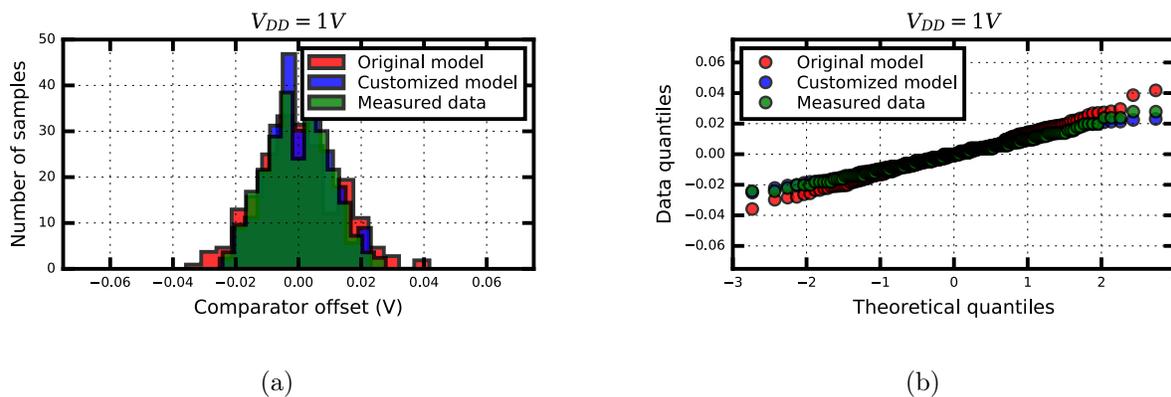


Figure 5.19: Comparison of (a) offset distribution and (b) corresponding quantile-quantile plot for the strong-arm comparator at nominal supply voltage.

ever it may be more practical for the designer to select a few supply/bias points to use. Fig. 5.22 compares the measured and simulated standard deviation of the offset across different supply voltages, but in this case the customized model was tuned using data from the nominal supply point. We observe that although at scaled supplies the offset prediction deviates from the measured data, the customized model still remains superior to the original one.

Similar results are achieved for all comparator topologies, as shown in Figure 5.23. It is important here to note that the exact same model for some topologies overestimates and for some underestimates the offset. This can be caused by layout-dependent effects, or topology-dependent variation effects that are not captured by post-layout simulation.

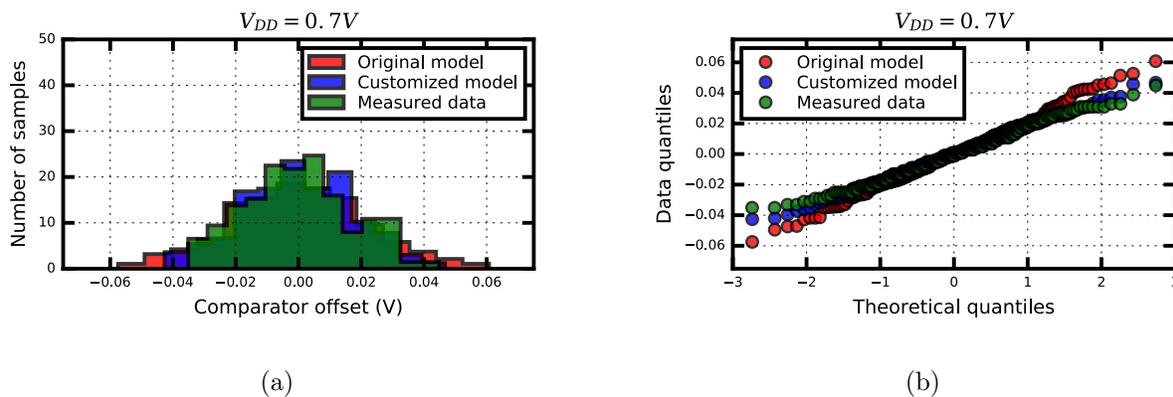


Figure 5.20: Comparison of (a) offset distribution and (b) corresponding quantile-quantile plot for the strong-arm comparator at scaled supply voltage.

