Master's Thesis

A Monte-Carlo Analysis of Random CMOS and Dual-Rail PLA for Sub-100nm Parameter Variations

（100nm以下でのパラメーターのばらつきに関するランダムCMOSと２線式PLAのモンテカルロ解析）

Zhicheng Liang

Student Number: 56473

Supervisor: Professor Kunihiro Asada

DEPARTMENT OF ELECTRONIC ENGINEERING
THE UNIVERSITY OF TOKYO

Summited on 2nd February, 2007

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 The Process Parameter Variation Challenge

As process geometries continue to shrink to the sub-100nm regime, increasing process parameter variation has been perceived as one of the major roadblocks faced by the semiconductor industry, which is partly shown in Figure 1.1 [1]. The ever increasing process parameter variation is one of the hot topics of recent semiconductor international conferences such as Symposium on VLSI Technology/Circuits 2006 etc (VLSI: Very-Large-Scale Integration). From 65nm generation, process parameter variations are so severe that only by process control during fabrication is no longer enough but requires designs with consideration to alleviate the impact of variations. When the semiconductor industry such as IBM, Texas Instruments etc. develops the next sub-65nm generation, main challenges already change from low power and noise reduction to process parameter variations correspondence.

1.1.2 Random and Regular Structure Implementation

Implementation methodology affects process parameter variations in sub-100nm regime. Generally, digital circuit implementation approaches can be divided into the random structure and the regular structure. The static CMOS standard-cell approach is an ad hoc implementation to lay out logic circuits and thus is with random structure (CMOS: Complementary Metal-Oxide-Semiconductor). We use the term random CMOS to refer to static CMOS standard-cell approach. On the other hand, array-based approaches, such as the Programmable Logic Array (PLA) [2]-[5] employ regular fabric to lay out logic circuits and thus is with regular structure. Process parameter variations of these two design approaches

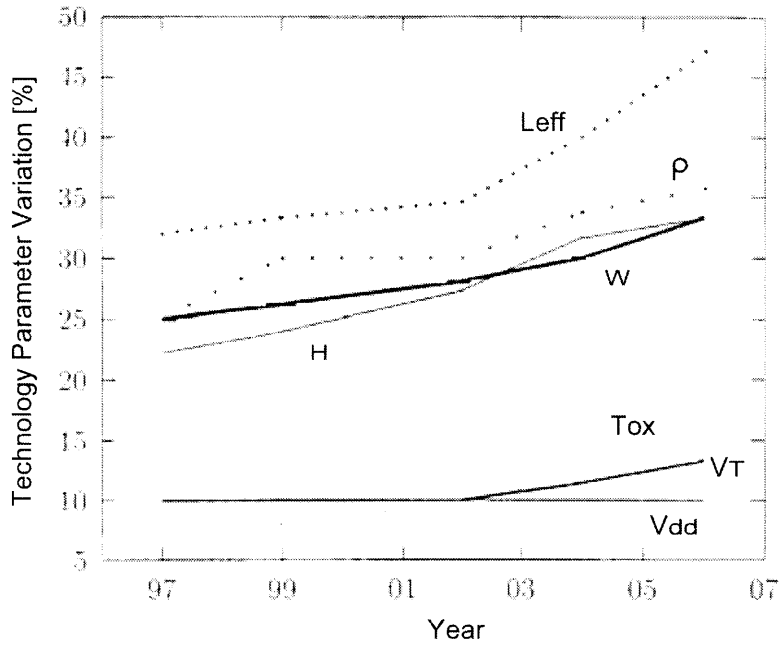


Figure 1.1 The ever increasing technology parameter variations [1] (L_{eff}: effective transistor length, V_{dd}: operating voltage, T_{ox}: oxide thickness, V_T: threshold voltage, ρ: wire resistivity, W:wire width, H: wire-to-base distance).

tend to be different, which will be analyzed in detail in Chapter 2. Compared with the random CMOS, the PLA is supposed to be able to alleviate the impact of process parameter variations on integrated circuits and with more manufacturability.

1.2 Research Objectives and the Thesis Organization

Since process parameter variation is an important consideration in the design of integrated circuits in sub-100nm technology, it is necessary to analyze, predict and alleviate the impact of these process parameter variations on integrated circuits before future generations of MOSFET technology are fully developed (MOSFET: Metal-Oxide-Semiconductor Field-Effect Transistor). In this thesis, we describe an approach to estimate the impact of process parameter variations on two design styles: the random CMOS and the PLA [2]-[5]. Our approach is built on accurate variation modeling, published data including the International Technology Roadmap for Semiconductors (ITRS) [6], Predictive Technology Models (PTM) [7][8], and Monte Carlo analysis to estimate and predict the trend of noise margins of the two design styles along with technology scaling.

This thesis is organized as follows. In Chapter 2, process parameter variations are analyzed

in detail. In Chapter 3 and 4, the random CMOS and the dual-rail PLA are respectively analyzed with respect to process parameter variations. In Chapter 5, techniques to improve the noise margin of the dual-rail PLA are proposed. In Chapter 6, speed and energy dissipation of the random CMOS and the dual-rail PLA with proposed techniques are analyzed. Chapter 7 concludes the thesis.

Chapter 2

Variation Models

2.1 Variation Decomposition

Process parameter variations depend on different scales in time and space. In general, these variations can be classified as lot-lot (ΔX_{L2L}), wafer-wafer (ΔX_{W2W}), die-die (ΔX_{D2D}) and intra-die or called within-die (ΔX_{WID}) [9]-[12]. A model can be expressed as

$$\Delta X = \Delta X_{L2L} + \Delta X_{W2W} + \Delta X_{D2D} + \Delta X_{WID} \quad (2.1)$$

$$= \Delta X'_{D2D} + \Delta X_{WID-SYS-layout} + \Delta X_{WID-SYS-spatial} + \Delta X_{WID-RAN} \quad (2.2)$$

$\Delta X'_{D2D}$: includes lot-lot, wafer-wafer and die-die components
$\Delta X_{WID-SYS-layout}$: within die, systematic layout pattern-dependent component
$\Delta X_{WID-SYS-spatial}$: within die, systematic spatial component
$\Delta X_{WID-RAN}$: random component

The first three components in Eq.(2.1) affect all the devices on the same chip in the same way, e.g., making the transistor gate lengths of devices on the same chip all larger or all smaller. Therefore, they are summarized into one term - $\Delta X'_{D2D}$ in Eq. (2.2).

