Response Surface Modeling of the SET Pulse Width Distribution with Operation Parameters and Process Variations

By

Chundong Liang

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

Master of Science

in

Electrical Engineering

August, 2014

Nashville, Tennessee

Approved by:

Professor Ronald D. Schrimpf

Professor Michael Alles

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Ronald Schrimpf from the bottom of my heart, for the support and guidance provided while completing my master's degree, for his wisdom and knowledge to inspire me to narrow down the goal for the project, for his patience and encouragement when results were difficult to come by. Whenever I meet with problems in this work, I would obtain guidance and advice from him immediately. Without this tremendous help, it is impossible for me to finish this work. I still remember his analogy of "dice theory" helps me a lot in defining the goal when I have trouble understanding the project in the preliminary stage.

Second, I would like to thank Dr. Michael Alles for being my second advisor. Dr. Alles is always considering my work and spends a lot of effort to help me with concepts and refresh me with ideas. He would sit down with me and help me analyze my presentation problems slide by slide. His novel perspective for mathematical modeling of the project inspires me to make great progress on statistical modeling methods and enriched my knowledge on EDA algorithms. Also, I would like to thank Dr. Dan Fleetwood, Dr. Lloyd Massengill, Dr. William Robinson and other professors for their feedback in group meetings. Especially, I appreciate Nihaar Mahatme's help for discussion on the topic.

The continued encouragement from my parents and my girlfriend Yanran Chen are also appreciated. The support from them provides me with the courage to pursue this goal. I am grateful for their wisdom and love. I will always remember Yanran's insightful perspectives to help me face various challenges in my life.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

I.1 Motivation

Microelectronic devices and integrated circuits (ICs) are exposed to a wide range of radiation environments in space applications. Single energetic particles in space may induce transients in combinational circuits and the resulting pulses could lead to soft errors in digital circuits. These undesired pulses are a threat to the correct functionality of ICs working in radiation environments. To determine the soft-error rate, the single-event transient (SET) pulse width is a key variable.

As technologies move into nano-scale dimensions, various complex device phenomena play important roles in digital circuit functionality and reliability. Additionally, variations in the circuit parameters such as process, voltage, and temperature (PVT) significantly affect the functionality of the devices and cause variability of the SET pulse widths in digital circuits [1]. Due to these variations of these circuit parameters, operating conditions and radiation parameters, SET-induced pulse widths are stochastic rather than deterministic in nano-scale digital circuits.

In order to quantify the SET pulse width distribution, a stochastic model is needed based on the underlying physical and electrical parameters. A standard method to achieve this goal is based on the Monte Carlo method using transistor level simulation. However, the number of parameters increases dramatically as semiconductor devices shrink, and the Monte Carlo method is time-consuming since it samples a large input parameter space. As a result, a fast model to

statistically evaluate the SET pulse width distribution is needed. The major objectives of this work are:

• To provide a framework/methodology for statistical analysis using Response Surface Modeling (RSM) and Design of Experiments (DOE) method to model the effects of process variations, design parameters, and environmental parameters on SET pulse width.

• To provide a fast modeling method to estimate the SET pulse width distribution under randomized input parameters.

I.2 Overview of the Thesis

The organization of the thesis is as follows:

The mechanisms related to SET pulse width are presented in Chapter 2. The variations in the SET pulse widths of standard cells due to PVT variations, design parameters, and environmental parameters are analyzed in Chapter 3. The application of design of experiment (DOE) methods and response surface modeling (RSM) to construct a predictive model for SET pulse width is discussed in Chapter 4. Chapter 5 covers estimating the distributions of SET pulse widths by the proposed model, and the result is compared with a standard Monte Carlo method. At the end, conclusions and possible future work are reported in Chapter 6.

CHAPTER II

SINGLE EVENT TRANSIENT MECHANISM

II.1 Single-event effect (SEE)

Single-event effects (SEE) occur when energetic particles (e.g., protons, neutrons, alpha particles, heavy ions) strike semiconductor devices and produce electrical effects. The sources of SEEs include cosmic rays from space, alpha particles in terrestrial environments, and neutrons that cause SEEs indirectly by secondary particles from nuclear interactions [2]. When a semiconductor device is struck by an energetic particle, the energy transferred from the particle may promote electrons to the conduction band and leave holes in the valence band. In this process, linear energy transfer (LET) describes the amount of energy transferred per unit of path length by the particles as normalized by the density of the material and has units of MeV-cm^2/mg. An ion with an LET of 97 MeV-cm^2/mg leads to charge deposition of approximately 1 pC/μm.

When excess charge is generated by energetic particles striking semiconductor material, all the mobile excess charge has to be either collected by the nodes of the circuit or recombine. Four primary mechanisms affect the charge collection process in bulk CMOS circuits:

1. Drift: Movement of holes and electrons in the presence of electric field. The electric field causes the holes to be swept into the p-region and the electrons into the n-region. Drift transport is a quick process and the duration is on the order of picoseconds.
2. Diffusion: Movement of holes and electrons induced by a charge concentration gradient. If the mobile charges exist outside the depletion region of a junction, they could diffuse to the depletion region before they recombine. The diffusion process is very slow compared to the

drift process. An illustration of drift and diffusion mechanisms in bulk transistor is shown in Fig. 1.

3. Bipolar-amplification: Charge movement can induce a potential drop in the n-well containing a PMOS device, which in turn forward biases the source-body junction of the parasitic bipolar structure. Additional charge will be injected from the source into the body. As a result, compared to a PN diode without bipolar-amplification, this mechanism in pMOSFETs adds amplified current to the original particle-induced current.
4. Recombination: The process by which electrons and holes annihilate one another if they are not collected.

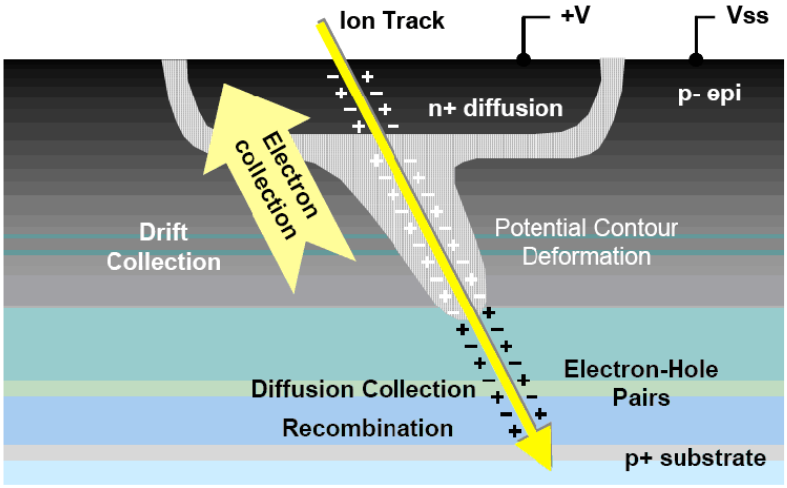

Fig. 1. Illustration of drift and diffusion mechanisms in a bulk transistor**[3]**

The four mechanisms in bulk CMOS devices dominate the charge collection process when the ion-induced charges are present. The ion-induced current transient occurs at the device terminals, and its shape depends on the corresponding charge collection and recombination mechanisms. The initial peak of the pulse shape results from the quick drift collection of the charges, and the tail of the curve is caused by the slow collection of the charge by diffusion, as shown in Fig. 2.

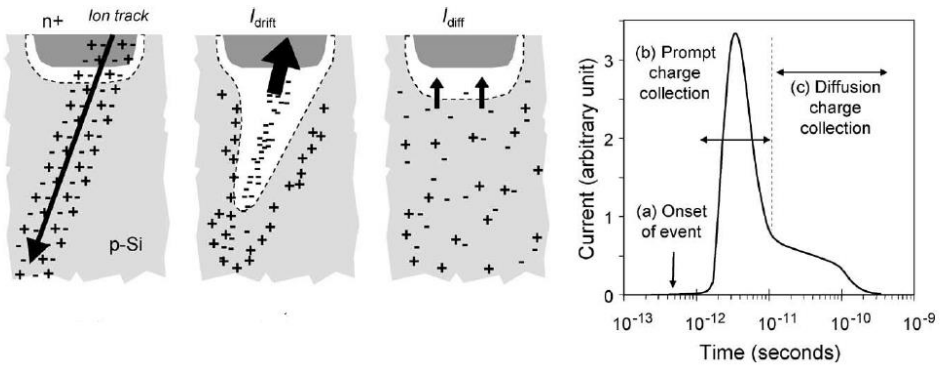

Fig. 2 Charge generation and collection phases in a reverse-biased junction and the resultant current pulse caused by the passage of a high-energy ion**[2]**

II.2 Single event Transient (SET) and pulse width

The current transient may flow through the circuit, inducing a voltage glitch at the struck node. This voltage perturbation caused by an energetic particle strike is defined as a Single Event Transient (SET). The voltage transient may cause malfunction in either combinational logic or storage cells.

Typical storage cells include latches and flip-flops. In this type of circuit the error rate due to single events is almost independent of the clock frequency of the circuit. The latch or flip-flop's state can be changed by an ionizing particle creating charge on a node regardless of the state of the clock signal at its input; on the other hand, for combinational logic circuits such as NAND gates, XOR gates, and inverters, SETs that are induced between storage cells can arrive at the input of the storage cell on the latching edge of the clock and be clocked in as erroneous data, as shown in Fig. 3. Typically, SET pulse width is defined as the time interval between the points at which the voltage is $V_{dd}/2$. If the single event-induced pulse does not reach $V_{dd}/2$, the pulse

width is treated as 0 in this work, as we assume it will not impact the state of the following logic gate.

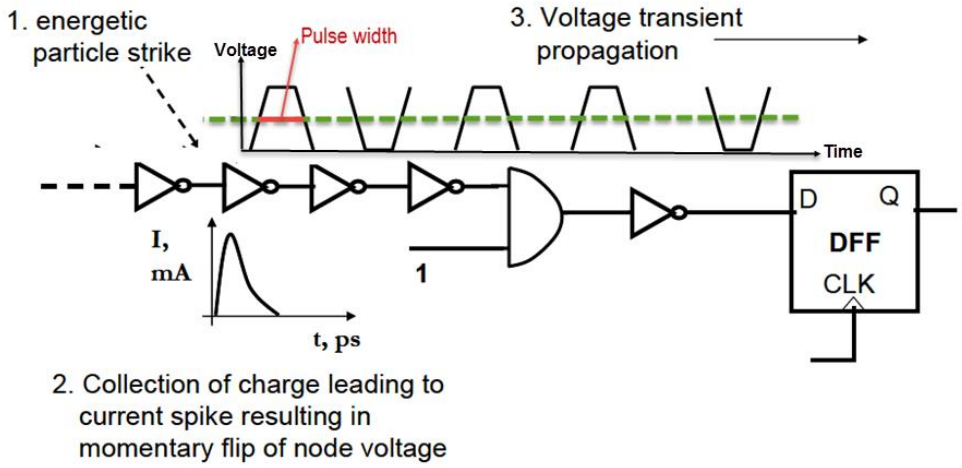

Fig. 3 Generation, propagation and latched of the SET pulse width

II.3 The influence of the SET pulse width to the logic circuits

In the case discussed in the previous section, the probability for a transient pulse to get latched as incorrect data depends directly on pulse width and clock frequency [4], [5]. The faster the clock, the more latching clock edges there are available to capture a transient signal; the wider the pulse width, the greater probability there is to capture the pulse width in the latch window. If the transient pulse becomes longer than the time period of the clock, then every induced transient pulse will be latched. Fig. 4 illustrates how the width of an SET determines the probability of whether or not the SET will be latched. In this figure, the data will latch on the clock's falling edge.

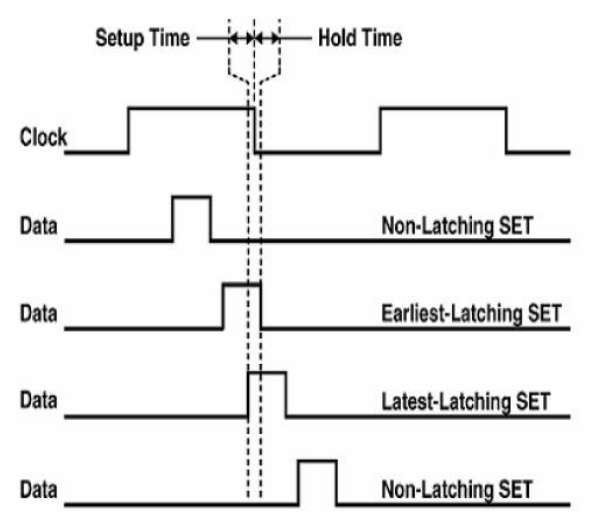

Fig. 4 SETs arriving at the latching edge of a clock can be recorded as incorrect bits**[6]**

From Fig. 4, one can see how a wider SET pulse will lead to a greater probability of the SET arriving on the latching edge of a clock signal. Whether or not SETs are latched depends on various factors, including electrical masking, latch window masking, and logical masking [5], [6]. Both latch window masking and logic masking are SET pulse width dependent.

As stated above, the SET pulse width distribution characterization in logic gates is important to predict the soft-error rate and evaluate the radiation sensitivity of digital circuits. Soft errors induced by SETs in combinational circuits are a strong function of SET pulse width [2],[7], and mechanisms (i.e., drift, diffusion, recombination, and bipolar effect) controlling SET pulse width are strong functions of circuit design parameters, operation parameters and device parameters. In the next chapter, we will discuss the impact of multiple parameters on the SET pulse width in standard cells.

CHAPTER III

SINGLE PARAMETER EFFECTS ON THE SET PULSE WIDTH

III.1 Simulation configuration

The impact of circuit parameters on SET pulse width in standard cells can be evaluated by circuit simulation. Circuit simulators are tools to estimate the expected behavior of a physical circuit. Mathematical models of each circuit component are required to perform a circuit simulation. In this section, the device models and a current source model to simulate SET pulse width in standard cells are discussed.