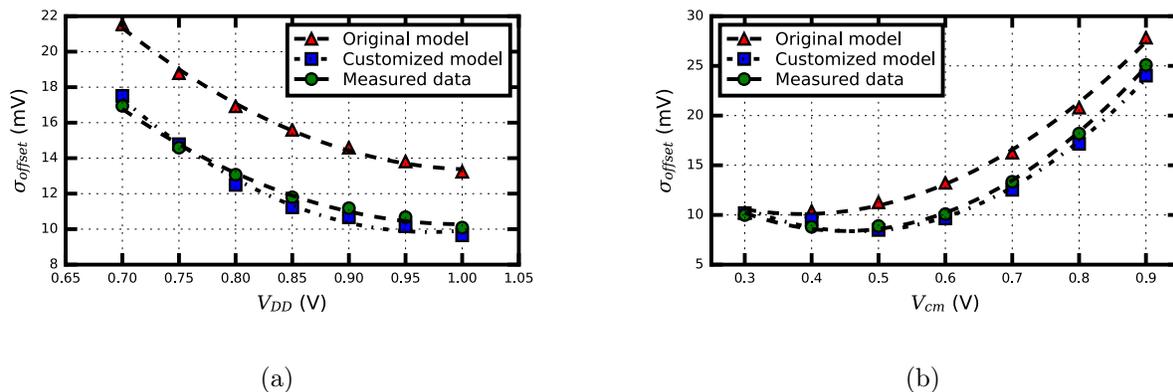


Figure 5.21: Comparison of offset standard deviation across (a) supply voltage and (b) input common-mode, when the customized model is calibrated at each voltage step.

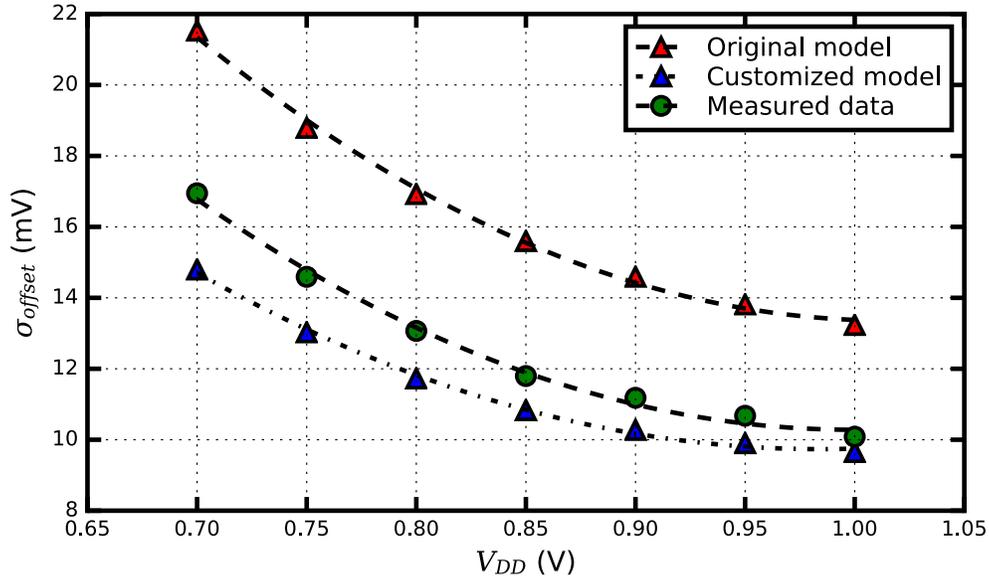
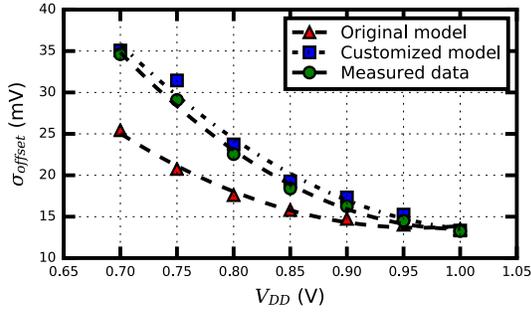


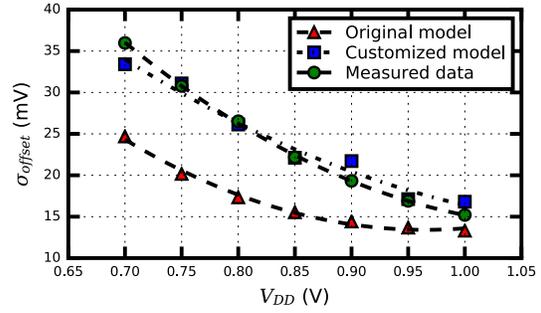
Figure 5.22: Comparison of offset standard deviation across supply voltage, when the customized model is calibrated only at nominal supply.