ΔX_{WID} in Eq.(2.1), on the other hand, may affect different devices differently on the same chip and they can be further divided into three components in Eq.(2.2). $\Delta X_{WID-SYS-layout}$ is the systematic layout pattern dependent variation which is mainly caused by the layout dependency of the process. $\Delta X_{WID-SYS-spatial}$ is the systematic spatial variation. It means that the variations of transistors lying near each other are more correlated, than the variations of transistors that are far apart. The last component $\Delta X_{WID-RAN}$, which is caused by process fluctuations, is totally random.

2.2 Characteristics Analysis of Random CMOS and PLA Variations

Table 2.1 Characteristics analysis of random CMOS and PLA variations. The difference lies in $\Delta X_{WID\text{-}SYS\text{-}layout}$. Early in the design stage, the layout of random CMOS is undecided. Hence, $\Delta X_{WID\text{-}SYS\text{-}layout}$ is assumed to be random and thus with spatial correlation ≈ 0. However, the layout of PLA is regular and decided; therefore $\Delta X_{WID\text{-}SYS\text{-}layout}$ can be assumed to be systematic, with spatial correlation ≈ 1 and with small magnitude [12][13].

	$\Delta X'_{D2D}$	$\Delta X_{WID\text{-}SYS\text{-}layout}$	$\Delta X_{WID\text{-}SYS\text{-}spatial}$	$\Delta X_{WID\text{-}RAN}$
random CMOS	systematic	random	systematic	random
spatial correlation within a chip	1	0	1	0
PLA	systematic	systematic	systematic	random
spatial correlation within a chip	1	1 (small)	1	0

The characteristics of random CMOS and PLA variations are summarized in Table 2.1. $\Delta X'_{D2D}$, $\Delta X_{WID\text{-}SYS\text{-}spatial}$ and $\Delta X_{WID\text{-}RAN}$ are substantially independent of the physical implementation of the integrated circuit. Hence, for both random CMOS and PLA, the characteristics of them are the same. $\Delta X'_{D2D}$ and $\Delta X_{WID\text{-}RAN}$ are components of variations with the spatial correlation of 1 and 0 within a chip, respectively. $\Delta X_{WID\text{-}SYS\text{-}spatial}$ is assumed to be systematic with spatial correlation=1 on small area in this study [9]-[11].

The difference between random CMOS and PLA regarding variations lies in $\Delta X_{WID\text{-}SYS\text{-}layout}$. Early in the design stage, the layout of random CMOS is undecided. Hence, $\Delta X_{WID\text{-}SYS\text{-}layout}$ is assumed to be random and thus with spatial correlation ≈ 0. However, the layout of PLA is regular and decided; therefore $\Delta X_{WID\text{-}SYS\text{-}layout}$ can be assumed to be systematic, with spatial correlation ≈ 1 and with small magnitude [12][13]. Figure 2.1 is an example of poly-gates. When poly-gates are placed randomly, width variations tend to be random and thus, relatively big. However, when they are placed regularly, width variations tend to be systematic and small.

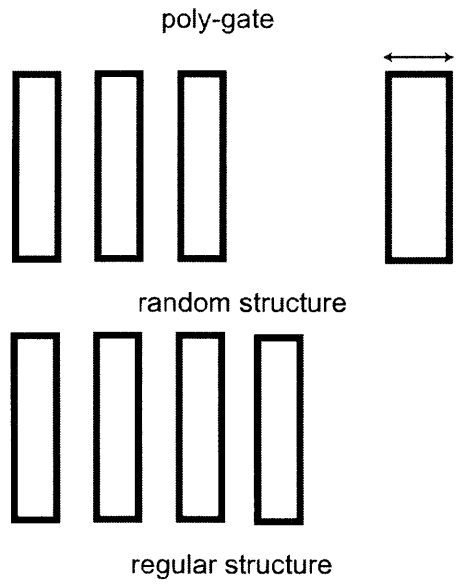


Figure 2.1 An example of poly-gates to analyze the $\Delta X_{WID-SYS-layout}$. When poly-gates are placed randomly, width variations tend to be random and thus, relatively big. However, when they are placed regularly, width variations tend to be systematic and small.

2.3 Sources of Variations Considered in This Study

In this thesis we consider process parameter variations in physical gate length (L_g) and width (W_g), effective gate length (L_{eff}), gate oxide thickness (T_{ox}), gate-source and gate-drain overlap length (L_{ovs}, L_{ovd}), mobility of carriers (μ) and threshold voltage (V_{th}). All the variations are assumed to follow the normal distribution and the standard deviations (σ) are listed in Table 2.2. 3σ of L_g, W_g, L_{eff}, T_{ox} are referred from ITRS (high performance) [6]. 3σ of L_{ovs} and L_{ovd} are calculated by Eq. (2.3). 3σ of μ is assumed to be 10% in this study. σ of each term in Eq. (2.2) can be expressed as Eq. (2.4) and are set as Table 2.3 [9]-[13]. Note that in Table 2.3, term B of the PLA is assumed to be 1/10 of the random CMOS [13].

$$\begin{aligned} L_{eff} &= L_g - L_{ovs} - L_{ovd} \\ \sigma_{L_{ovs,d}} &= \sqrt{\frac{\sigma_{L_g}^2 + \sigma_{L_{eff}}^2}{2}} \qquad (2.3) \\ \sigma_X^2 &= \sigma_{X'_{D2D}}^2 + \sigma_{X_{WID-SYS-layout}}^2 + \sigma_{X_{WID-SYS-spatial}}^2 + \sigma_{X_{WID-RAN}}^2 \qquad (2.4) \\ &= (A + B + C + D)\sigma_X^2 \end{aligned}$$

Moreover, variance of the gate-source and gate-drain overlap capacitance (C_{ovs}, C_{ovd}) is estimated to be the combined variance of T_{ox} and L_{ovs}, or T_{ox} and L_{ovd}, based on the parallel

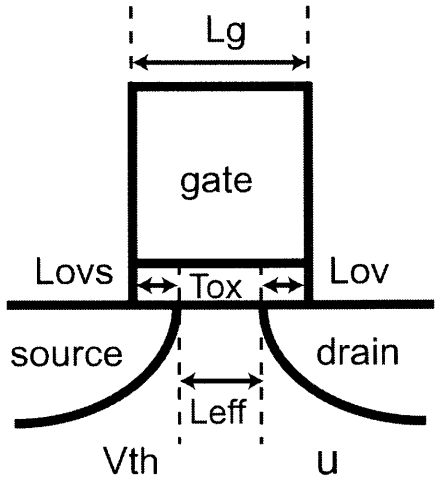


Figure 2.2 Parameters considered with respect to process variations. we consider physical gate length (L_g) and width (W_g), effective gate length (L_{eff}), gate oxide thickness (T_{ox}), gate-source and gate-drain overlap length (L_{ovs}, L_{ovd}), mobility of carriers (μ) and threshold voltage (V_{th}).