III.1.1 Predictive Technology Models (PTM) and 45nm bulk PDK

Due to device shrinking, various physical effects (like short channel effect, gate leakage, etc.) play significant roles in modern device behavior. The number of parameters of a MOSFET model is increasing significantly to represent the complex behavior of modern MOSFETs. BSIM (Berkeley Short-channel IGFET Model)[8] from the University of California Berkeley is a well-known industry standard model. BSIM4[9], a variant of the BSIM model, is widely used in state-of-the-art integrated circuit simulations. The parameters of these models are extracted by characterization of the MOSFET and they are fab and technology dependent. The ITRS (International Technology Roadmap for Semiconductor) is actively involved in defining the future technology nodes. MOSFET models are also required for future technology nodes for use in research activities.

The PTM (Predictive Technology Model) is a well-known technology model for transistors as specified by ITRS. It is an accurate, customizable, and predictive model for transistors and interconnect technologies provided by the nano-scale Integration and Modeling (NIMO) Group at ASU [10]. The models are compatible with various standard circuit simulators, and scalable with a wide range of process variations. With PTM, circuit design and performance evaluation can be started even before the advanced semiconductor technology is fully developed. PTM provides a list of features for research toward the 7-nm regime.

- Predictions of various transistor structures, such as sub-45nm bulk technology nodes used in this work.
- New methodology of prediction, which is more physical, scalable, and continuous over technology generations.
- Predictive models for emerging variability and reliability issues, such as NBTI.

The source code to generate PTM models is available from [10]. The transistor models used in this work correspond to the 45-nm technology node. Additionally, FreePDK45 design contains technical files, design rules, display resources and scripts to permit design and rule checking for a generic 45-nm process. Schematic creation and Spectre simulation for these models are supported by Cadence Virtuoso and the Cadence Analog Design Environment (ADE). In this work, MATLAB is also used for data processing.

Since the input model files for Spectre are not the same as for SPICE, appropriate modifications have been made in this work on the PTM model such that it can be supported by Cadence Spectre and the Cadence Analog Design Environment (ADE).

III.1.2 Bias current source model

In this work, single-event current was injected using the bias dependent compact model published by Kauppila et al. [11], which helps eliminate the possibility of injecting unphysical amounts of current into the circuit. This model is a modified current source that responds dynamically to circuit conditions. The current source checks the bias condition on the connected node, and adjusts the amount of injected current accordingly. The configuration parameters of the current source model are calculated to fit the SET current pulse shape obtained using TCAD [11].

III.2 The impact of single parameter on SET pulse width

In order to determine the parameters that may significantly impact the SET pulse width in logic gates, a broad literature review is conducted to identify relevant parameters. From [12]-[13], PVT (process, voltage and temperature) variations, design parameters (width/length ratio, loading capacitance), and environmental parameters (LET) affect SET pulse width. However, the conclusions in the literature are sometimes conflicting. For example, a parameter that is identified to be important for pulse width may not be considered statistically significant if other parameters are included. In the real world, to evaluate the pulse width distribution of a particular type of logic gate, multiple parameters should be considered simultaneously. In this work, a series of parameters that may affect SET pulse width is investigated, statistically modeled and discussed, and this serves as the basis for modeling the pulse width distribution in the following chapters. In this example, a NAND2 with input “10” is selected for illustration. The sensitive device is the cut-off nMOSFET, as shown in Fig. 5.

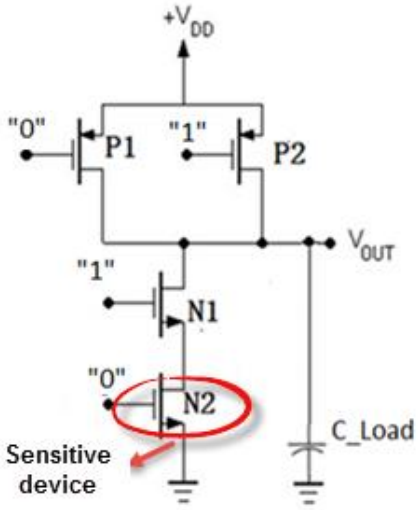

Fig. 5 Simulation setup of NAND2

Other input conditions and other standard logic cells could be easily evaluated in a similar way. The parameter variations used in the detailed discussion are listed in Table 1.

Table 1 parameter variations

Parameters	Nominal Value	Variations	Distribution
Leff (Effective channel length)	22.5nm	$\sigma = \pm 6\%$	Gaussian
V_{thn} (Threshold voltage of NMOS)	0.466V	$\sigma = \pm 14\%$	Gaussian
V_{thp} (Threshold voltage of PMOS)	-0.411V	$\sigma = \pm 14\%$	Gaussian
LET	8MeV/mg/cm^2	6-10MeV/mg/cm^2	Uniform
V_{dd} (Power supply)	1.0V	0.8-1.2V	Uniform
W/L	2 X minimum	1X – 3X minimum	Uniform

Cap (Load capacitance)	2 X minimum inv	1X – 3X minimum inv	Uniform
Temp (Temperature)	62.5 Celsius	25-100 Celsius	Uniform

III.2.1 Process variations

The semiconductor manufacturing process induces variations in the physical parameters, resulting in electrical variations of the CMOS devices. These variations are statistical in nature, and they can be categorized as either intra-die or inter-die [14], as shown in Fig. 6. The inter-die process variations are the variations across the wafers or wafer-lots and the intra-die variations are the variations within the die. The physical and electrical parameters of MOSFETs vary due to the fabrication process and the device wear-out mechanisms.

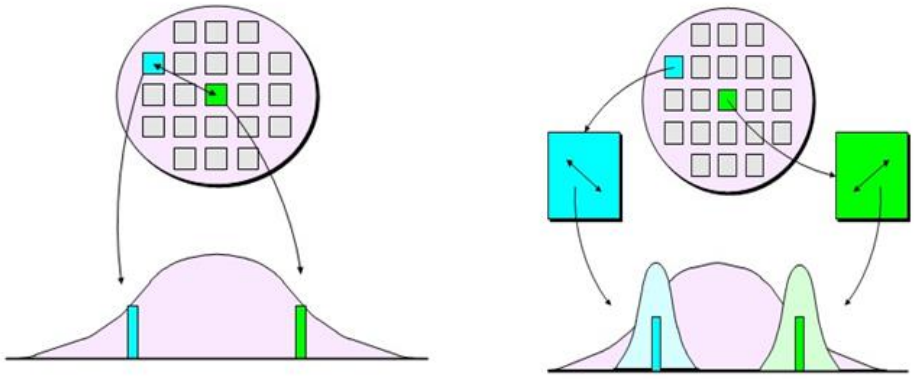

Fig. 6 Inter-die(left) and Intra-die(right) process variances [14]

These variations affect the yield of a CMOS process and it is important to properly estimate the amount of variation for each parameter. A typical approach for modeling variations begins with I–V measurements obtained either from a wafer prober or an on-chip measurement system [15]. Then the measured data are used to extrapolate process parameter distributions. It is widely

accepted that the process variations generally follow a Gaussian distribution. This Gaussian distribution assumption is used in this work to investigate the impact of process variations on the performance of the circuits.

With shrinking feature sizes, the variations are becoming significant and this affects the behavior of logic gates. According to [16] ,the variation in threshold voltage and channel length for a 45-nm CMOS process affects the soft-error response, as shown in Fig. 7. Thus, it is necessary to include process variables in evaluating the distribution of SET pulse widths.

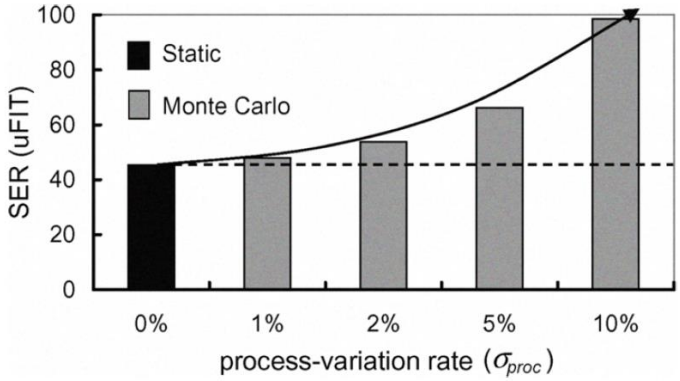

Fig. 7 Dependence of SER on process-variations [16]

In this work, channel length and threshold voltage of both nMOSFETs and pMOSFETs are selected for illustration of the modeling work. The number of process parameters considered could easily be expanded if necessary.

III.2.1.1 Channel length

Channel length is a process variable and the Gaussian distribution is often assigned to this random variable. Theoretically, while channel length is a factor affecting both nodal capacitance and drain current, the effect on drain current is the dominant effect. According to equation (1),

$$I \propto C_{ox}\mu\left(\frac{width}{length}\right) \qquad (1)$$

Increased effective channel length decreases the current through the PMOS transistors that provide current to reinforce the perturbed output voltage, which leads to an increase of the SET pulse width.

In Spectre simulations, the effective channel length is determined by adding a variable, XL, to the drawn length in the PTM models. The variable varies between the nominal value ±3σ, which equals 18% of the nominal value [17]. In order to investigate the relationship of the SET pulse width to the effective channel length, a variable XL is modified in the PTM model file so that the effective channel length values are selected at multiple values within ±3σ of the nominal value. All the other parameters are at their nominal values. Then the SET pulse widths are obtained by Spectre simulation and PTM models with the selected XL. The simulation results of the SET pulse width vs. effective channel length are shown in Fig. 7, which confirms that the SET pulse width is positively correlated to the channel length, as discussed before.

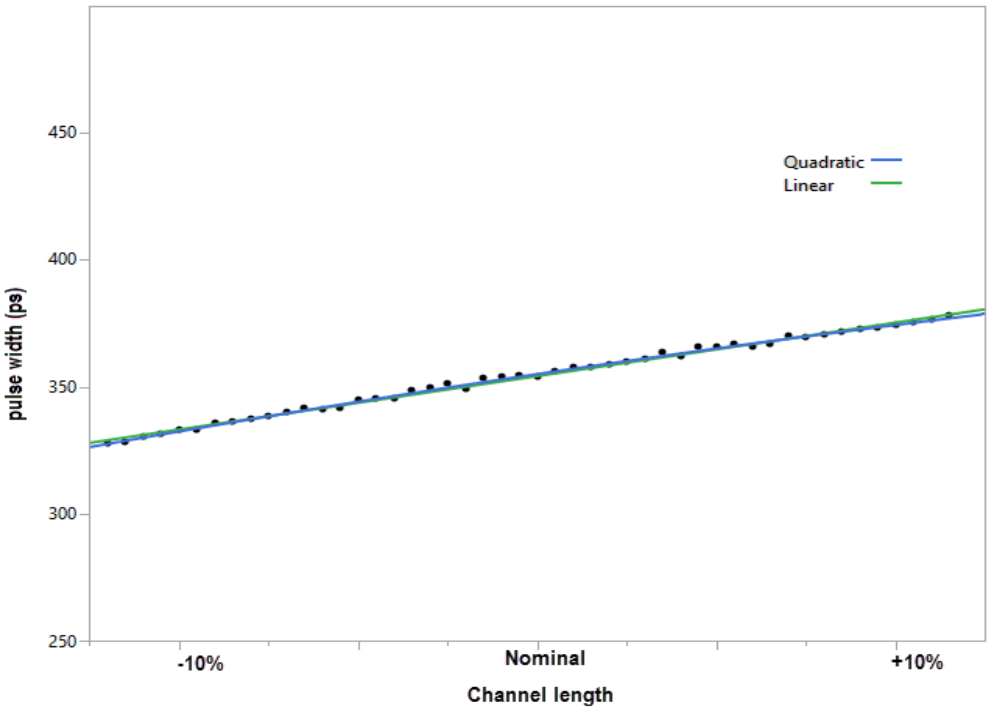

Fig. 8 SET pulse width vs. Effective channel length for 45nm bulk NAND2

In order to statistically model SET pulse width vs. channel length, both linear and quadratic polynomial models are used here, as shown in Table 2. Here, the mean square error (MSE) is used to evaluate the fitness of the model to the data. The root mean square error indicator (RMSE) measures the average difference between the model and the simulated data.

$$\mathrm{MSE}(f_p) = \frac{1}{n}\sum_{i=1}^{n} \left(f_p(X_i) - Y_i\right)^2 \qquad (2)$$

where n is the number of simulated data points, X_i is the input vector, and Y_i is the response corresponding to input vector X_i .

By comparing the RMSE of the linear and quadratic models, SET pulse width vs. channel length is better described by the linear model.

Table 2 Linear and quadratic model of the SET pulse width with channel length

Model	MSE	RMSE
Linear	0.742215	0.861519
Quadratic	1.241982	1.114442

III.2.1.2 PMOS threshold voltage

Threshold voltage variations result from oxide thickness and dopant fluctuation issues. To account for the parameter variations in Spectre simulations, the PMOS threshold voltage is specified in the PTM model files and the variable lies within the range of nominal value $\pm 3\sigma$. All the other parameters stay at their nominal values. The simulation results of the SET pulse width vs. PMOS threshold voltage are shown in Fig. 9. The mechanism is that the conducting PMOS device in the NAND2 provides the path through which the restoring current flows. An increase in threshold voltage in the PMOS device (a decrease of threshold voltage magnitude) induces a stronger restoring current and shortens the SET pulse.