In order to quantify the improvement of the customized models, we need to estimate the model prediction error. Although this is not a typical model selection problem, in the sense that we are not building a model from scratch but rather modifying an existing model, the problem is formulated in such a way that a parallel with a typical polynomial curve-fitting problem can be made [71]. In our case, we use the holdout-method for cross-validation and separate the data in two sets, D1 and D2. Half of the available data is used to train the model and the rest, also called the validation set, is used to validate the model and estimate the prediction error. The process is repeated after exchanging the training and validation sets. The model with the best predictive performance is selected.

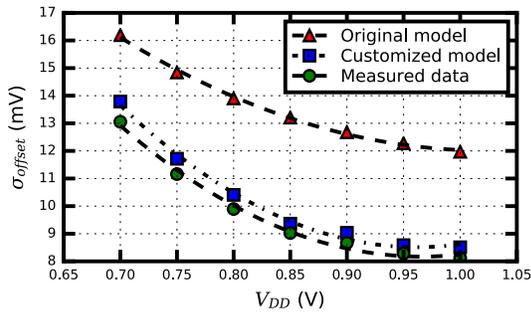
Figure 5.24 shows the estimated prediction error of the customized model for each data set. The minimum and maximum error across different supplies and topologies is shown. It is up to the designer to optimize for a specific topology and supply voltage or bias point. From the available models, the designer can select the one that has the best predictive performance for the characteristics of their specific design, provided the initial data set size is adequate. In the case where the supply of data for training and testing is limited, in order to build good models as much of the available data as possible should be used for training. However, if the validation set is small, it will give a relatively noisy estimate of predictive performance. In order to address this dilemma, higher-order cross-validation can be used. This involves partitioning the available data into S groups. Then $S-1$ groups are used to train the model and the remaining group is used to evaluate the model and estimate the prediction error. This procedure is repeated for all possible combinations.



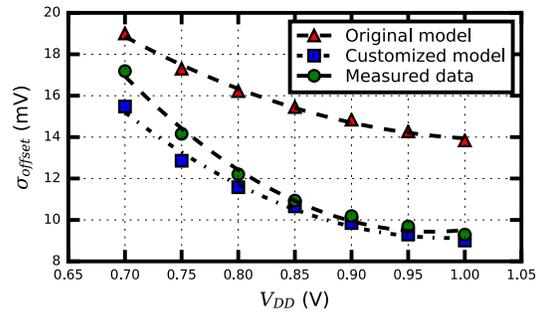
(a) INT1



(b) INT2



(c) DSA1



(d) DSA2

Figure 5.23: Comparison of offset standard deviation across supply voltage when the customized model is calibrated at each voltage step for various comparator topologies.

We will define here as the optimal model the model with the best predictive performance at nominal supply for each topology. The estimated prediction errors for each topology are summarized in Table 5.5. We observe a dramatic improvement over the original models for topologies that had a large prediction error to begin with, but there is also an improvement in cases where the original model prediction was fairly accurate.

Table 5.5: Mean absolute prediction error comparison at nominal supply voltage

	SA	INT1	INT2	DSA1	DSA2
Original model	31.2%	1.9%	12.5%	47.4%	48.9%
Customized model	4.0%	0.5%	6.8%	4.8%	0.4%