Table 2.2 3σ settings of parameter variations

Physical gate length (nm)	90	65	45	32
$3\sigma_{L_g}/L_g$, $3\sigma_{W_g}/L_g$	10%	10%	10%	12%
$3\sigma_{L_{eff}}/L_{eff}$	20%	20%	20%	20%
$3\sigma_{T_{ox}}/T_{ox}$	4%	4%	4%	4%
$3\sigma_{L_{ovs}}/L_{ovs}$, $3\sigma_{L_{ovd}}/L_{ovd}$	15%	15%	15%	21%
$3\sigma_{\mu}/\mu$	10%	10%	10%	10%

Table 2.3 σ^2 settings of each component in Eq.(2.4) [9]-[13] (A: $\frac{\sigma^2_{X'_{D2D}}}{\sigma^2_X}$, B: $\frac{\sigma^2_{X_{WID-SYS-layout}}}{\sigma^2_X}$, C: $\frac{\sigma^2_{X_{WID-SYS-spatial}}}{\sigma^2_X}$, D: $\frac{\sigma^2_{X_{WID-RAN}}}{\sigma^2_X}$)

Component	A	B	C	D
random CMOS	40%	30%	20%	10%
PLA	40%	3%	20%	10%

capacitor across the gate oxide as Eq. (2.5) .

$$\begin{aligned} C_{ovs,d} &= L_{ovs,d}\frac{\epsilon_{ox}}{T_{ox}} \\ \Delta C_{ovs,d} &= (L_{ovs,d} + \Delta L_{ovs,d})\frac{\epsilon_{ox}}{T_{ox} + \Delta T_{ox}} - L_{ovs,d}\frac{\epsilon_{ox}}{T_{ox}} \end{aligned} \tag{2.5}$$

Random Fluctuation of Dopants

In scaled CMOS devices, there exists a statistical fluctuation in the number of dopants, resulting in a threshold voltage variation (ΔV_{th}). This discrete dopant effect on ΔV_{th} is estimated as follows [14].

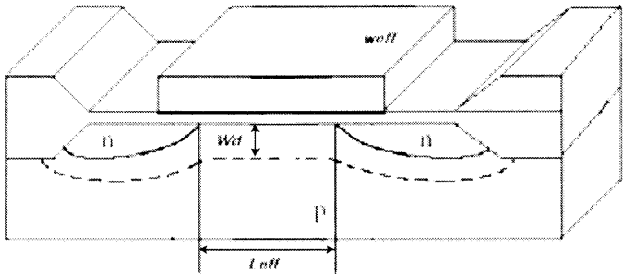


Figure 2.3 MOS structure to estimate the threshold voltage variation from a statistical fluctuation in the number of dopants.

For the transistor structure like Figure 2.3, the classical expression of the threshold voltage is

$$V_{th} = V_{fb} + 2\phi_f + \frac{\sqrt{2\epsilon_{si} q N_{ch} (|2\phi_f + V_{BS}|)}}{C_{ox}} \tag{2.6}$$

where V_{fb} is the flat band voltage, ϕ_f is the Fermi potential, ϵ_{si} is the silicon permittivity, q is the electron charge, N_{ch} is the channel doping concentration, V_{BS} is the body bias voltage, C_{ox} is the gate oxide capacitance per unit area, respectively.

The number of dopants in the depletion layer can be calculated by $N = N_{ch} L_{eff} W_{eff} W_d$. It follows the poison distribution, which can be approximated to the normal distribution. Hence $\sigma_N = \sqrt{N}$ can be derived. The maximum depletion layer thickness can be expressed by $W_d = \sqrt{2\epsilon_{si}(|2\phi_f + V_{BS}|)/qN_{ch}}$. Therefore the standard deviation ($\sigma_{V_{th}}$) of ΔV_{th} can be derived as

$$\begin{aligned} \sigma_{V_{th}} &= \frac{\partial V_{th}}{\partial N_{ch}} \sigma_{N_{ch}} \\ &= \frac{1}{C_{ox}} \sqrt{\frac{\epsilon_{si} q \phi_f}{N_{ch}}} \sqrt{\frac{N_{ch}}{L_{eff} W_{eff} W_d}} \\ &= \frac{1}{C_{ox}} \sqrt[4]{\frac{\epsilon_{si} q^3 N_{ch} (|2\phi_f + V_{BS}|)}{8}} \frac{1}{\sqrt{L_{eff} W_{eff}}} \end{aligned} \tag{2.7}$$

$V_{BS} = 0$ is assumed in the following analysis.

The variations of L_g, W_g, L_{eff}, T_{ox}, L_{ovs} and L_{ovd} follow Eq. (2.4) and the variations of μ and V_{th} are totally random in this study.

Based on the above variation modeling, we analyze the noise margins of the random CMOS and the dual-rail PLA through 1000 runs of Monte-Carlo simulation with SPICE, using the PTM from 90nm to 32nm technology [7][8] in the following chapters. Vdd is set from 1.2v to 0.9v for the four processes, respectively. Vdd noise is set by $\sigma_{Vdd} = 25mV$ and $\sigma_{Vdd} = 50mV$, respectively. Based on the simulation results, we use 3σ and 6σ assurance for fault-free operations. 3σ assurance means out of 1000 gates 1 ~ 2 gates exceed the predetermined margin of safe operations, and 6σ assurance means out of 1 billion gates only 1 gate exceeds the predetermined margin of safe operations. Table 2.4 summarizes the major characteristics of the PTM [8] released on February 22, 2006.