Table 3 Linear and quadratic model of the SET pulse width with PMOS threshold voltage

Model	MSE	RMSE
Quadratic	1.311712	1.1453
Linear	19.83998	4.454209

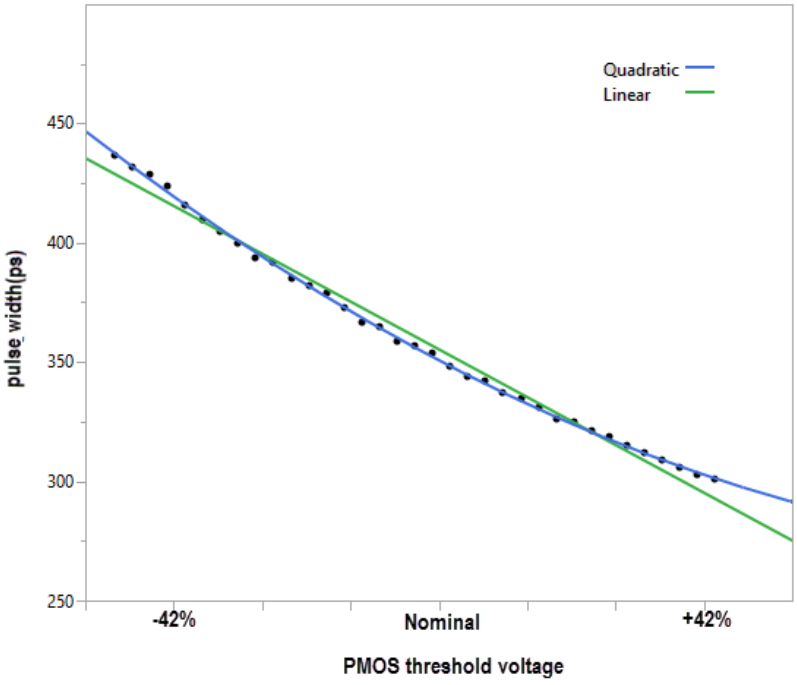

Fig. 9 SET pulse width vs. PMOS threshold voltage for 45nm bulk NAND2

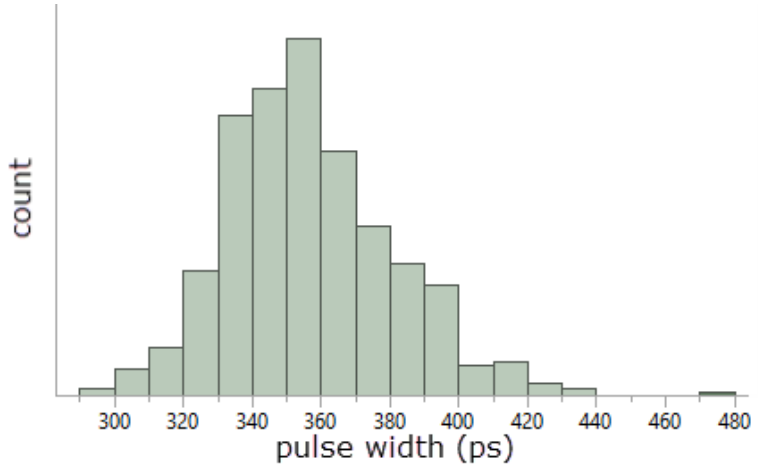

Fig. 10 SET pulse width distribution due to V_{th} variation of the pMOSFET

The relationship of SET pulse width to PMOS device threshold is approximately quadratic. The quadratic function of a Gaussian-distributed variable follows a non-Gaussian distribution [17].

This means that if the threshold voltage follows the Gaussian distribution, the pulse width distribution will not be Gaussian due to their non-linear relationship. In order to check the distribution of the SET pulse width with a Gaussian distributed Vth_p, 1000 Monte Carlo simulations are conducted and the result is plotted in Fig. 10. The SET pulse width is on the X-axis and count is on the Y-axis. The result shows that the simulated SET pulse width distribution is different from a Gaussian distribution, which is induced by the non-linear relationship between the SET pulse width and the threshold voltage. The approximated Gaussian distribution uses the same value of μ and σ as the simulated pulse width distribution, and it is reported in Table 4.

Table 4 Pulse width variation due to V_{thp} variation

	Value
Mean(ps)	356.4074
Sigma(ps)	23.7349
$3\sigma/\mu$	19.98%

III.2.1.3 NMOS threshold voltage

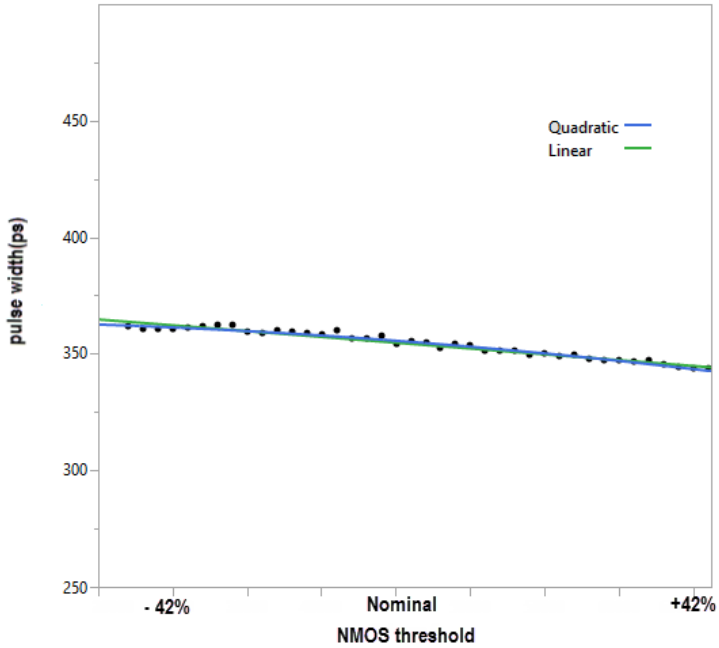

Fig. 11 SET Pulse width vs. NMOS threshold for 45nm bulk NAND2

Similarly, the simulated SET pulse width vs. NMOS threshold is shown in Fig. 11. This result shows the impact of NMOS threshold voltage on the SET pulse width is negligible. Because the NMOS device driven by “0” is working in the cut off region and the two NMOS devices are connected in series, the threshold voltage of the NMOS devices trivially affects the collected charge or restoring current under the “1 0” input bias condition. The illustration is shown in Fig. 12.

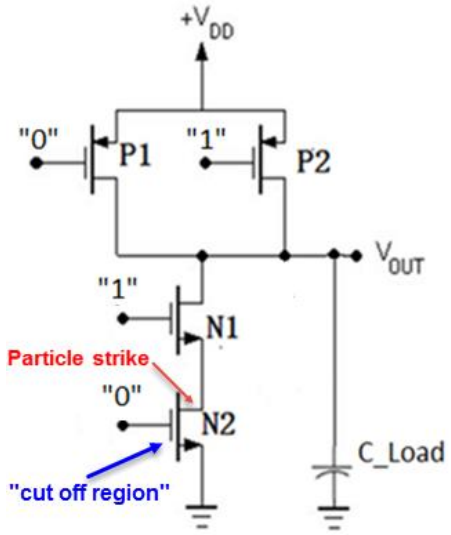

Fig. 12 nMOSFET N2 stays in cut off region under input “1 0” condition

III.2.2 Power Supply

Power supply is another factor that affects the SET pulse width of a logic gate. In 45nm bulk PTM devices, the nominal voltage is 1.0 V. In application, the voltage can be set to a value smaller than 1.0 V if the circuit is working in low-power mode or a value larger than 1.0 V in high-performance mode. In this work, the nominal voltage is configured as 1.0 V and the variation is set as a uniform distribution within 0.8 V – 1.2 V. In the Spectre simulations, the power supply voltage is specified in the netlist files and the variable lies within 0.8 V – 1.2 V. All the other parameters stay at their nominal values.

A plot of the simulated SET pulse width as a function of V_{dd} is shown in Fig. 13. Here, V_{dd} is on the X-axis and the SET pulse width is on the Y-axis. A decreasing supply voltage decreases the drive currents and hence it takes a longer time to restore the node back to its original state. As a result, a lower supply voltage leads to a longer SET pulse. The result shows that SET pulse width

is not a linear function of V_{dd} and the quadratic model describes the relationship better than the linear model.

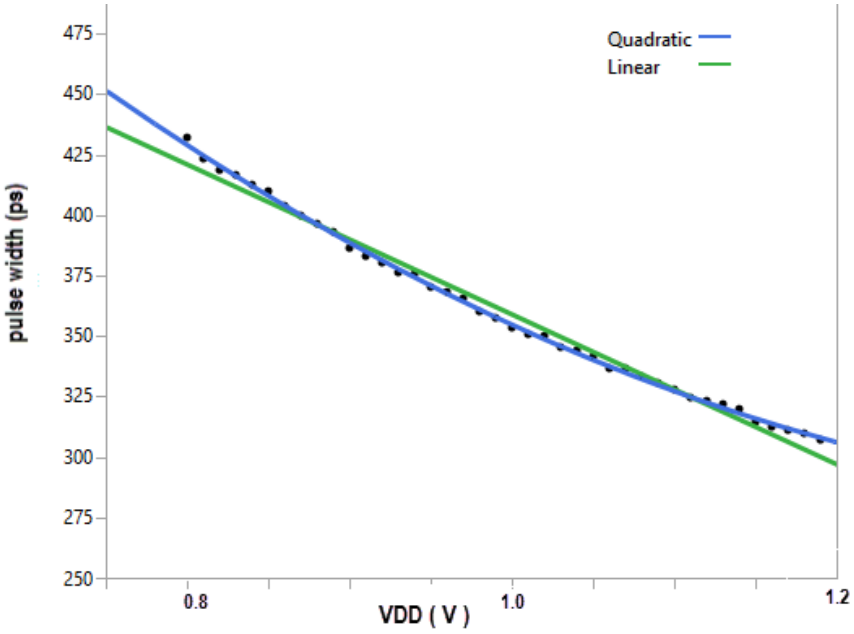

Fig. 13 SET pulse width vs. V_{dd} for 45nm bulk NAND2

Table 5 Linear and quadratic model of the SET pulse width with V_{dd}

Model	MSE	RMSE
Linear	16.96068	4.118335
Quadratic	1.602269	1.265808

III.2.3 Temperature:

In this work, the temperature of interest ranges from 25 Celsius to 100 Celsius with a uniform distribution. The temperature variable is specified in the netlist files and the variable lies within

25 Celsius to 100 Celsius. All the other parameters stay at their nominal values. The simulated SET pulse width vs. temperature is shown in Fig. 14.

For off-NMOS strikes in a NAND2 under the "10" input bias, the restoring transistor is the conducting PMOS transistor (and vice versa for a PMOS strike). The drive strength of the restoring transistor depends on the channel mobility. With increasing temperature the hole mobility in the channel of the pMOSFET decreases [17]. This weakens the drive strength of the restoring pMOSFET, which leads to a larger SET pulse width, as confirmed by Fig.14. Table 6 shows a linear relationship of the SET pulse width to the temperature, which is consistent with Gadlage's experimental result in [18].

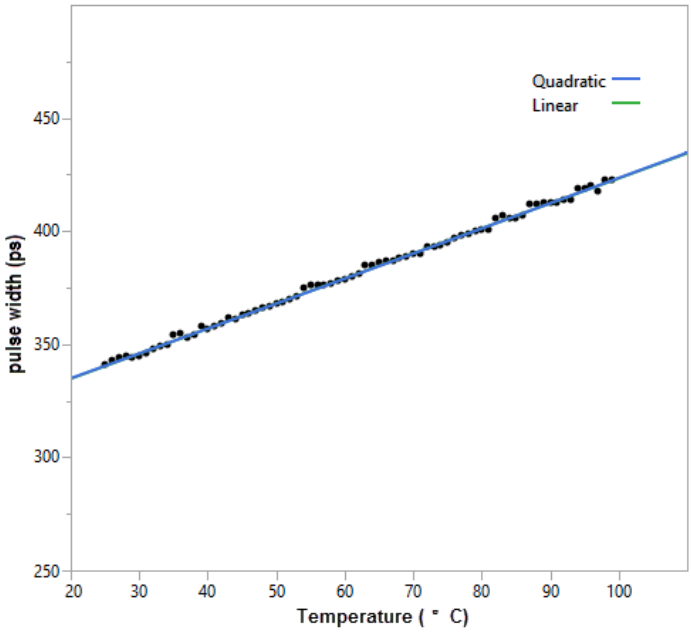

Fig. 14 Pulse width vs. temperature for 45nm bulk NAND2

Table 6 Linear and quadratic modeling error for temperature

Model	MSE	RMSE
Linear	1.373656	1.172031
Quadratic	1.389988	1.178977

III.2.4 LET

In this work, the collected charge from a single event strike is assumed to be proportional to LET and the range of interest is chosen from 6 MeV-cm^2/mg to 10 MeV-cm^2/mg. The reason of the LET range selection is for the ease of validating the proposed modeling approach which will be discussed in Chapter IV. Biased-current model [11] is used to convert the LET variable to the corresponding current source files in Spectre simulation. The LET variable is sampled between 6 MeV-cm^2/mg and 10 MeV-cm^2/mg and all the other parameters stay at their nominal values. The simulation result of the SET pulse width vs. LET is shown in Fig. 15. A larger LET means a larger collected charge, which lead to longer time for the current to restore the perturbed voltage back to normal. Table 7 indicates that the quadratic model outperforms the linear model to describe the relationship of the SET pulse width to the LET, even in the relatively small range selected in this work.