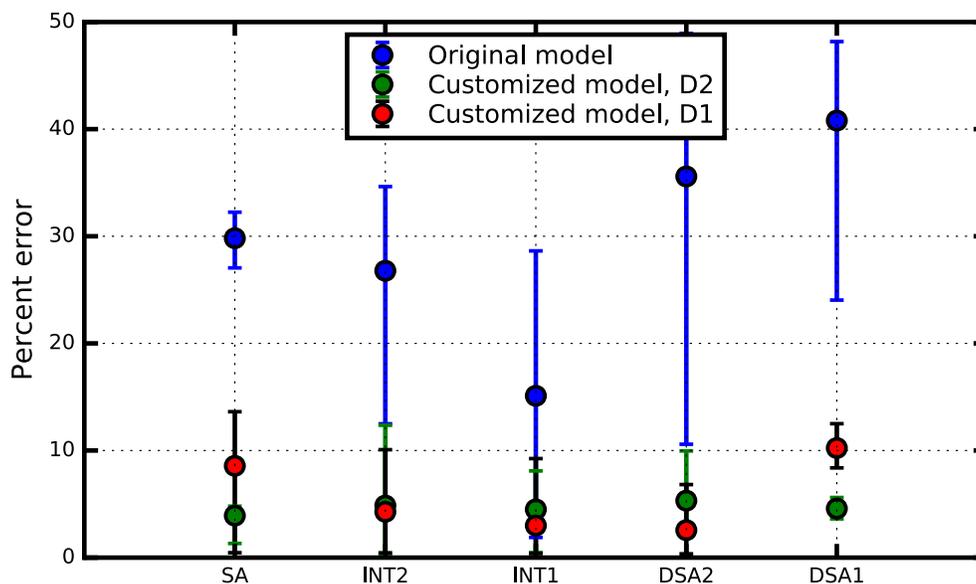


Figure 5.24: Comparison of mean absolute percent prediction error across different supply voltages for all topologies.

5.5 Summary

In this work, we addressed the problem of rising variability and insufficient variability modeling in two ways. Firstly, by characterizing variability in a deeply-scaled technology node, and secondly by demonstrating a methodology for simple, fast model tuning for design-specific yield optimization.

Technology characterization was achieved by measuring a set of dedicated test structures in a 28nm FDSOI technology. Test structures included both device characterization as well as high-speed comparator characterization, with a focus on design-dependent, layout-dependent and topology-dependent sources of variation. Worst-case within-die variation was measured to be approximately 11%, following area scaling rules. Systematic die-to-die device current variation up to 46% was measured across different dies. Layout-dependent systematic effects did not appear to be significant in this technology. Several comparator topologies were also measured, showing a direct link between comparator sensitivity and measured offset.

Yield optimization was achieved by model customization to a specific design. The methodology that was developed in Chapter 3 was implemented on measured data. The methodology was shown to have the ability to tune models to variability structure measurements, decreasing the estimated prediction error from $\sim 30\%$ to $<4\%$.

Chapter 6

Conclusions

In this chapter, we will provide a summary of the key points of this work, highlight its key contributions and expand on them by discussing future work and the future of circuit design in general.

6.1 Key contributions

Over the course of this work we have identified several problems related to variability in deeply-scaled technologies. Firstly, variability is rising, as any small perturbation from the nominal conditions has an amplified effect on device performance when the dimensions and voltages are scaled. Aggressively scaled dimensions also make the manufacturing process harder to control, further increasing variability. Secondly, the nature of variability keeps changing and will keep changing as new types of materials and devices are introduced. It is shown through measurement and simulation results that variability modeling fails to capture variability in an accurate way that will allow the designer to both push the performance limits and achieve high yield. It is also shown that the performance of statistical models depends on topology and design-specific effects, making the designer's attempt to design centering even more difficult.

The key goal of this work is to provide the groundwork for a reliable and effective solution to the aforementioned issues. This is achieved firstly by assessing technology variability and the quality of device models in silicon. Measurements of device voltage-current characteristics quantify the increase in variability and highlight a very large systematic current variation. Comparator measurements and sensitivity analysis illustrate how large device current variations translate to design-specific variation in practice. Topologies with higher sensitivities to the carrier mobility parameter are shown to have higher overall variation.

Additionally, the yield prediction capability of the device models is evaluated and found to significantly deviate from the measurements.

Finally, after recognizing and measuring the insufficiencies of existent variability modeling, we propose model customization as a promising solution. We develop a mathematical framework based on a combination of existent variability modeling techniques and deep-learning concepts. This allows for a post-manufacturing model tuning using data from specific test structures. The customized models are shown to outperform the original models.

The main advantage of this technique is its adaptability to specific designs. It is shown that design-specific variation is a significant cause of increased model prediction error, however very little research has focused on dealing with this effect. Although having reliable generic models is very important for early-stage design, a design-specific approach seems necessary bearing in mind that a measured $\sim 30\%$ deviation in individual device performance can cause anything from 2% to $\sim 50\%$ error in comparator performance, depending on the comparator topology and selected bias point. When the goal is performance-yield optimization, a design-specific approach in modeling will yield the best results.