Table 2.4 Summary of PTM [8]

Physical gate length (nm)	90	65	45	32
V_{dd} (V)	1.2	1.1	1	0.9
T_{ox} (nm)	1.4	1.2	1.1	1.0
L_{eff} (nm)	35	24.5	17.5	12.6
V_{th} (V)	0.263	0.258	0.257	0.242
R_{dsw} (Ω /μ m)	185	165	145	135
I_{on} (μ A/μ m)	1105	1180	1230	1260
I_{off} (nA/μ m)	80	150	220	450
CV/I (ps)	1.03	0.73	0.50	0.34

Chapter 3

The Variation Analysis of the Random CMOS

3.1 Introduction

There are numerous circuit styles to implement a given logic function. The common design metrics by which a gate is evaluated are area, speed, energy and power. Depending on the application, the emphasis will be on different metrics. In addition to these metrics, recently robustness to noise and reliability also have become an important concern because of the decreasing power supply voltage and the ever increasing levels of process parameter variations, as introduced in Chapter 1.

The most widely used logic style is the static CMOS, which is implemented randomly in standard-cell approach. The static CMOS style is really an extension of a static CMOS inverter to multiple inputs. Hence, the inverter is truly the nucleus. The electrical behavior of complex circuits such as logic gates, adders, multipliers, and microprocessors can be almost completely derived by extrapolating the results obtained for inverters. The analysis of an inverter can be extended to explain the behavior of more complex gates such as NAND (Not AND), NOR (Not OR) or XOR (Exclusive OR), which in turn form the building blocks for modules such as multipliers and processors [17].

In this thesis, we use the term random CMOS instead of static CMOS to refer to the ad hoc approach to lay out logic circuits, as a comparison to a regular structured design approach such as the Programmable Logic Array (PLA), which will be introduced in the next chapter.

In this chapter, in order to estimate how the increasing process parameter variations affect the robustness and performance of the random CMOS design methodology, we analyze the trend of the noise margin of a CMOS inverter (INV) along with scaling of the technology.

3.2 Noise Margin Definitions of Random CMOS

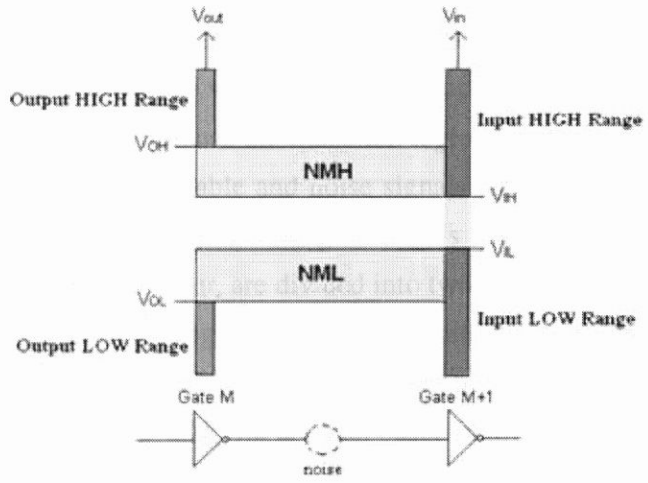


Figure 3.1 The definition of noise margins [15]

A noise margin is used to study the stability or vulnerability of a logic gate. It is a parameter closely related to the input-output voltage characteristics. This parameter permits one to determine the allowable noise voltage on the input of a gate so that the output will not be affected. The specification most commonly used to specify noise margin (or noise immunity) is in terms of two parameters - the LOW noise margin, NM_L, and the HIGH noise margin, NM_H. Figure 3.1 shows the basic concept of the noise margins [16]. Since the output of M gate acts as the input of gate M+1, when the output of gate M is in the low state, NM_L is defined as

$$NM_L = V_{IL} - V_{OL}$$

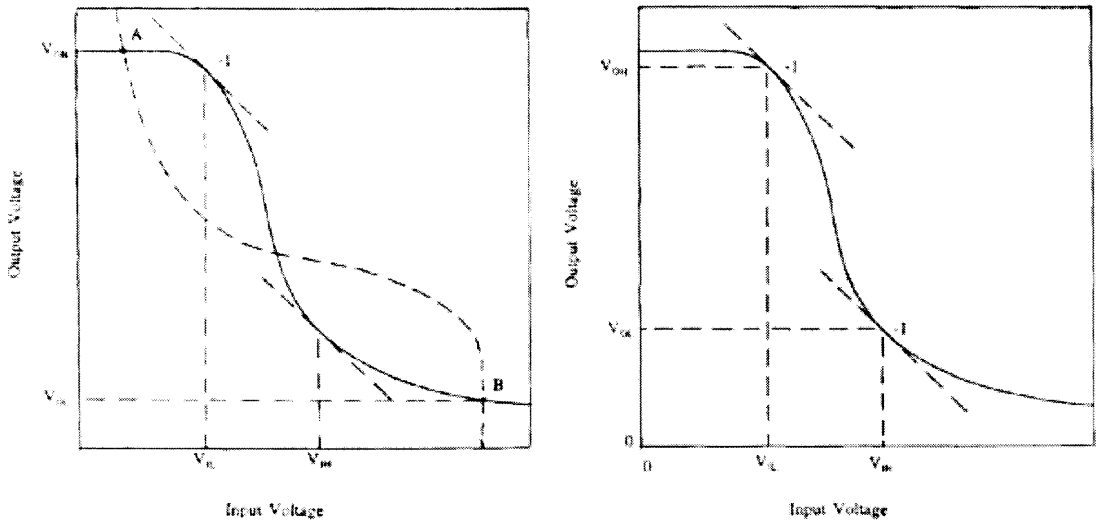
where V_{IL} is the maximum input voltage level that can be treated as the low level by gate M+1, and V_{OL} is the maximum output low voltage level of gate M. The shaded area in NM_L describes the dc noise that can be sustained by the logic gate M+1 to keep its input low state. If the noise level exceeds NM_L, the low output of gate M may be treated as high input of gate M+1 and cause a malfunction. Similarly, when the output of gate M is in the high state, NM_H is defined as

$$NM_H = V_{OH} - V_{IH}$$

where V_{IH} is the minimum input high voltage and V_{OH} is the minimum output high voltage. NM_H describes the dc noise that can be tolerated by the gate to keep its high state.

3.2.1 The Unity Gain Definition of Noise Margins

In almost any digital design textbook [16][17], the unity gain definition is exclusively used as the definition of the noise margin of the random CMOS [15]. The transition points V_{IL} and V_{IH} are defined as the input voltage values where the slope of the voltage transfer curves is -1, as illustrated in Figure 3.2, because the transition region between V_{IL} and V_{IH} has a gain greater than 1 and, hence, is not stable and noise signals falling into that transition region will be amplified to pass through following gates. This tends to cause malfunction. The definitions for V_{OL} and V_{OH}, however, are divided into two camps. Some textbooks [16][17] define V_{OL} and V_{OH} as the stable logic states of an infinite chain of the same gates, which, in the case of random CMOS gates, are V_{DD} and 0, respectively, as shown in Figure 3.2(a). In others, the definitions for V_{OL} and V_{OH} are also based on the -1 slope points, as shown in Figure 3.2(b).