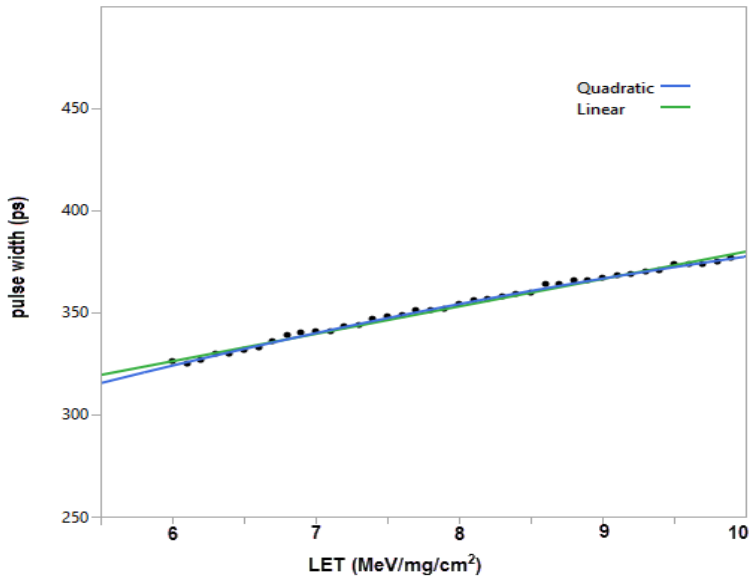

Fig. 15 SET Pulse width vs. LET for 45nm bulk NAND2

Table 7 Linear and quadratic modeling error for LET

Model	**MSE**	**RMSE**
Quadratic	0.750756	0.866462
Linear	1.827925	1.352008

III.2.5 Load Capacitance:

In Spectre simulation, the selected loading inverter size ranges from 1X to 3X minimum inverter size. The load capacitance variable is specified between 1X to 3X of the input capacitance of a minimum size inverter. All the other parameters stay at their nominal values. The simulated SET pulse width vs. loading capacitance is shown in Fig. 16. A larger capacitance results in a larger time constant according to (3), and this causes a larger SET pulse width because it takes more time for the perturbed node to be charged to the original voltage.

$$\tau = R_{equ} C_{load} \quad (3)$$

where R_{equ} is the equivalent resistance through which the restoring current flows to recharge the perturbed voltage

A possible reason for the relatively small dependence of the SET pulse width on the loading capacitance is the fact that the loading capacitance determines the rise and fall time of the transient, but the duration of the SET pulse is governed by the restoring drive current rather than the loading capacitance within the selected capacitance range [19].

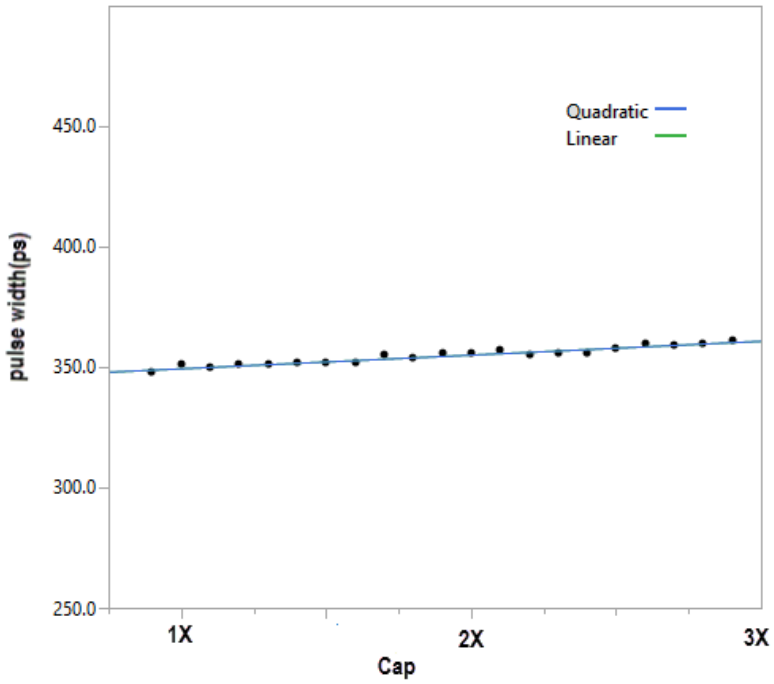

Fig. 16 SET pulse width vs. loading capacitance for 45nm bulk NAND2

Table 8 Linear and quadratic modeling error for loading cap

Model	MSE	RMSE
Linear	0.922237	0.960332
Quadratic	0.973096	0.986456

III.2.6 W/L Ratio

The ratio of width to length is an important design parameter in logic circuits. In order to obtain a balanced rise/fall time, the ratio of width to length for the PMOSFET is twice of that for the NMOSFET. In this work, the ratio of width to length for both the PMOSFETs and NMOSFETs are altered simultaneously to keep the balanced rise/fall time, and the range of the ratio varies from the minimum ratio to three times this ratio. All the other parameters stay at their nominal values.

The simulated SET pulse width vs. W/L ratio is shown in Fig. 17. From equation (1), a larger W/L ratio means a higher restoring current, which shortens the SET pulse. Although the corresponding output loading capacitance also changes with W, the dominant factor is the restoring current since the output capacitance has little impact on pulse width, which is verified by the simulation results in the previous section.

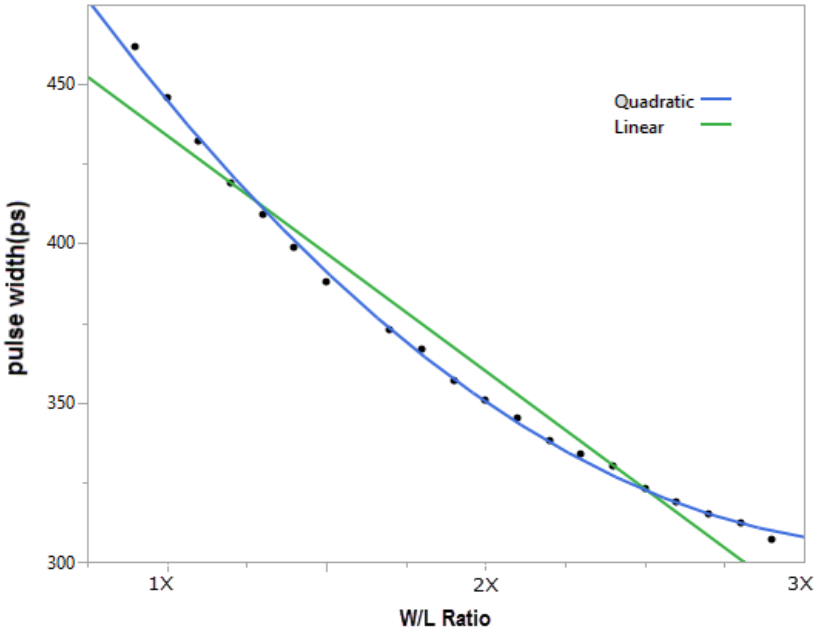

Fig. 17 SET pulse width vs. W/L ratio for 45nm bulk NAND2

Table 9 Linear and quadratic modeling error for W/L ratio

Model	**MSE**	**RMSE**
Quadratic	4.4946948	2.1200695
Linear	85.699545	9.2574049

Another observation is that a quadratic relationship better describes the relationship of the SET pulse width to the W/L ratio. It indicates that the SET pulse width does not follow a uniform distribution even if W/L follows a uniform distribution because of the non-linear relationship between the two variables.

III.3 Conclusion

In this chapter, the impact on the SET pulse widths of process variations and operation parameters is investigated. First, the parameters that may statistically impact the SET pulse width are identified by literature review. Then Spectre simulations are conducted to evaluate the impact on the SET pulse width of these parameters individually. The mechanisms are discussed based on the results. Interestingly, non-linear relationships between the SET pulse width and input parameters are identified, which indicates that the SET pulse width does not necessarily follow a Gaussian distribution even if the process variations follow Gaussian distributions. Therefore, it is necessary to quantitatively describe the SET pulse width distribution, which will be discussed in the next chapters.

CHAPTER IV

RESPONSE SURFACE MODELING OF SET PULSE WIDTH WITH MULTIPLE PARAMETERS

In radiation environments, accurate statistical modeling of SET pulse widths is very important in digital circuits since process variations and environmental and operational variables all impact the SET pulse width. In this chapter, we introduce Design of Experiments (DOE) as a cost-effective way to statistically rank the impact of multiple parameters on the SET pulse width and screen unimportant parameters, and then a response surface of the SET pulse width is modeled with the identified important parameters.

IV.1 Parameter screen by Design of Experiment (DOE) method

With feature sizes shrinking, the number of device parameters increases dramatically. In order to reduce modelling complexity and reduce the input parameter dimensions, assumptions are made by the modeler to determine which inputs are more likely to be important than the others, or one-at-a-time searching for the vital factors is used. However, input parameter selection by intuition rather than quantifying techniques is not reliable and the one-at-a time method is costly and time-consuming, especially for high-dimension input parameters.

Design of Experiment (DOE) techniques, which were originally developed for physical experiments, can be applied here to efficiently determine the impact of the input parameters on the output response. A properly designed experimental plan can drastically reduce the number of experimental runs that would otherwise be required if the one-at-a-time method is used. In this section, the background knowledge of screening design and the DOE method are introduced in

the first part, and then a NAND2 example is used to illustrate the application of the screening technique.

IV.1.1 Full factorial and fractional factorial DOE methods

The definitions of some terms related to DOE are listed here:

- Factors: input parameters; N stands for the number of input parameters
- Level: the discretized value that a factor takes; K stands for the number of levels
- Coded variables: It is convenient to transform the natural variables to coded variables, which are usually defined with zero mean and the same standard deviation:

  -1 level represents the smallest value in the range of the factor and +1 level represents the largest value in the range of the factor

- Sampling: The process of choosing values for input parameter combinations is called sampling. The points chosen using the DOE theory ensures less experiment/simulations are conducted and the best possible precision of the mathematical response surface can be obtained.
- Run: Simulation or experiment with selected input parameter combination
- Confound: The effects of some interactions cannot be distinguished from other interactions

The 2-level fractional factorial (FF) method is an effective approach for screening factors; i.e., it can detect important factors correctly, as long as the assumptions are not violated [20]. A comparison between the exhaustive method and the FF method is listed in Table 10.

Table 10 Comparison between exhaustive method and FF method

	Exhaustive method	FF method
Application	Evaluate parameters' impacts on the response when the closed-form function is not known	
Accuracy	The effect of each parameter and the interaction between them are evaluated	High order interactions between parameters are neglected
Number of tests	K^N, which is a huge number when K or N is large	A subset of exhaustive method

In the 2-level FF method, each factor takes two levels: largest value and smallest value. The full factorial method includes 2^K runs and a FF method includes $2^{(k-p)}$ runs, where p is related to the confounding. Generally speaking, fewer test runs induce more confounding. Thus, p should be selected to balance the number of runs and the accuracy of the result. A 3-input 2-level FF method is shown here as an example to illustrated the underlying mathematical principles.

Table 11 The 2^3 experiment design plan

run	Factor level			Interaction level				Response
	X1	X2	X3	X1X2	X1X3	X2X3	X1X2X3	
1	+1	+1	+1	+1	+1	+1	+1	Y1
2	-1	+1	+1	-1	-1	+1	-1	Y2
3	+1	-1	+1	-1	+1	-1	-1	Y3
4	-1	-1	+1	+1	-1	-1	+1	Y4
5	+1	+1	-1	+1	-1	-1	-1	Y5

6	-1	+1	-1	-1	+1	-1	+1	Y6
7	+1	-1	-1	-1	-1	+1	+1	Y7
8	-1	-1	-1	+1	+1	+1	-1	Y8

The 2^3 full factorial design shown in**Error! Reference source not found.** allows estimation of ll main effects (X1 , X2 , X3), all two factor interactions(X1 , X1X3 and X2X3), as well as the three factor interaction (X1X2X3). In order to find out how a specific factor influences the response, we evaluate the differences between the average of the response when the factor is high and the average of the response when the factor is low. That is, the main (or individual) effect U_i of a factor x_i to the response is defined as half the difference of the average response values:

$$U_i = \frac{1}{2}\left\{\frac{\sum_{k\in K_i^+} y_k}{n_r/2} - \frac{\sum_{k\in K_i^-} y_k}{n_r/2}\right\} = \frac{1}{n_r}\sum_{k=1}^{n_r} x_{ik} \times y_k \qquad (4)$$

where n_r is number of experimental runs

y_k :response of the kth run

K_i^+ :set of run indexes where x_i is +1

K_i^- :set of run indexes where x_i is -1

x_{ik}: k_{th} element of x_i

In order to determine how two factors jointly affect the response, we compute the difference between the average value of the response when both factors are at the same level, i.e., both high and both low, and the average value of the response when both factors take on different levels.

Thus the interaction (or joint) effect, $U_{i\times j}$, of factors x_i and x_j, i $\neq$ j, is defined as half the difference of the effects due to x_i given that x_j is at + 1 and - 1, respectively. That is,

$$U_{i\times j} = \frac{1}{2}\{U_{i|j=+1} - U_{i|j=-1}\} = \frac{1}{n_r}\sum_{k=1}^{n_r} x_{ik} \times x_{jk} \times y_k \quad (5)$$

The size of a full factorial matrix increases exponentially with the number of factors; this leads to an unmanageable number of runs. Fractional factorial designs are used when many factors are present. A half fraction of the 2^3 full factorial designs is shown in Table 12. The algorithm for systematically deriving such a design plan can be found in the JMP 11 pro software [21].

Table 12 2^{3-1} experiment design plan

run	Factor level			Interaction level				Response
	X1	X2	X3	X1X2	X1X3	X2X3	X1X2X3	
1	+1	+1	+1	+1	+1	+1	+1	Y1
2	-1	+1	-1	-1	+1	-1	+1	Y2
3	+1	-1	-1	-1	-1	+1	+1	Y3
4	-1	-1	+1	+1	-1	-1	+1	Y4

It is noted that the values in column X1X2X3 of Table 12 are now identical. From (3) and (4), we are no longer able to distinguish the impact from the third-order interactions. Such effects are defined as being aliased with the others and quantified as resolution or degree of confounding. The trade-off is the number of runs and the resolution of the result. Two standard resolutions are listed here:

- Resolution = 4 means that main effects are not confounded with other main effects or two-factor interactions. However, two-factor interactions are confounded with other two-factor interactions.
- Resolution $\geqslant$ 5 means there is no confounding between main effects, between two-factor interactions, or between main effects and two-factor interactions. So the impact of each parameter and all the second-order interactions on the output can be evaluated.

In general, the more accurate the model is, the higher the overhead cost needed to build them. For fractional factorial designs with fewer runs, we have to pay the price of lower resolution. Typically, the impact of high-order interactions can be assumed to be negligible, compared to the impact of the main factors and low-order interactions [22]. The loss of ability to distinctly estimate the impact of all-order interactions is an acceptable tradeoff against the expenses of experiments/simulations.