The main disadvantage of this technique is that it requires one tape-out iteration in order to produce a customized model. This first design is based on the original models, and therefore yield cannot be guaranteed at this point. Additional resources need to be used to design the desired test-structures, which will not be directly part of the final product. Although at first glance it may seem wasteful, it is likely more cost-effective than the alternative. When working in a new technology, the first design iteration rarely meets all the required specifications, and the design has to be iterated more times. Addition of test structures to the design can help rapidly improve the models and therefore reduce the number of iterations. Therefore model customization can and should become the norm for circuit design by adding process monitoring structures to any tape-out. As a result, constant model improvement would be possible, either design-specific or not, and yield prediction would be significantly improved.

6.2 Future work

This work has laid the groundwork for design-specific device modeling, and shown its performance on a small-scale example. Further refinements are necessary before this can be widely employed as a yield-optimization methodology.

The software developed here can be refined to become faster and more flexible for a wider range of models and circuits. This can be achieved by revisiting code, improving communication channels between the software and simulator, expanding functionality and implementing it in a faster language like C/C++. An automated way for implementing S-fold cross-validation is a feature that is still missing, but would be necessary for a commercial tool. A graphical interface with commercial design tools can make the methodology much easier to use.

Regarding test structure design, it is difficult to automate the process because it is design and case-specific. However, the addition of a response surface modeling step to describe a system performance with respect to the performance of its sub-components would make it a lot easier to identify the most sensitive blocks to characterize. There is a lot of prior research in performance modeling, presented in Chapter 2, which could be incorporated to this work in order to automate the test structure selection process.

The end vision for this work would be to implement a tool that performs two tasks: identify the most-sensitive block of a design that needs to be characterized, and implement the model customization methodology based on the measured results.

6.3 Conclusions

In conclusion, a methodology has been successfully applied to customize models and improve yield prediction in a 28nm FDSOI technology. The methodology creates design-specific models in order to correct insufficiencies in the existent device models. The estimated prediction error of the customized models ranges from 0.4% to 6.8%, depending on topology; a vast improvement over the original models which have a prediction error that goes up to $\sim 50\%$ for the same design topologies.

As devices are scaling and layout rules become more restrictive, the number of possible layout and topology configurations shrinks. This makes design-specific methodologies become much more practical and more necessary. This work prepares the ground for better communication between the modeling and design sides of circuit design in order to mitigate the effects of variability, rapidly improve device models and achieve high yielding designs in a small number of design iterations.

Bibliography

- [1] Semiconductor Industry Association, “International Roadmap for Semiconductors (ITRS),” <http://www.itrs.net/>, 2009.
- [2] Intel Corporation, “Intel Developer Forum: The Evolution of a Revolution,” <http://www.intel.com/pressroom>, 2007.
- [3] B. Murmann, “ADC Performance Survey 1997-2015,” <http://web.stanford.edu/~murmman/adcsurvey.html>.
- [4] L.-T. Pang, K. Qian, C. Spanos, and B. Nikolic, “Measurement and Analysis of Variability in 45 nm Strained-Si CMOS Technology,” *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 8, pp. 2233–2243, 2009.
- [5] A. Asenov, “Simulation of Statistical Variability in Nano MOSFETs,” in *VLSI Technology, 2007 IEEE Symposium on*, June 2007, pp. 86–87.
- [6] C.-H. Lin, “Performance-aware corner model for design for manufacturing,” *Electron Devices, IEEE Transactions on*, vol. 56, no. 4, pp. 595–600, 2009.
- [7] H. Zhang, T.-H. Chen, M.-Y. Ting, and X. Li, “Efficient design-specific worst-case corner extraction for integrated circuits,” pp. 386–389, 2009.
- [8] S. R. Nassif, “Modeling and analysis of manufacturing variations,” pp. 223–228, 2001.
- [9] S. R. Nassif, “Delay variability: sources, impacts and trends,” pp. 368–369, Feb 2000.
- [10] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, “High-performance cmos variability in the 65-nm regime and beyond,” *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 433–449, July 2006.
- [11] A. Papadopoulou, “Characterization of Variability in Deeply-Scaled Fully Depleted SOI Devices,” Master’s thesis, University of California, Berkeley, 2011.
- [12] D. M. Bohling and L. A. O’Neill, “An interactive computer approach to tolerance analysis,” *IEEE Transactions on Computers*, vol. C-19, no. 1, pp. 10–16, Jan 1970.