(a) (b)

Figure 3.2 Typical choices for defining V_{OL} and V_{OH} used in most logic circuit textbooks [18]. (a) Bistable logic state definition of an infinite chain of the same gates. (b) -1 slope definition

Because a negative noise margin can be relatively easily obtained from otherwise perfectly valid MOS inverter characteristics, several highly respected digital electronics textbooks [16][17] have abandoned the definition of Figure 3.2(b), but used the definition of Figure 3.2(a). Therefore, in this thesis, we use the definition of Figure 3.2(a), as illustrated in Figure 3.3 in more detail.

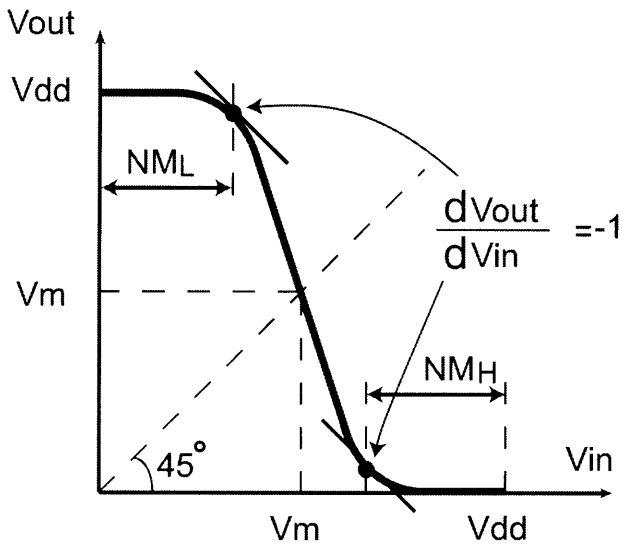


Figure 3.3 Unity gain definition of LOW noise margin (NM_L) and HIGH noise margin (NM_H) of an inverter on the voltage transfer curve. They are used for estimation in this thesis.

3.2.2 The Maximum Embedded Rectangle Definition of Noise Margins

As a comparison to the unity gain definition, the maximum embedded rectangle definition of noise margins will be introduced in this subsection. It is related to an infinite long chain of logic gates, and the noise margin of the gate in this long chain is the maximum noise magnitude before any malfunction occurs. Lohstroh et al. [19] discussed two kinds of noise: 1) noise that is only present somewhere far in the gate chain, as shown in Figure 3.4(a), and 2) noise that is present at all inputs in the chain with all the low-gate and high-gate noise sources contributing in such a way as to cause the maximum tendency to upset the logic levels. If the logic states of the gates in the chain are as shown in Figure 3.4(b), the worst-case noise is present to increase logic low state and decrease logic high state. The noise gives the maximum possibility to upset the logic states.

The best-case noise margin only describes the noise propagation characteristics. The worst-case noise margin is what people are concerned about. Further analysis shows that the infinitely long chain of inverters is equivalent to two cross-coupled inverter pair forming a bistable latch, as shown in Figure 3.5.

With the presence of noise signal δV, $V1'$ is not equal to $V2$, and $V2'$ is not equal to $V1$ anymore. The transfer curves of A1 and A2 shift according to the value of δV. The VTC of A1 shifts vertically, and the VTC of A2 shifts horizontally as shown in Figure 3.6(a). If the noise is so large that point X (stable point) and point Z (metastable point) coincide, as shown

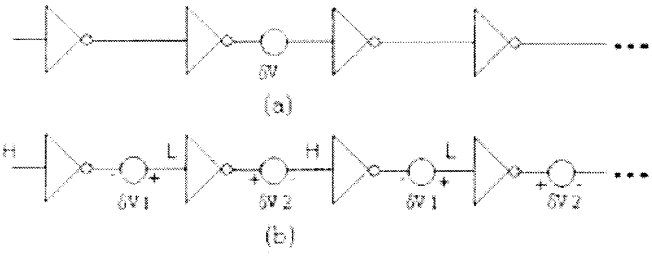


Figure 3.4 (a) Best-case noise and (b) Worst-case noise of an infinite long chain of inverters [15]

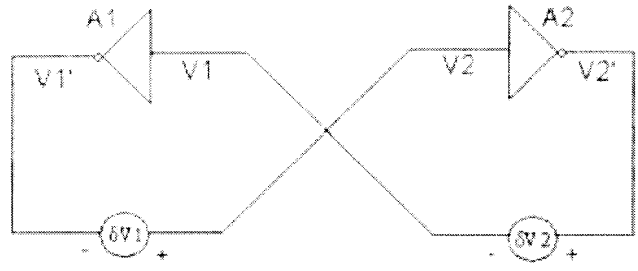
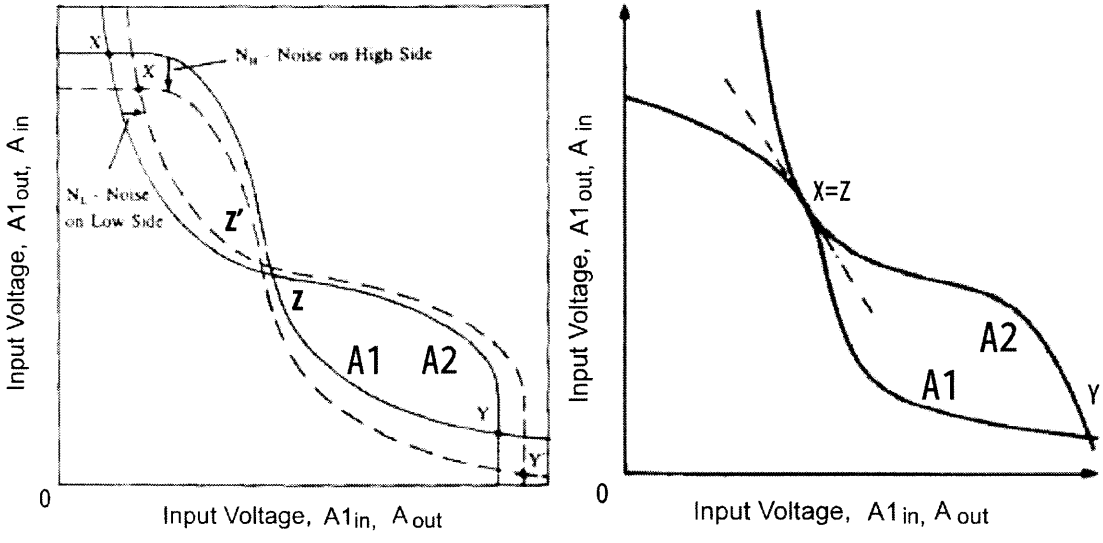


Figure 3.5 Equivalent circuit of Figure 3.4(b) [15]

in Figure 3.6(b), with a little more noise the VTC will have only one intersection (Y). Then the two cross-coupled inverter pair will switch to the wrong state.