IV.1.2 Application of fractional-factorial method to SET induced pulse width

To illustrate the application of fractional factorial design to screen unimportant input parameters, a NAND2 is shown here as an example.

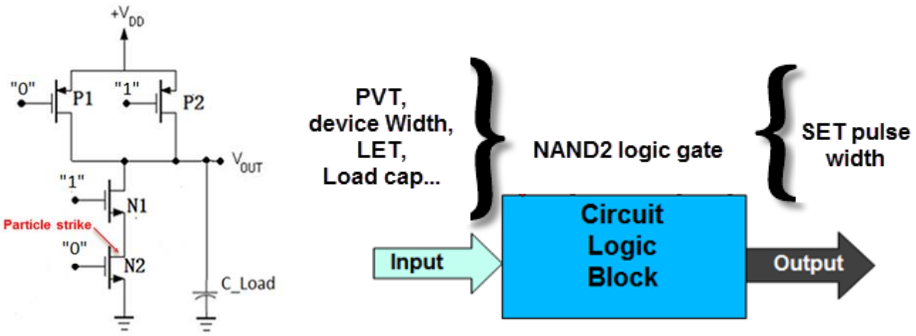

Fig. 18 Example of applying FF to NAND2 gate for parameter screening

For the purpose of illustrating the process of screening, eight parameters are selected as the input parameters. The distributions of the effective channel length, threshold voltage are based on international technology roadmap for semiconductors report [23]; the voltage supply selection is based on the PDK used in the work; the ratio of channel width over length and loading capacitance are empirical design parameters. It is noticed in simulations that if the LET changes in a larger range, the effect of the LET is much larger than the other parameters and all the other parameters are screened. As a result, the LET value selection here is adopted for the purpose of evaluating SET pulse widths as a multivariate. JMP (pro 11) is used to generate the required experimental designs and conduct statistical analysis. The parameter candidates are listed in

Table 13.

Table 13 parameter list for screening

Parameters	Low Value	High level	Variations
leff (Effective channel length)	18.45nm (-3 σ)	26.55nm (+3 σ)	σ = ± 6% Gaussian

V_{thn} (Threshold voltage of NMOS)	0.27028V (-3 σ)	0.66172V (+3 σ)	σ = ± 14% Gaussian
V_{thp} (Threshold voltage of PMOS)	-0.58362V (-3 σ)	-0.23838V (+3 σ)	σ = ± 14% Gaussian
LET	6MeV- cm^2/mg	10MeV- cm^2/mg	6-10 MeV- cm^2/mg Uniform
V_{dd} (Power supply)	0.8V	1.2V	0.8-1.2V Uniform
W/L	1X minimum	3X minimum	1X – 3X minimum Uniform
Cap (Load capacitance)	1X minimum inv	3X minimum inv	1X – 3X minimum inv Uniform
Temp (Temperature)	25Celsius	100 Celsius	25-100 Celsius Uniform

In order to evaluate the impact of these parameters, an 8-factor-2-level design is required. A design matrix of resolution V is selected from JMP. A design of resolution V provides the ability to evaluate all the main factors and two factor interactions without aliasing each other. The resolution V FF design requires 64 simulations. Part of the design matrix is shown in **Error! eference source not found.**. The -1/+1 means the smallest/largest value a parameter could take

within its distribution. Each row indexed with the natural number stands for a simulation with the assigned input parameter settings. The SET pulse widths are simulated by Spectre simulator with corresponding PTM model files.

Table 14 Parameter screen design matrix for the NAND2

	Pattern	leff	Vth0n	vth0p	LET	VDD	width	cap	temp	pulse width (s)
1	−−−−−−−+	-1	-1	-1	-1	-1	-1	-1	1	6.40969e-10
2	+−+++++−	1	-1	1	1	1	1	-1	1	3.85504e-10
3	++−−+++−	1	1	-1	-1	1	1	1	-1	3.86154e-10
4	+−−++++−	1	-1	-1	1	1	1	1	1	2.9445e-10
5	++−++−+−	1	1	-1	1	1	-1	1	-1	6.29204e-10
6	+++−−−−−	1	1	1	-1	-1	-1	-1	-1	5.18965e-10
7	−−+−++++	-1	-1	1	-1	1	1	1	1	5.06077e-10
8	+−−+−−++	1	-1	-1	1	-1	-1	1	1	2.76905e-10
9	−+−+−−−−	-1	1	-1	1	-1	-1	-1	-1	4.46276e-10
10	+−+−−−+−	1	-1	1	-1	-1	-1	1	-1	2.20003e-10
11	−−−−+−+−	-1	-1	-1	-1	1	-1	1	-1	2.75661e-10
12	++−+++−+	1	1	-1	1	1	1	-1	1	1.87486e-10
13	++−−+−−+	1	1	-1	-1	1	-1	-1	1	4.18073e-10
14	−+++−−+−	-1	1	1	1	-1	-1	1	-1	3.35348e-10
15	+−+−+++−	1	-1	1	-1	1	1	1	-1	3.99253e-10
16	++−+−−−+	1	1	-1	1	-1	-1	-1	1	1.85201e-10
17	−−+++−++	-1	-1	1	1	1	-1	1	1	6.51911e-10
18	−−−−++−+	-1	-1	-1	-1	1	1	-1	1	3.62572e-10
19	−−+−−−++	-1	-1	1	-1	-1	-1	1	1	3.52919e-10
20	−−−+−−+−	-1	-1	-1	1	-1	-1	1	-1	2.8825e-10

After the 64 simulations listed in the matrix are conducted, the impact rank of the input parameters on the SET pulse width is obtained by numerical calculations and shown in Fig. 19. The Y axis is the normalized pulse width variant induced by a single parameter across all the other parameter combinations, as calculated by equation 4.

As we can see from Fig. 19, the threshold voltage of the PMOSFET, V_{dd}, and device width are negatively correlated to the SET pulse width; the dependence of the load capacitance and the NMOS threshold voltage on the SET pulse width is relatively small. By comparing 3^8 =6561 exhaustive simulations, the fractional factorial method successfully identifies important parameters, with only a small number of simulations. The conclusions are consistent with the mechanisms discussed in chapter III.

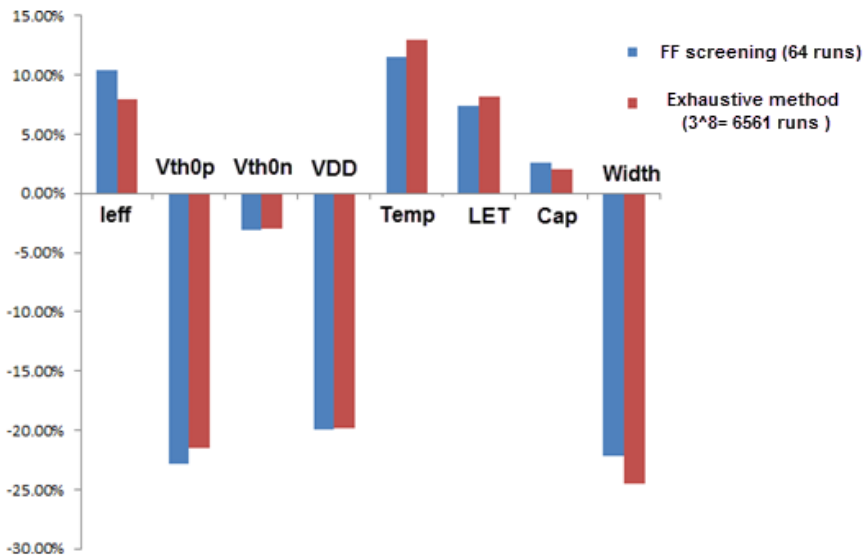

Fig. 19 Parameter rank by FF method and exhaustive method

IV.2 Response Surface Modeling (RSM) of the SET pulse widths

Mathematically, the response surface is a multivariate of random variables which represent input parameters. RSM is conducted after the important factors are selected from unimportant ones (screening) to find a mathematical equation relating input parameters $\vec{x}$ and output response $\vec{y}$:

If the true function between the input and the output is

$$\vec{y} = f(\vec{x}), \quad (6)$$

then the RSM of the true function is

$$\hat{y} = g(\vec{x}) \quad (7)$$

$$\vec{y} = \hat{y} + \varepsilon, \quad (8)$$

where ε represents both the error of approximation and measurement (random) errors.

The most common RSM approach is to apply the DOE to identify an efficient set of input parameter settings $(x_1, x_2, \ldots, x_n)$ and then use regression analysis to create a polynomial approximation of the underlying physics models, as illustrated in the block diagram of Fig. 20 (the output response is the SET pulse width in our case). A typical flow of building the response surface involves

1) Choosing sampled input parameter combinations
2) Obtaining the output pulse widths at the selected parameter combinations using simulations/experiments
3) Calculating the coefficients in the model with input parameters—output SET pulse width

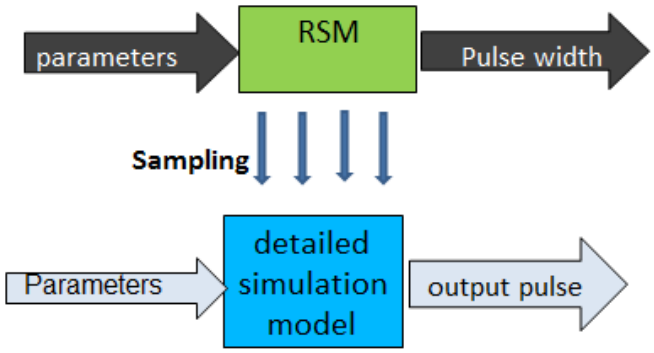

Fig. 20 Modeling input-output RSM for pulse width distribution

Model accuracy may be sacrificed if the number of simulation runs is reduced. A proper modeling strategy allows balancing of the conflicting requirements of reducing simulation runs and providing model accuracy at the same time.

IV.2.1 Central composite design (CCD) for quadratic RSM

The central composite design (CCD) is one of the most common methods to build a quadratic response surface. In the CCD scheme, data points are sampled in a fashion that combines both a cube and a star shape in the design space. Each factor in the CCD scheme has five levels, 0, ±1, and ±α [24]. Fig. 21 shows the CCD plan pictorially for the case of two factors. Factorial points and axial points are sampled as input parameters combinations for RSM. The square sub plan in blue circles is a two-level fractional factorial design discussed in section IV.1. In order to accurately evaluate all the first-order and cross-factor interactions in RSM, this fractional factorial plan must be of resolution V or higher [25]. In addition, a supplementary design plan is needed to estimate the coefficients of the pure quadratic terms in the RSM. One such supplementary plan is the star design plan, shown as red diamonds in Fig. 21. Each factor in the star plan can take three levels, i.e., 0 and ±α, and this plan requires a total of 2m + 1 experimental runs. The value of α is chosen so that the composite plan exhibits the rotatability property[26].

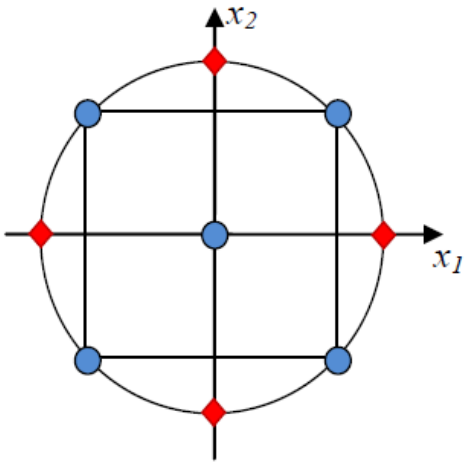

Fig. 21 CCD design for two factors [27]

Note that for values of α other than 1, each factor is evaluated at five levels. The total number of runs is

$$N = 2^m + 2m + 1 \quad (9)$$

where 2^m represent factorial points, $2m$ represent the axial points, and 1 is the center point.

The main advantage of the composite scheme described above is that it allows for the estimation of all the coefficients of a second-order model but requires fewer runs than a three-level fractional factorial design. Another advantage is that the simulated/experimental data used for previous variable screening can be reused for RSM since the matrix of CCD design overlaps with the matrix of 2-level factorial design used in the variable screening step.

With the obtained data sampling points and corresponding response, a second order analytical model could be derived with matrix calculation. That is

$$f(x) = x_i^{\boldsymbol{T}} \boldsymbol{A} x_i + \boldsymbol{B}^{\boldsymbol{T}} x_i + \boldsymbol{C} \quad (10)$$

where $\begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{B} \end{bmatrix} = (X^{\boldsymbol{T}} X)^{-\mathbf{1}} X^{\boldsymbol{T}} \boldsymbol{Y}$ and $X = \begin{bmatrix} 1 & x_1^{\boldsymbol{T}} \\ \vdots & \vdots \\ 1 & x_n^{\boldsymbol{T}} \end{bmatrix}$

IV.2.2 Methods for building highly non-linear response surface

Screening design combined with central composite design (CCD) is sufficient to model second order output response of logic gates. When the second order RSM could not provide satisfactory accuracy, (for example, when the parameters vary over a very large range, the second order RSM may not work well), there are also methods to describe a highly non-linear response. In this

section, two highly non-linear modeling methods are discussed as supplements for the quadratic RSM discussed in the previous sections.

IV.2.2.1 Neural Network

The concept of neural networks comes from the way that human brain processes information. A typical human brain has approximately 10 billion neurons in the cortex and 60 trillion synapses which connect the neurons together. This arrangement is a highly complex, nonlinear, and parallel structure. Consequently, human brains are efficient for information processing and learning. Artificial neural networks emulate the way the human brain solves problems and ANN can be used to model input-output relationships [28]. It is widely used to model high-dimensional, non-linear data. The model has several layers:

- Input layer: connected with original input variables
- Hidden layer: a set of constructed variables
- Output layer: response variables

Each of the variables in a layer is called a node and Fig. 22 shows a typical three-layer artificial neural network. The number of input and output neurons is defined by the specification of the problem at hand. The number of hidden units to use is not known in advance and must be specified or determined through experimentation. A good starting point is to use one hidden layer, and the number of neurons in the hidden layer is varied to find the best fit to the data.