- [13] J. Power, B. Donnellan, A. Mathewson, and W. Lane, "Relating statistical MOSFET model parameter variabilities to IC manufacturing process fluctuations enabling realistic worst case design," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 7, no. 3, pp. 306–318, Aug 1994.
- [14] R. K. Brayton, G. D. Hachtel, and A. L. Sangiovanni-Vincentelli, "A survey of optimization techniques for integrated-circuit design," *Proceedings of the IEEE*, vol. 69, no. 10, pp. 1334–1362, Oct 1981.
- [15] S. R. Nassif, A. J. Strojwas, and S. W. Director, "A methodology for worst-case analysis of integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 5, no. 1, pp. 104–113, January 1986.
- [16] Y. Deguchi, N. Ishiura, and S. Yajima, "Probabilistic ctss: analysis of timing error probability in asynchronous logic circuits," in *Design Automation Conference, 1991. 28th ACM/IEEE*, June 1991, pp. 650–655.
- [17] S. Devadas, H. F. Jyu, K. Keutzer, and S. Malik, "Statistical timing analysis of combinational circuits," in *Computer Design: VLSI in Computers and Processors, 1992. ICCD '92. Proceedings, IEEE 1992 International Conference on*, Oct 1992, pp. 38–43.
- [18] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Computer Aided Design, 2003. ICCAD-2003. International Conference on*, Nov 2003, pp. 900–907.
- [19] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical timing analysis: From basic principles to state of the art," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 4, pp. 589–607, April 2008.
- [20] G. G. E. Gielen and R. A. Rutenbar, "Computer-aided design of analog and mixed-signal integrated circuits," *Proceedings of the IEEE*, vol. 88, no. 12, pp. 1825–1854, Dec 2000.
- [21] Z. Wang and S. W. Director, "An efficient yield optimization method using a two step linear approximation of circuit performance," in *European Design and Test Conference, 1994. EDAC, The European Conference on Design Automation. ETC European Test Conference. EUROASIC, The European Event in ASIC Design, Proceedings.*, Feb 1994, pp. 567–571.
- [22] A. Dharchoudhury and S. M. Kang, "Worst-case analysis and optimization of vlsi circuit performances," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 4, pp. 481–492, Apr 1995.
- [23] W. Daems, G. Gielen, and W. Sansen, "Simulation-based generation of posynomial performance models for the sizing of analog integrated circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 5, pp. 517–534, May 2003.

- [24] X. Li, P. Gopalakrishnan, Y. Xu, and L. T. Pileggi, "Robust analog/RF circuit design with projection-based performance modeling," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 1, pp. 2–15, Jan 2007.
- [25] T. Kiely and G. Gielen, "Performance modeling of analog integrated circuits using least-squares support vector machines," in *Design, Automation and Test in Europe Conference and Exhibition, 2004. Proceedings*, vol. 1, Feb 2004, pp. 448–453 Vol.1.
- [26] H. Liu, A. Singhee, R. A. Rutenbar, and L. R. Carley, "Remembrance of circuits past: macromodeling by data mining in large analog design spaces," in *Design Automation Conference, 2002. Proceedings. 39th*, 2002, pp. 437–442.
- [27] T. McConaghy, T. Eeckelaert, and G. Gielen, "Caffeine: template-free symbolic model generation of analog circuits via canonical form functions and genetic programming," in *Design, Automation and Test in Europe, 2005. Proceedings*, vol. 2, March 2005, pp. 1082–1087.
- [28] T. McConaghy and G. G. E. Gielen, "Template-free symbolic performance modeling of analog circuits via canonical-form functions and genetic programming," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 8, pp. 1162–1175, Aug 2009.
- [29] K. Qian and C. Spanos, "A comprehensive model of process variability for statistical timing optimization," in *Proc. SPIE, Design for Manufacturability through Design-Process Integration III*, vol. 6925, 2008.
- [30] B. N. K. Qian and C. Spanos, "Hierarchical modeling of spatial variability with a 45nm example," in *Proc. SPIE, Design for Manufacturability through Design-Process Integration III*, vol. 7275, 2009.
- [31] K. Q. Y. Qiao and C. Spanos, "Variability aware compact model characterization for statistical circuit design optimization," in *Proc. SPIE, Design for Manufacturability through Design-Process Integration III*, vol. 8327, 2012.
- [32] Y. Ben and C. J. Spanos, "Estimating the probability density function of critical path delay via partial least squares dimension reduction," in *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, March 2011, pp. 1–7.
- [33] C. C. McAndrew, "Statistical modeling for circuit simulation," in *Quality Electronic Design, 2003. Proceedings. Fourth International Symposium on*, March 2003, pp. 357–362.
- [34] X. Li, C. C. McAndrew, W. Wu, S. Chaudhry, J. Victory, and G. Gildenblat, "Statistical modeling with the psp mosfet model," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 4, pp. 599–606, April 2010.
- [35] C. H. Lin, M. V. Dunga, D. D. Lu, A. M. Niknejad, and C. Hu, "Performance-aware corner model for design for manufacturing," *IEEE Transactions on Electron Devices*, vol. 56, no. 4, pp. 595–600, April 2009.