(a) (b)

Figure 3.6 VTC of two cross-coupled inverter pair [18]. (a) with moderate worst-case noise applied and (b) with marginal worst-case noise applied

The inverter retains its state as long as the shifted curves continue to intersect in three

points [18]. Therefore, the valid maximum shift distance before the bistable point disappears represents the noise margin of the gate. Graphically, it is the largest rectangle that can be embedded into the VTC loop of the inverter pair, as illustrated in Figure 3.7. In this case, the area of the rectangle $NM_L \cdot NM_H$ is maximized. This method is called the maximum rectangle method (MRM) or maximum product criteria (MPC). If the equal noise margins ($NM_L = NM_H$)have to be maintained , the rectangle becomes a square, and the method is called the maximum square method (MSM), or maximum equal criteria (MEC). For VTCs that are nearly symmetrical, MPC and MEC give very close results. Usually, noise margins defined by MPC are smaller than those defined by the unity-gain definition.

In this thesis, as a comparison to the unity gain definition which will be used, we will also show the voltage transfer characteristic of a CMOS inverter.

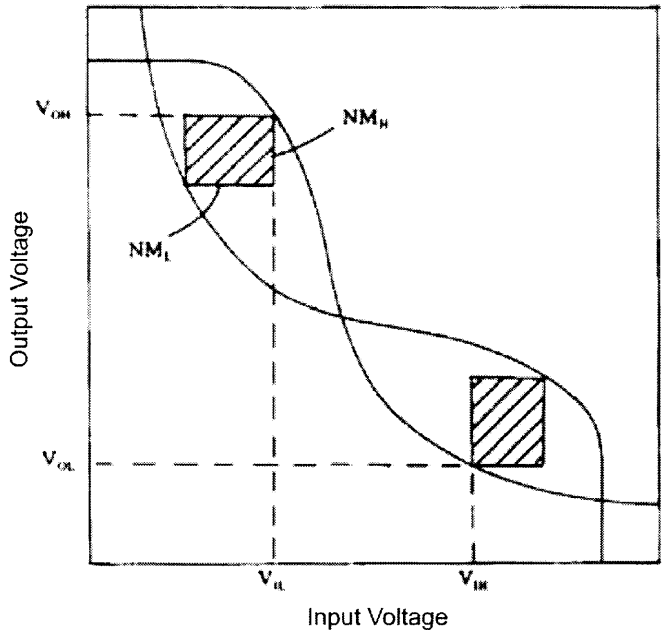


Figure 3.7 Noise margins for a given inverter characteristic, using an area embedded within the transfer curve loop [18].

3.3 Sizing of a CMOS Inverter

We construct a symmetrical inverter, which is an inverter where the NMOS (Negative-Channel Metal Oxide Semiconductor) and the PMOS (Positive-Channel Metal Oxide Semiconductor) are sized such that the switching threshold (V_m, it will be explained in the next paragraph) is equal to Vdd/2, to analyze the impact of the increasing process parameter vari-

ations.

$W_g/L_g = 4$ is chosen for the NMOS based on Figure 3.8, which indicates when $W_g/L_g > 2$, the normalized variation of the on current (Ion) of an NMOS approaches a stable value. After deciding the size of the NMOS, the size of the PMOS is chosen to make $V_m = Vdd/2$, where V_m is the switching threshold of an inverter. V_m is defined as the point where $V_{in} = V_{out}$. Its value can be obtained graphically from the intersection of the VTC with the line given by $V_{in} = V_{out}$ (refer to Figure 3.3). $V_m = Vdd/2$ is generally desired since this results in comparable values for the LOW and HIGH noise margin.

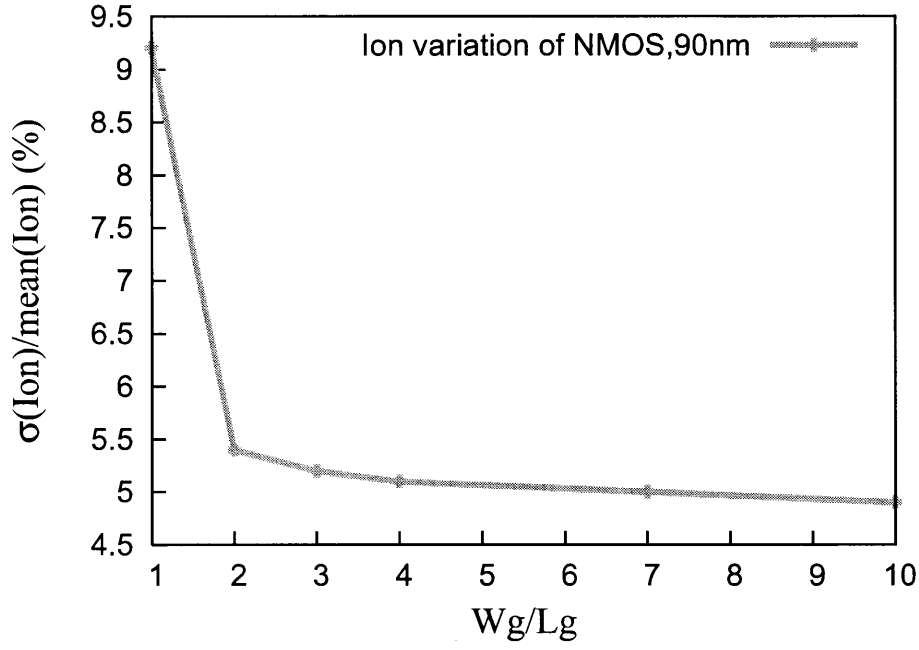


Figure 3.8 $W_g/L_g = 4$ is chosen for NMOS of the inverter because when $W_g/L_g > 2$, the normalized variation of Ion approaches a stable value.