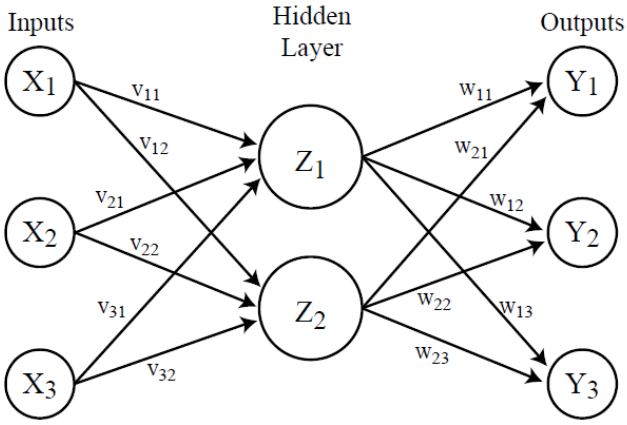

Fig. 22 Diagram of typical ANN for function approximation[29]

A node takes a transformed linear combination of the outputs from the nodes in the lower layer as its input. That is, hidden layer nodes takes a_u as a linear combination of the input variables:

$$a_u = \sum_{j=1}^{p} \left(v_{ju}\, x_j + \theta_u\right) \qquad (11)$$

where the w_{ju} are unknown parameters that need to be trained by data and θ_u is a parameter that plays the role of an intercept in linear regression.

Then it sends a transformed function as its outputs to the nodes in the next layer. The transformation function is usually sigmoidal or linear, which is defined by an activation function or transfer function and noted as

$$z_u = g(a_u) \qquad (12)$$

$$g(x) = 1/(e^{-x} + 1) \qquad (13)$$

Finally,

$$y_k = \sum_{j=1}^{p} w_{tk}\, x_j \,.\, g\Big(\sum_{j=1}^{p} v_{ju}\, x_j + \theta_{1j}\Big) + \theta_{2u} \qquad (14)$$

The model is flexible to fit universal approximation functions. However, large amounts of data are needed to train the coefficients of the neural network. In addition, there is a possibility of over-fitting, in which condition the neural network could provide perfect fitting to training data, but behave badly to predict new data [28].

IV.2.2.2 Kriging modeling method

Kriging models, or spatial correlation models, evolved in the field of geostatistics, and have recently become popular in the area of statistical modeling [30]. In Kriging models, spatial correlation functions are assigned for the outputs of the multi-dimensional input parameters. Response is predicted based on correlation functions weighted by the simulated/experimental output responses [31]. The general form of the stochastic response as a function of the input vector x is:

$$f(x) = \sum_{n=1}^{M} a_m b_m(x) + Z(x) \quad (15)$$

where a_m is the unknown coefficient for b_m, and the stochastic component $Z(x)$ is a random process, commonly assumed to be Gaussian, with zero mean and covariance

$$\mathrm{Cov}[\mathrm{Z(x)}, \mathrm{Z(x')}] = \sigma^2 R(x, x') \quad (16)$$

where σ^2 is the process variance and $R(x, x')$ is the correlation function

One typical selection for the correlation function is of the form

$$R(x, x') = \prod_{j=1}^{p} \exp\{-\theta_j \left| x_j - x'_j \right|^{\rho_j}\} \quad (17)$$

where $\theta_j \geq 0$ and $1 \leq \rho_j \leq 2$

By the least-square method, b_m could be estimated with given input vectors and corresponding output response. A primary advantage of kriging is that exact predictions are achieved at the observed values. However, a drawback is that the estimated kriging model parameters are computationally intensive to obtain.

IV.2.3 An example of using Response Surface Modeling (RSM) and DOE for predicting SET pulse width

In the previous sections in this chapter, parameters are screened by the fractional factorial method based on DOE theory, and then several RSM building methods are investigated. The algorithm of RSM for SET pulse width with multiple input parameters involves three steps and is illustrated in Fig. 23:

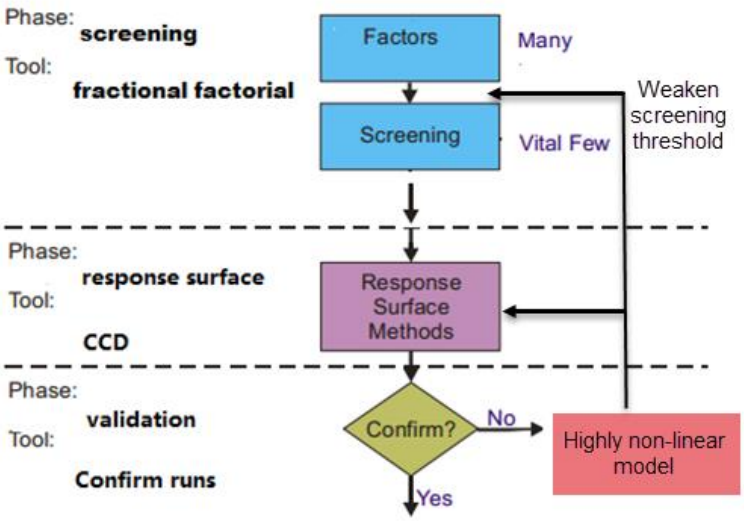

Fig. 23 RSM flow with reduced input parameter dimensions

In this work, RSMs of the SET pulse widths in NAND2 are investigated. The input parameters and their distributions are the same as those used in the screening step. In section IV.1.2, V_{thn} and the loading capacitance are screened by the FF method and the remaining six parameters are used for RSM. The parameter selections are shown in Table 15.

Table 15 Parameter value for CCD

Parameters	Nominal level	Low level	High level
leff (Effective channel length)	22.5nm	18.45nm (-3 σ)	26.55nm (+3 σ)
V_{thp} (Threshold voltage of PMOS)	-0.411V	-0.58362V (-3 σ)	0.23838V (+3 σ)
LET	8MeV/mg/cm^2	6MeV/mg/cm^2	10MeV/mg/cm^2
V_{dd} (Power supply)	1.0V	0.8V	1.2V
W/L	2 X minimum	1X minimum	3X minimum
Temp (Temperature)	62.5 Celsius	25Celsius	100 Celsius

Similarly to the fractional factorial design matrix, JMP 11 pro is utilized to obtain the CCD design matrix for the six input parameter combinations. A total number of 46 SET pulse widths are required for the RSM and 33 of them have already been obtained in the screening step. Spectre simulations are used to obtain the SET pulse widths for the remaining 46-33=13 parameter combinations. Matrix calculation is conducted with Wolfram Mathematica 8.0 [32] to derive the quadratic response surface model.

$$Pulse\ width = x_i^{\boldsymbol{T}} \boldsymbol{A} x_i + \boldsymbol{B}^{\boldsymbol{T}} x_i + \boldsymbol{C} \quad (18)$$

$$x_i = [leff\ \ vth0p\ \ Vdd\ \ temp\ \ LET\ \ width] \quad (19)$$

$$A = \begin{bmatrix} -1.7495 & -12.75 & 0.5625 & -7.375 & -3.1875 & 5.9375 \\ & 14.2505 & 1.78E-15 & 28.5625 & 7.25 & -10.875 \\ & & -1.7495 & -0.375 & -1.4375 & 0.9375 \\ & & & 10.7505 & 3.125 & -2.5 \\ & & & & 27.2505 & -3.0625 \\ & & & & & 0.750503 \end{bmatrix}$$

$$B = \begin{bmatrix} 38.61765 \\ -85.4412 \\ 27.14706 \\ -74.2941 \\ -82.5588 \\ 41.64706 \end{bmatrix} \text{ and } \quad C = 366.119726$$

In this equation, the diagonal of matrix A contains the coefficients of the second order input parameters, the others are the coefficients of the interactions of the input parameters; the values in vector B are the coefficients of the first order input parameters. Both the second order of the input parameters and the interaction between the parameters impact the SET pulse width and cannot be neglected.

In order to validate the accuracy of the RSM, another 50 input parameter combinations are simulated with random sampled input parameter combinations. The error between the simulated SET pulse widths and those obtained by RSM is less than 8%. Consequently, highly-nonlinear RSM is not required in this example.

IV.2.3 Conclusion

In this chapter, design of experiment and response surface modeling are discussed in detail. Parameters that identified to affect the SET pulse width in Chapter III could be screened by design of experiment techniques by a small number of simulation runs at certain parameter combinations, and then central composite design or high non-linear methods can be used to build the response surface. An example of a NAND2 logic gate is illustrated to show parameter screening and the RSM of the SET pulse width. The result shows that by RSM with reduced input parameter dimensions, a predictive model could be constructed at only a fraction of the traditional simulation cost. The proposed method in this chapter serves as a basis for estimating the pulse width distribution in Chapter V.

CHAPTER V

ESTIMATE THE SET PULSE WIDTH DISTRIBUTION WITH MONTE CARLO METHOD AND RSM METHOD

Nowadays, designers spend a considerable amount of computing budget on ensuring that circuits perform under all possible conditions. Usually they perform the worst case analysis on all the possible corners, and this method may lead to overly pessimistic results for estimating the circuit response [33]. Similarly, the SET pulse width in nano-scale ICs is not deterministic due to process variations, voltage, and temperature [34]. As a result, the statistical distribution of the SET pulse width in standard cells is needed for reliability concerns.

In this chapter, first we discuss the standard MC simulations of the SET pulse width in section 1. Then in section 2 we develop a fast statistical moment estimation method based on RSM. Then the results from the proposed fast statistical method are compared with the standard Monte Carlo method. At the end, a summary is presented in Section 3.

V.1 Evaluate the SET pulse width with Standard Monte Carlo method

The Monte Carlo method is used in a variety of applications to approximate the output signal distributions [35]. It can be used in simulations to estimate the mean, variance, and other statistical moments of output variables. Typically, three steps to conduct a standard Monte Carlo method in circuit simulations are:

1. Input parameters of interest are sampled based on their distribution. This procedure takes more samples around the parameter values with higher probability and fewer samples with lower probability values.
2. After that, the transistor-level simulation is conducted for each combination of the sampled parameters. The desired simulation output is measured in each simulation, leading to the sample set of measured values. This process requires relatively long simulation time and large memory to store all the data.
3. The moments of the circuit simulation outputs are calculated using standard moment estimation equations on the generated sample set.

The accuracy of a standard Monte Carlo method is mathematically specified by the confidence interval. α-level confidence interval is defined as an interval where the statistical measurement falls with the probability α. For a given α, the more sampled points are obtained by Monte Carlo method, the closer the obtained moment is to the true value [36]. Practically the samples are on the order of thousands to tens of thousands [36] to ensure the accuracy of the Monte Carlo method. In this work, 10000 simulations are conducted with eight parameters sampled according to their distributions, as listed in Table 16. Spectre simulator and the PTM models are used to conduct the Monte Carlo simulation.

Table 16 Parameter variations and distributions

Parameters	Nominal Value	Variations	Distribution
leff (Effective channel length)	22.5nm	$\sigma = \pm 6\%$	Gaussian

V_{thn} (Threshold voltage of NMOS)	0.466V	$\sigma = \pm 14\%$	Gaussian
V_{thp} (Threshold voltage of PMOS)	-0.411V	$\sigma = \pm 14\%$	Gaussian
LET	8MeV/mg/cm^2	6-10MeV/mg/cm^2	Uniform
V_{dd} (Power supply)	1.0V	0.8-1.2V	Uniform
W/L	2 X minimum	1X – 3X minimum	Uniform
Cap (Load capacitance)	2 X minimum inv	1X – 3X minimum inv	Uniform
Temp (Temperature)	62.5 Celsius	25-100 Celsius	Uniform

The analog design environment (ADE) provided by Cadence Spectre can be used to conduct Monte Carlo simulations. It supports Monte Carlo simulation by adding statistics blocks in the circuit netlist files [37]. However, two significant issues occur during simulations: first, it is observed in the simulation experiments that the pulse width does not change for a channel width smaller than 45 nm while keeping the rest of the parameters at their nominal values; second, the pulse does not change by altering threshold voltage in the statistical block. A possible reason might be Cadence Spectre ADE does not support PTM models in the statistical block.

So instead of using statistical blocks, Spectre is used to simulate a netlist with the corresponding model files one by one to solve these two technical issues. As a result, 10000 netlist files with corresponding models should be generated. Due to the large amount of data processing, MATLAB2012b is used to generate the parameter value according to their distributions, and Python files are used as parsers to write the generated parameters from MATLAB to the

corresponding model files in Spectre; the simulated SET pulse widths data are sent back to Matlab for data processing.

In this work, the SET pulse width distributions are generated under all the input vectors and the results are shown in

Fig. 24. The red curves fit the data using nonparametric density estimation (kernel density estimation). The approximated normal distributions plotted with green curves use the same value of μ and σ of the SET pulse width distribution. It is observed in each input state that the obtained distribution is different from the normal distribution. This result is quite consistent with the published experimental data [38].