- [36] D. Divekar, *FET Modeling for Circuit Simulation*. Springer US, 1988.
- [37] G. A. F. Seber, *Multivariate Observations*. John Wiley and Sons, 1984.
- [38] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics*, vol. 4, pp. 1035–1038, 1963.
- [39] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
- [40] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.
- [41] X. Li, "Finding deterministic solution from underdetermined equation: Large-scale performance modeling by least angle regression," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, July 2009, pp. 364–369.
- [42] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.
- [43] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," 2013.
- [44] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar)," in "*Lecture Notes in Control and Information Sciences*", Blondel, V., Boyd, S. and Kimura, H., Ed. Springer, 2008, pp. 95–110.
- [45] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: SRAM evaluation through norm minimization," in *IEEE/ACM International Conference on Computer-Aided Design, 2008. ICCAD 2008*, Nov. 2008, pp. 322–329.
- [46] M. Qazi, M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, "Loop flattening and spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, Mar. 2010, pp. 801–806.
- [47] T. Kobayashi, K. Nogami, T. Shirotori, Y. Fujimoto, and O. Watanabe, "A current-mode latch sense amplifier and a static power saving input buffer for low-power architecture," in *, 1992 Symposium on VLSI Circuits, 1992. Digest of Technical Papers*, Jun. 1992, pp. 28–29.
- [48] J. Montanaro, R. T. Witek, K. Anne, A. J. Black, E. M. Cooper, D. W. Dobberpuhl, P. M. Donahue, J. Eno, W. Hoepfner, D. Kruckemyer, T. H. Lee, P. C. M. Lin, L. Madden, D. Murray, M. H. Pearce, S. Santhanam, K. J. Snyder, R. Stehpany, and S. C. Thierauf, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1703–1714, Nov. 1996.

- [49] H. Johansson and C. Svensson, "Time resolution of NMOS sampling switches used on low-swing signals," *Solid-State Circuits, IEEE Journal of*, vol. 33, no. 658625, pp. 237–245, 1998.
- [50] J. Kim, B. Leibowitz, and M. Jeeradit, "Impulse sensitivity function analysis of periodic circuits," *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, no. 4681602, pp. 386–391, 2008.
- [51] A. Hajimiri, S. Limotyrakis, and T. Lee, "Jitter and phase noise in ring oscillators," *Solid-State Circuits, IEEE Journal of*, vol. 34, pp. 790–804, 1999.
- [52] M. Jeeradit, J. Kim, B. Leibowitz, P. Nikaeen, V. Wang, B. Garlepp, and C. Werner, "Characterizing sampling aperture of clocked comparators," *VLSI Circuits, 2008 IEEE Symposium on*, no. 4585955, pp. 68–69, 2008.
- [53] S. Ohkawa, M. Aoki, and H. Masuda, "Analysis and characterization of device variations in an LSI chip using an integrated device matrix array," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 2, pp. 155–165, May 2004.
- [54] D. Schinkel, E. Mensink, E. Klumperink, E. v. Tuijl, and B. Nauta, "A Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup+Hold Time," in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, Feb. 2007, pp. 314–605.
- [55] M. Miyahara, Y. Asada, D. Paik, and A. Matsuzawa, "A low-noise self-calibrating dynamic comparator for high-speed ADCs," in *Solid-State Circuits Conference, 2008. ASSCC '08. IEEE Asian*, Nov. 2008, pp. 269–272.
- [56] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.
- [57] L.-T. Pang and B. Nikolic, "Measurements and Analysis of Process Variability in 90 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 4907316, pp. 1655–1663, 2009.
- [58] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, "Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 131–144, Jan. 2008.
- [59] R. A. Bianchi, G. Bouche, and O. Roux-dit Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," in *Electron Devices Meeting, 2002. IEDM '02. International*, Dec. 2002, pp. 117–120.
- [60] N. Damrongplisit, "Study of Variability in Advanced Transistor Technologies," Ph.D. dissertation, University of California, Berkeley, 2014.