3.4 Simulation Results

The size setting of the inverter and some simulation results without Vdd noise are listed in Table 3.1, from which we estimate the trend of the noise margin of the inverter quantitatively.

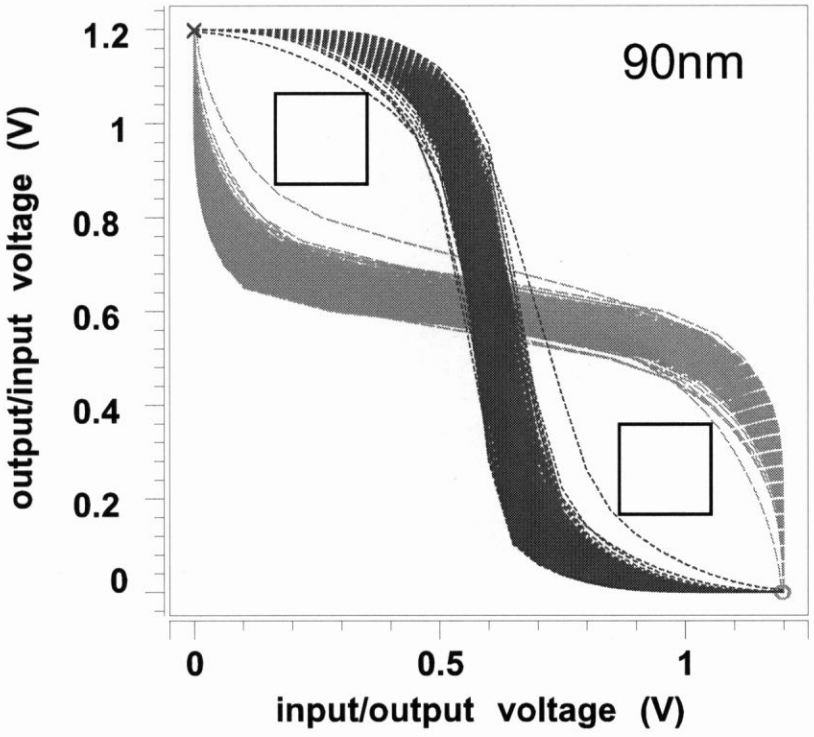
For comparison to the unity gain noise margin definition, simulation results are shown in Figure 3.9 in a form of voltage transfer characteristic. It indicates that a pair of maximum rectangles can no longer be embedded inside the loop when the process shrinks to 32nm.

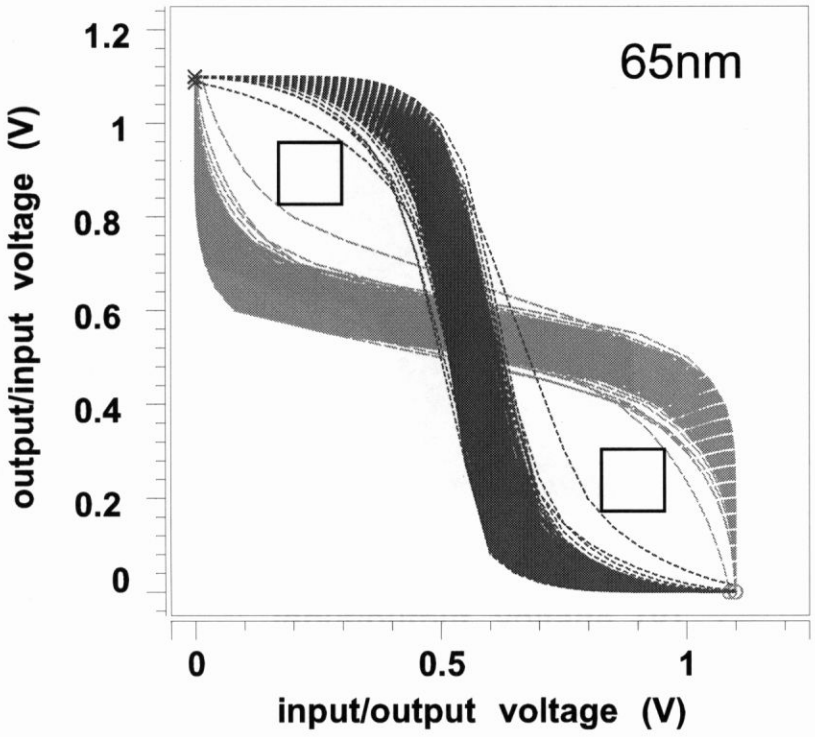
For unity gain definition of noise margins as described in Figure 3.3, noise margin profiles

Table 3.1 Size settings of the inverter and some simulation results without Vdd noise. $W_g/L_g = 4$ is chosen for the NMOS based on Figure 3.8. After deciding the size of the NMOS, the size of the PMOS is chosen to make $V_m = Vdd/2$, where V_m is the switching threshold of an inverter. V_m is defined as the point where $V_{in} = V_{out}$ (refer to Figure 3.3). $V_m = Vdd/2$ is generally desired since this results in comparable values for the LOW and HIGH noise margin.

Physical gate length (nm)	90	65	45	32
$W_{eff-p}(nm)$ ($W_{eff-p} = 11L_g - 10nm$)	962	692	512	336
$W_{eff-n}(nm)$ ($W_{eff-n} = 4L_g - 10nm$)	350	250	170	118
σ of V_{th-p} (mV)	5.5	7.6	10.6	15.5
σ of V_{th-n} (mV)	9.5	13	18.7	26.6
threshold voltage (Vm) mean (mV)	602	552	501	452
threshold voltage (Vm) σ (mV)	18	20	22	35
LOW noise margin mean (mV)	456	413	363	307
LOW noise margin σ (mV)	19	21	25	38
HIGH noise margin mean (mV)	481	431	376	311
HIGH noise margin σ (mV)	26	28	30	43

of the inverter are shown in Figure 3.10. They are histograms with an interval of 5mV. The trend of noise margins of the INV with and without Vdd noise is shown in Figure 3.11. It shows that there is no NM_H left for 6σ assurance for a CMOS INV at 32nm process when $\sigma_{Vdd} = 50mV$. Insufficient noise margins mean high possibility of malfunction and thus we predict the random CMOS can not work reliably at 32nm process. Hence, other circuit design styles that suppress the impact of process parameter variations should be investigated.