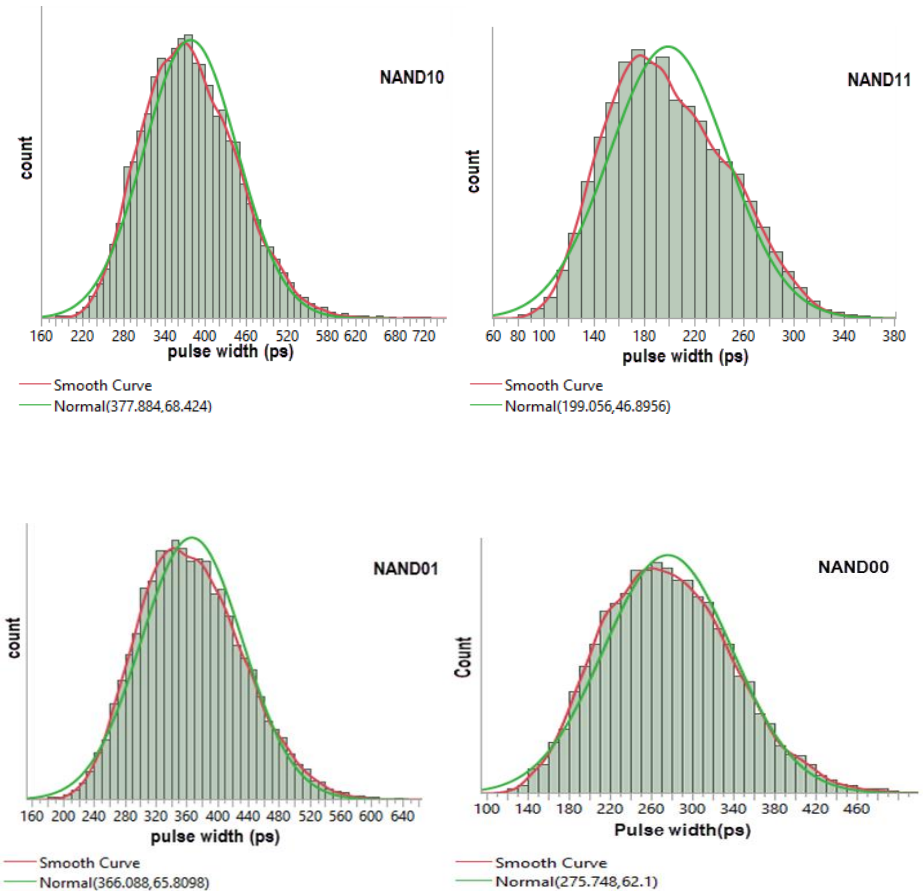

Fig. 24 The SET pulse width distribution of NAND2 gate under all input vectors

A method to quantify the discrepancy between the SET pulse width distribution and corresponding normal distribution uses statistical moments: skewness (γ) and excess kurtosis (κ). It is known that μ and σ are the first and second moments of a distribution. Similarly, γ and κ are the third and fourth moments. The mean measures the location of the distribution, standard deviation measures its spread, skewness measures the symmetry of the distribution and kurtosis measures the flatness or peakedness of the distribution. A Gaussian distribution is perfectly symmetric; therefore its skewness is zero. A positive skewness means the data shift to the left in comparison with a normal distribution, and vice versa. A positive Kurtosis indicates a "sharper" distribution than the normal distribution, and vice versa. Therefore, the third and fourth moments can be used to quantify the deviation of a non-Gaussian distribution from the corresponding Gaussian distribution with the same μ and σ.

Table 17 statistical distribution of the SET pulse width of NAND gate under different input

Input vector	mean	standard deviation	skewness	excess kurtosis
00	275.748	62.10	0.3139	-0.152141
01	366.088	65.09	0.3311	-0.156598
10	377.884	68.42	0.3893	0.097117
11	199.056	46.89	0.3197	-0.441802

A conclusion that can be drawn from the results is that the SET pulse width distribution varies significantly with different input vectors. The average value of pulse width distributions varies

from 199.056 ps under input vector “11”, to 377.884ps with input “10”. Also the spread (variation) increases by 50% from input “11” to “10”.

Similarly, the SET pulse-width distributions of a NOR gate for all input vectors are plotted in Fig. 25. The four moments of the data distribution are shown in Table 18. It also shows that the SET pulse width distributions do not follow a normal distribution, although the input parameters follow either normal distributions or uniform distributions. Interestingly, by comparing the statistical distributions of the NAND2 and NOR2 gate, the mean values of the SET pulse width for the NOR2 gate are larger but the spread is smaller than those of the NAND2 gate.

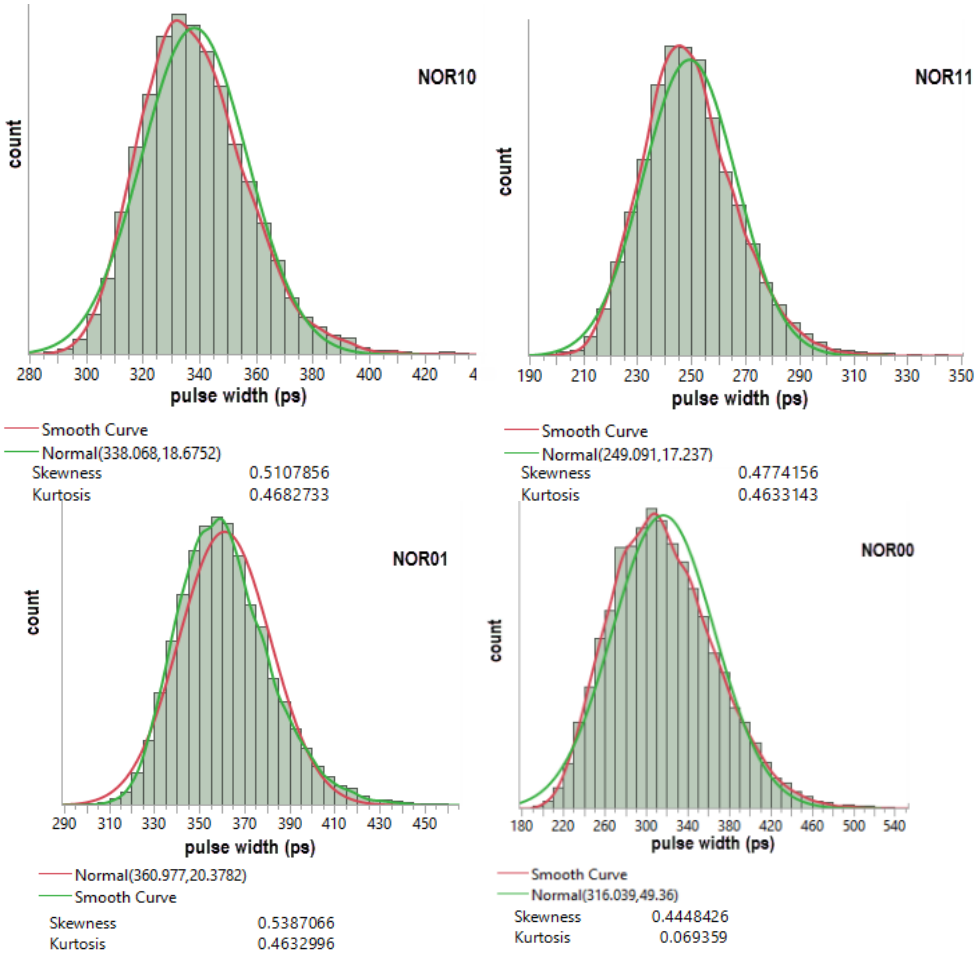

Fig. 25 The SET pulse width distribution of NOR2 gate under all input vectors

Table 18 statistical distribution of the SET pulse width of NOR gate under different input

Input vector	mean	standard deviation	skewness	excess Kurtosis
00	316.039	49.36	0.4448426	0.069359
01	360.977	20.38	0.5387066	0.463299
10	338.068	18.68	0.5107856	0.468273
11	249.091	17.24	0.4774156	0.463314

V.2 Evaluate the SET pulse width with RSM-DOE based model

V.2.1 Motivation for efficient estimation of the SET pulse width distribution

In order to obtain output single event pulse width distributions by transistor-level simulations, a large number (thousands to tens of thousands) of simulation iterations are required by the principles of MC analysis for high confidence [39]. Due to the large number of cells in standard cell libraries and long simulation times for advanced transistors with large numbers of input parameters, the necessity of thousands of simulation iterations results in a long simulation time for the characterizations of the SET pulse width distribution.

Additionally, due to the nature of semiconductor manufacturing processes, circuit operation environments, and circuit behavior, the PVT and other parameters do not necessarily follow Gaussian distributions. The non-linear relationship between pulse width and input parameters may also contribute to the non-Gaussian distribution of output pulses. However, SPICE-like

simulators such as Cadence Spectre are only compatible with Gaussian, lognormal and uniform input distributions by their embedded statistical functions.

In this section, aiming to efficiently estimate the statistical moments of the output pulse signal, we propose a fast method based on RSM with fractional factorial DOE technique. The proposed method requires significantly fewer simulations than the standard MC method to evaluate pulse width distribution for the standard logic cells. Furthermore, it can be used for any type of input parameter pdf irrespective of the simulation tools. Last but not least, the application of the method is not limited to 45 nm logic gate pulse width distribution estimation; it could be expanded for other technology nodes, or other kinds of output signal modeling.

The proposed method involves:

- Design of experiment technique to reduce the input parameter number so that the dimension of the input parameters is reduced.
- In contrast to sampling the input distribution randomly, RSM designs are used to select certain input parameter combinations to conduct simulation.
- The SET pulse widths are obtained by Spectre simulations.
- The coefficient matrixes of the response surface are calculated with the input parameters and the corresponding SET pulse widths.
- Using the derived symbolic RSM, the output distribution resulting from input parameter variations are calculated by numerical calculations, bypassing the corresponding transistor-level simulations.

V.2.2 An example of estimating the SET pulse width by RSM

Here, we use the NAND2 used in Chapter IV as an example to illustrate the flow of estimating the SET pulse width distributions. Assume that input parameters include:

- Process parameters (effective channel length L , threshold voltage V_{th})
- Design parameters (output capacitance, ratio of width to length, voltage supply)
- Environment parameters (Temperature, LET)

Table 19 Input parameters of the RSM

Parameters	Nominal Value	Variations	Distribution
leff (Effective channel length)	22.5nm	$\sigma = \pm 6\%$	Gaussian
V_{thn} (Threshold voltage of NMOS)	0.466V	$\sigma = \pm 14\%$	Gaussian
V_{thp} (Threshold voltage of PMOS)	-0.411V	$\sigma = \pm 14\%$	Gaussian
LET	8MeV/mg/cm^2	6-10MeV/mg/cm^2	Uniform
V_{dd} (Power supply)	1.0V	0.9-1.1V	Uniform
W/L	2 X minimum	1X – 3X minimum	Uniform
Cap (Load capacitance)	2 X minimum inv	1X – 3X minimum inv	Uniform
Temp (Temperature)	62.5 Celsius	25-100 Celsius	Uniform

In contrast to randomly sampling, selective input parameter combinations are sampled by design of experiment theory and RSM to ensure a second order response surface can be constructed with as few simulation runs as possible. Each simulation is considered as a discrete random event and the probability of each simulation event is equal to the joint probability of the input parameters. The illustration of the method is shown in Fig. 26. In this example, we consistently use NAND2 with input vector "10".

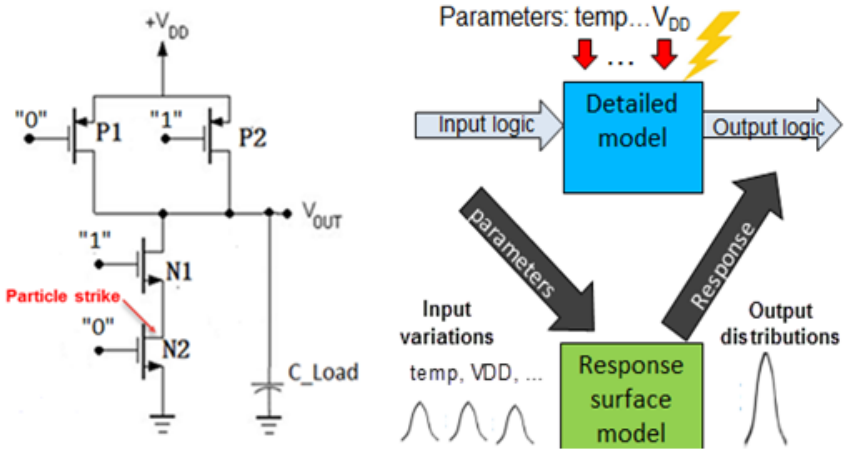

Fig. 26 Illustration of using RSM to estimate the SET pulse width distributions

The steps for estimating the distribution of the SET pulse width with selected input parameters are:

1. In order to build a model with as few simulations as possible, the first step is to use parameter screening to reduce input parameter dimensions. Using fractional factorial design of experiment technique, we can obtain required input parameter sampling points. By 64 Spectre simulations at the selected points, we can screen unimportant parameters (V_{thn} and loading capacitance in this case) to reduce the modeling building effort while

providing enough accuracy. The impacts of each input parameter on the output SET pulse width are ranked in Fig. 27.

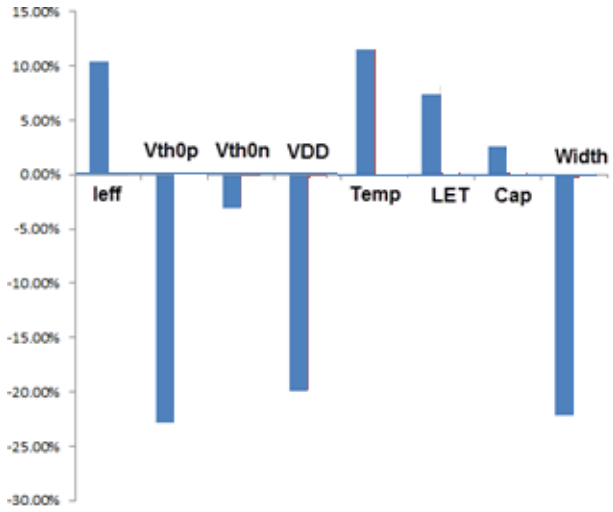

Fig. 27 Rank of the impact from relevant parameters

2. The second step is to add sampling points of the unscreened parameters by central composite design theory in order to describe the non-linear relationship between the input parameters and output pulse width. Then a quadratic symbolic closed-form equation is derived based on the sampled simulations and corresponding input parameters. The total required simulation runs are 46; moreover, some of the input parameter combinations are included in the FF designs used in step 1 and only 13 additional simulations are needed.
3. With these results, the correlation of the circuit output depends on simulation runs of the input parameter, which in turn depends on the each input parameter (x_i) sampled. After a quadratic symbolic multivariate is derived, anther 50 validation runs by random sampling are conducted to evaluate the accuracy of the second-order RSM. Here the error is limited

to 8%, which provides enough accuracy (if the model cannot provide enough accuracy, and then the non-linear model discussed in Chapter IV is needed). The second-order RSM is obtained here as

$$Pulse\ width = x_i^{\boldsymbol{T}} \boldsymbol{A} x_i + \boldsymbol{B}^{\boldsymbol{T}} x_i + \boldsymbol{C} \quad (20)$$

4. The derived symbolic mathematical equation is used to derive the output distribution with input variables. The symbolic numerical calculation from input variable distribution to output SET pulse width distribution is done by Mathematica 8.0 and Matlab2012B[40].