- [61] A. Ortiz-Conde, F. G. Sanchez, J. Liou, A. Cerdeira, M. Estrada, and Y. Yue, "A review of recent {MOSFET} threshold voltage extraction methods," *Microelectronics Reliability*, vol. 42, pp. 583 – 596, 2002. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0026271402000276>
- [62] O. Weber, F. Andrieu, J. Mazurier, M. CassÃr, X. Garros, C. Leroux, F. Martin, P. Perreau, C. Fenouillet-BÃranger, S. Barnola, R. Gassilloud, C. Arvet, O. Thomas, J. P. Noel, O. Rozeau, M. A. Jaud, T. Poiroux, D. Lafond, A. Toffoli, F. Allain, C. Tabone, L. Tosti, L. BrÃrard, P. Lehnen, U. Weber, P. K. Baumann, O. Boissiere, W. Schwarzenbach, K. Bourdelle, B. Y. Nguyen, F. BÃsuf, T. Skotnicki, and O. Faynot, "Work-function engineering in gate first technology for multi-VT dual-gate FDSOI CMOS on UTBOX," in *2010 International Electron Devices Meeting*, Dec. 2010, pp. 3.4.1–3.4.4.
- [63] N. Planes, O. Weber, V. Barral, S. Haendler, D. Noblet, D. Croain, M. Bocat, P. O. Sassoulas, X. Federspiel, A. Cros, A. Bajolet, E. Richard, B. Dumont, P. Perreau, D. Petit, D. Golanski, C. Fenouillet-Beranger, N. Guillot, M. Rafik, V. Huard, S. Puget, X. Montagner, M. A. Jaud, O. Rozeau, O. Saxod, F. Wacquand, F. Monsieur, D. Barge, L. Pinzelli, M. Mellier, F. Boeuf, F. Arnaud, and M. Haond, "28nm FDSOI technology platform for high-speed low-voltage digital applications," in *2012 Symposium on VLSI Technology (VLSIT)*, Jun. 2012, pp. 133–134.
- [64] J. Mazurier, O. Weber, F. Andrieu, C. L. Royer, O. Faynot, and M. Vinet, "Variability of planar Ultra-Thin Body and Buried oxide (UTBB) FDSOI MOSFETs," in *2014 IEEE International Conference on IC Design Technology*, May 2014, pp. 1–4.
- [65] A. Misaka, A. Goda, K. Matsuoka, H. Umimoto, and S. Odanaka, "A statistical critical dimension control at CMOS cell level," Dec. 1996, pp. 631–634.
- [66] P. Nuzzo, F. D. Bernardinis, P. Terreni, and G. V. d. Plas, "Noise Analysis of Regenerative Comparators for Reconfigurable ADC Architectures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1441–1454, Jul. 2008.
- [67] B. Razavi, "The StrongARM Latch [A Circuit for All Seasons]," *IEEE Solid-State Circuits Magazine*, vol. 7, no. 2, pp. 12–17, 2015.
- [68] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, Jul. 2004.
- [69] J. A. Power, A. Mathewson, and W. A. Lane, "MOSFET statistical parameter extraction using multivariate statistics," in *Proceedings of the 1991 International Conference on Microelectronic Test Structures*, Mar. 1990, pp. 209–214.
- [70] C. H. Lin, M. V. Dunga, D. Lu, A. M. Niknejad, and C. Hu, "Statistical Compact Modeling of Variations in Nano MOSFETs," in *2008 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, Apr. 2008, pp. 165–166.
- [71] C. Bishop, *Information Science and Statistics*, Oct. 2006.