The statistical parameters (four moments) are listed in Table 20, in comparison with those obtained by the standard Monte Carlo method. Fig. 28 shows the pdf obtained by RSM. The results shows the estimated output SET pulse width distribution by the proposed RSM accurately approximates the results obtained using the standard Monte Carlo method.

Table 20 Estimated distribution parameters of the SET pulse width by standard MC and RSM

	Mean	Standard deviation	Skewness	Kurtosis
Standard MC	377.9	68.42	0.3889	3.0974
RSM	379.1	77.59	0.3861	2.9230
Error	0.3%	13.4%	0.72%	5.6%

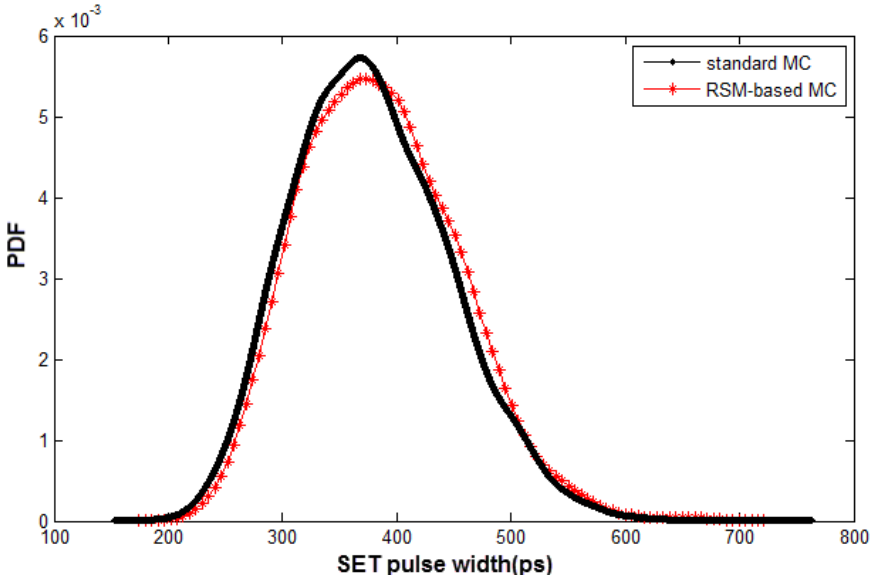

Fig. 28 PDF obtained by standard MC and RSM

The simulation expense to obtain the distribution is about 1/60 of the effort required for standard simulations. The computation time is listed in Table 21. The simulations are done on the Linux workstation of Vanderbilt ISDE and the modeling work is done on a PC with an Intel 1.7 GHz Core i5 CPU.

Table 21 Effort for obtaining the SET pulse width distributions

Method	Input combinations	Simulation time	Modeling time	Total
RSM method	127	3 min 40s	1 min 33s	5 min 3s
Standard Monte Carlo	10000	5 hours 13s	0	5 hours 13s

V.2.3 Discussion

The standard Monte Carlo simulation of the SET pulse widths is time-consuming because of the sampling strategy of the MC method. A large number (thousands) of samples are normally required to generate enough samples for accuracy concerns. This is especially critical in the case of characterizing SET pulse width distributions for the large number of standard cells fabricated in various technology nodes.

Monte Carlo analysis based on RSM is fast due to the fact that the generated model is analytical equations reproducing the simulation behavior. The application of the methodology is not limited to characterizing the SET pulse width for standard cells; it can also be used for large circuits (such as for oscillators) where large numbers of simulations are required by the standard Monte Carlo method.

A potential issue with the proposed method is that, when the input parameter dimension is very high, the number of screened parameters may be large, and this may bring in a relatively large modeling error to evaluate the output response. In addition, for highly non-linear RSM, the mathematical form of the model may be very complex and it requires a relatively large modeling time for the numeric calculations.

CHAPTER VI

SUMMARY

The impact on the SET pulse width of input parameters is investigated by transistor-level simulations with a 45 nm PTM model. A non-linear relationship of the SET pulse width to the input parameters is identified. This leads to a non-Gaussian or uniform distribution of the SET pulse width even if the input parameters are uniform or Gaussian distributed. Therefore, it is necessary to quantitatively describe the distribution by statistical modeling methods.

In order to predict the SET pulse width statistically with the minimum simulation cost, fractional factorial methods and response surface modeling techniques are proposed here to model the pulse width as a multi-variant function with multiple input parameters. The result shows that the method identifies important input parameters with only a fraction of the simulation cost, compared to the one-factor-at-a-time method. It also could be used as an efficient method for modeling the SET pulse width of other technology nodes in the future.

Based on the model combining RSM and DOE technique, a fast statistical moment estimation method is proposed, which shows the proposed method provides accurate results to predict the statistical distributions of the SET pulse width in a NAND2 gate. In comparison with the standard Monte Carlo simulation to estimate the SET pulse width distribution, the proposed method has three advantages: first, this method can use any kind of probability density function (pdf) irrespective of the simulation tools, and second, this method requires two orders of magnitude fewer simulation iterations and provides 60X speedups in the SET pulse width distribution characterization for standard library cells. Last but not least, the application of the

method is not limited to 45-nm logic gate pulse width distribution estimation; it could be expanded to other technology nodes, or other kinds of output signal modeling.

REFERENCES

[1] A. V Kauppila, S. Member, L. W. Massengill, S. Member, W. T. Holman, G. L. Vaughn, and J. S. Kauppila, "Frequency-Domain Analysis of Analog Based on Energy Spectral Density," vol. 51, no. 6, pp. 3537–3545, 2004.

[2] R. C. Baumann, "Radiation-Induced Soft Errors in Advanced Semiconductor Technologies," vol. 5, no. 3, pp. 305–316, 2005.

[3] J. Barth, "Space, atmospheric, and terrestrial radiation environments," *Nucl. Sci. IEEE Trans.*, 2003.

[4] R. A. Reed, M. A. Carts, P. W. Marshall, C. J. Marshall, S. Buchner, M. La Macchia, B. Mathes, and D. McMorrow, "Single Event Upset cross sections at various data rates," *IEEE Trans. Nucl. Sci.*, vol. 43, no. 6, pp. 2862–2867, 1996.

[5] L. W. Massengill, A. E. Baranski, D. O. Van Nort, J. Meng, and B. L. Bhuva, "Analysis of single-event effects in combinational logic-simulation of the AM2901 bitslice processor," *IEEE Trans. Nucl. Sci.*, vol. 47, no. 6, pp. 2609–2615, 2000.

[6] D. Mavis and P. Eaton, "Soft error rate mitigation techniques for modern microcircuits," *IEEE Int. Reliab. Phys. Symp.*, 2002.

[7] S. Buchner and M. Baze, "Single-event transients in fast electronic circuits," *IEEE NSREC Short Course*, 2001.

[8] http://www-device.eecs.berkeley.edu/~bsim3.

[9] http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4

[10] http://ptm.asu.edu/.

[11] J. S. Kauppila, A. L. Sternberg, M. L. Alles, a. M. Francis, J. Holmes, O. a. Amusan, and L. W. Massengill, "A Bias-Dependent Single-Event Compact Model Implemented Into BSIM4 and a 90 nm CMOS Process Design Kit," *IEEE Trans. Nucl. Sci.*, vol. 56, no. 6, pp. 3152–3157, Dec. 2009.

[12] M. J. Gadlage, J. R. Ahlbin, V. Ramachandran, P. Gouker, C. a. Dinkins, B. L. Bhuva, B. Narasimham, R. D. Schrimpf, M. W. McCurdy, M. L. Alles, R. a. Reed, M. H. Mendenhall, L. W. Massengill, R. L. Shuler, and D. McMorrow, "Temperature Dependence of Digital Single-Event Transients in Bulk and Fully-Depleted SOI Technologies," *IEEE Trans. Nucl. Sci.*, vol. 56, no. 6, pp. 3115–3121, Dec. 2009.

[13] R. Naseer, J. Draper, Y. Boulghassoul, M. Del Rey, S. Dasgupta, and A. Witulski, "Critical Charge and SET Pulse Widths for Combinational Logic in Commercial 90nm CMOS Technology," pp. 227–230, 2007.

[14] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, "Path-Based Statistical Timing Analysis Considering Inter- and Intra-Die Correlations," pp. 1–6, 2002.

[15] R. Rao and A. Srivastava, "Statistical estimation of leakage current considering inter-and intra-die process variation," *ISLPED*, 2003.

[16] H.-K. Peng, H.-M. Huang, Y.-H. Kuo, and C. H.-P. Wen, "Statistical Soft Error Rate (SSER) Analysis for Scaled CMOS Designs," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 17, no. 1, pp. 1–24, Jan. 2012.

[17] P. D. Friedberg, "Spatial Modeling of Gate Length Variation for by," 2007.

[18] M. Gadlage, "Impact of Temperature on Single-Event Transients in Deep Submicrometer Bulk and Silicon-On-Insulator Digital CMOS Technologies," 2010.

[19] S. DasGupta and A. Witulski, "Effect of well and substrate potential modulation on single event pulse shape in deep submicron CMOS," *Nucl. Sci.* , 2007.

[20] J. Kleijnen, "State-of-the-art review: a user's guide to the brave new world of designing simulation experiments," *INFORMS J.*, 2005.

[21] M. Gaudard, P. Ramsey, and M. Stephens, "Interactive Data Mining and Design of Experiments : the JMP® Partition and Custom Design Platforms," 2006.

[22] B. Jones and C. J. Nachtsheim, "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects," vol. 43, no. 1, pp. 1–15, 2011.

[23] A. V Kauppila, S. Member, B. L. Bhuva, S. Member, L. W. Massengill, W. T. Holman, and D. R. Ball, "Impact of Process Variations and Charge Sharing on the Single-Event-Upset Response of Flip-Flops," vol. 58, no. 6, pp. 2658–2663, 2011.

[24] N. V. Queipo, C. J. Arévalo, and S. Pintos, "The integration of design of experiments, surrogate modeling and optimization for thermoscience research," *Eng. Comput.*, vol. 20, no. 4, pp. 309–315, Oct. 2004.

[25] J. P. C. Kleijnen, "EXPERIMENTAL DESIGNS FOR SENSITIVITY ANALYSIS OF SIMULATION MODELS By Jack P . C . Kleijnen February 2001 ISSN 0924-7815 SIMULATION MODELS," 2001.

[26] L. E. Grade and D. E. D. En, “Application of Design-of-Experiment Methods and Surrogate Models in Electromagnetic Nondestructive Evaluation,” 2011.

[27] V. Fedorov, “Optimal experimental design,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 26, p. 055008, 2010.

[28] V. Janakiraman, A. Bharadwaj, and V. Visvanathan, “Neural Networks,” vol. 29, no. 7, pp. 1056–1069, 2010.

[29] V. C. P. Chen, R. R. Barton, and J. K. Allen, “A review of design and modeling in computer experiments,” vol. 22, no. 2003, pp. 231–261.

[30] D. Gorissen, L. De Tommasi, W. Hendrickx, J. Croon, and T. Dhaene, “RF circuit block modeling via Kriging surrogates,” no. 1.

[31] B. Ankenman, B. L. Nelson, and J. Staum, “Stochastic Kriging for Simulation Metamodeling,” *Oper. Res.*, vol. 58, no. 2, pp. 371–382, Dec. 2009.

[32] S. Mangano, *Mathematica cookbook*, vol. 129. 2010, p. 2865.

[33] A. Dharchoudhury, S. Member, and S. M. Kang, “Worst-Case Analysis and Optimization of VLSI Circuit Performances.”

[34] R. Garg, C. Nagpal, and S. P. Khatri, “A fast, analytical estimator for the SEU-induced pulse width in combinational designs,” *Proc. 45th Annu. Conf. Des. Autom. - DAC ’08*, p. 918, 2008.

[35] M. Merrett, P. Asenov, Y. Wang, M. Zwolinski, D. Reid, C. Millar, S. Roy, Z. Liu, S. Furber, and A. Asenov, “Modelling Circuit Performance Variations due to Statistical Variability : Monte Carlo Static Timing Analysis.”

[36] X. Li, J. Le, and L. T. Pileggi, *Statistical Performance Modeling and Optimization*. Now Publishers Inc, 2007, p. 160.

[37] D. O. Riordan, “Recommended Spectre Monte Carlo Modeling Methodology,” pp. 1–12, 2006.

[38] J. Yao, S. Member, Z. Ye, M. Li, Y. Li, R. D. Schrimpf, D. M. Fleetwood, and Y. Wang, “Statistical Analysis of Soft Error Rate in Digital Logic Design Including Process Variations,” vol. 59, no. 6, pp. 2811–2817, 2012.

[39] D. Basu, J. Guha, P. Hatab, P. Vaidyanathan, C. Mouli, and S. K. Groothuis, “Using TCAD, Response Surface Model and Monte Carlo Methods to Model Processes and Reduce Device Variation,” *2009 Int. Conf. Simul. Semicond. Process. Devices*, pp. 1–4, Sep. 2009.

[40] S. R. Otto and J. P. Denier, “An Introduction to Programming and Numerical Methods in MATLAB,” *Sci. York*, vol. 3, pp. 27–33, 2005.