90nm
output/input voltage (V)
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
input/output voltage (V)


65nm
output/input voltage (V)
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
input/output voltage (V)

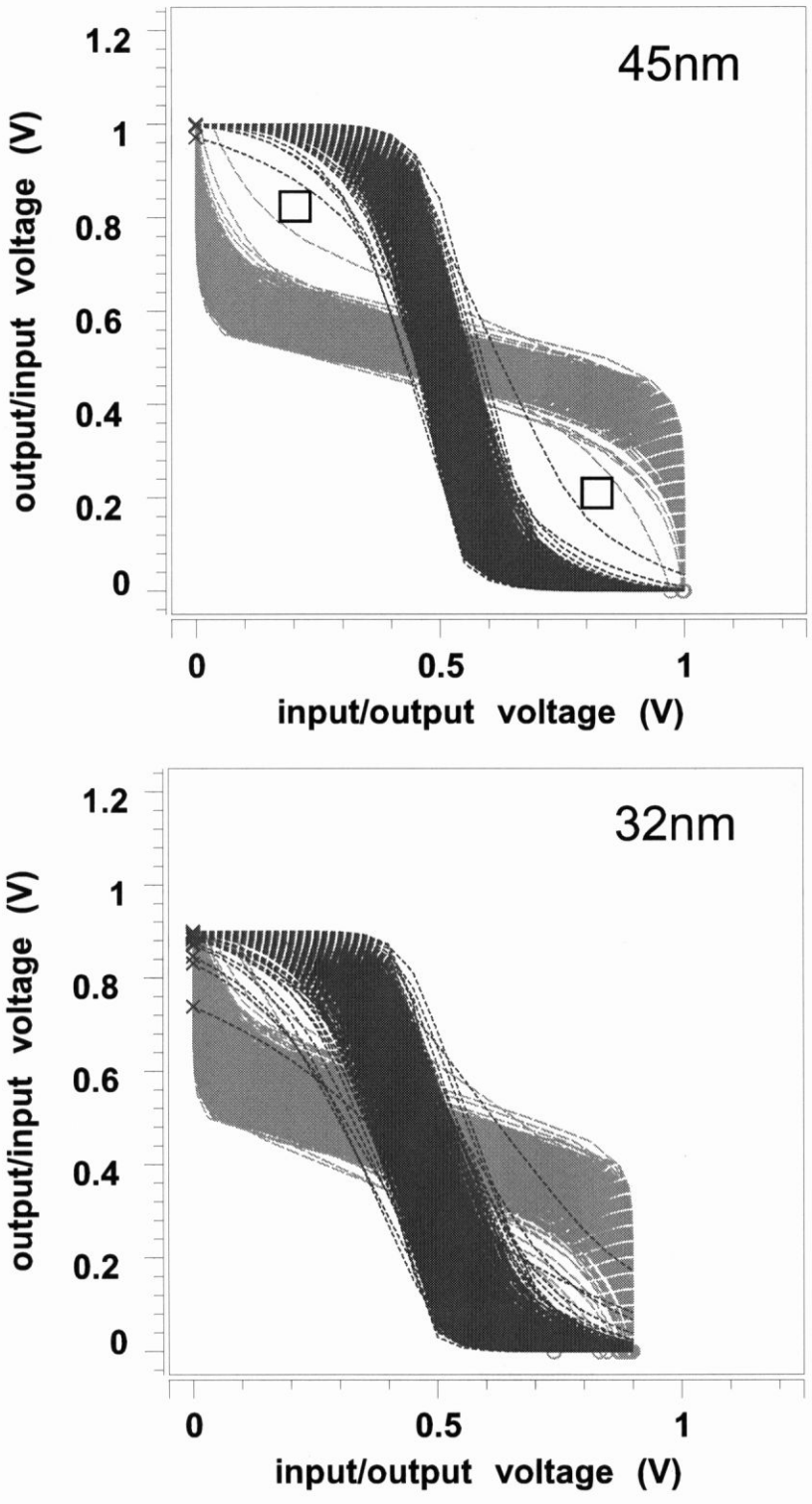
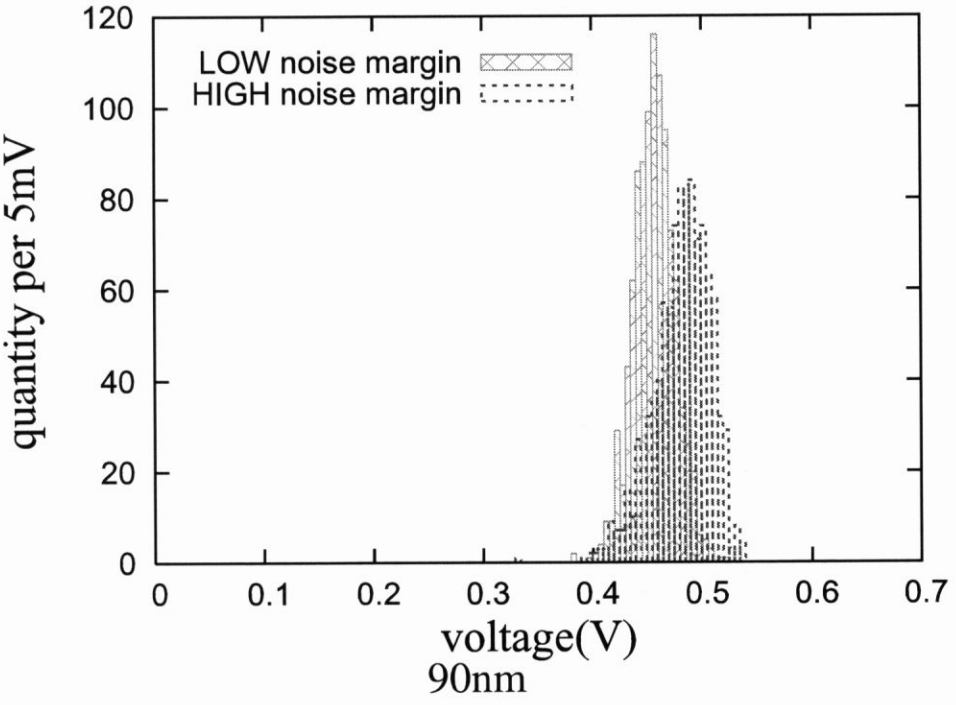
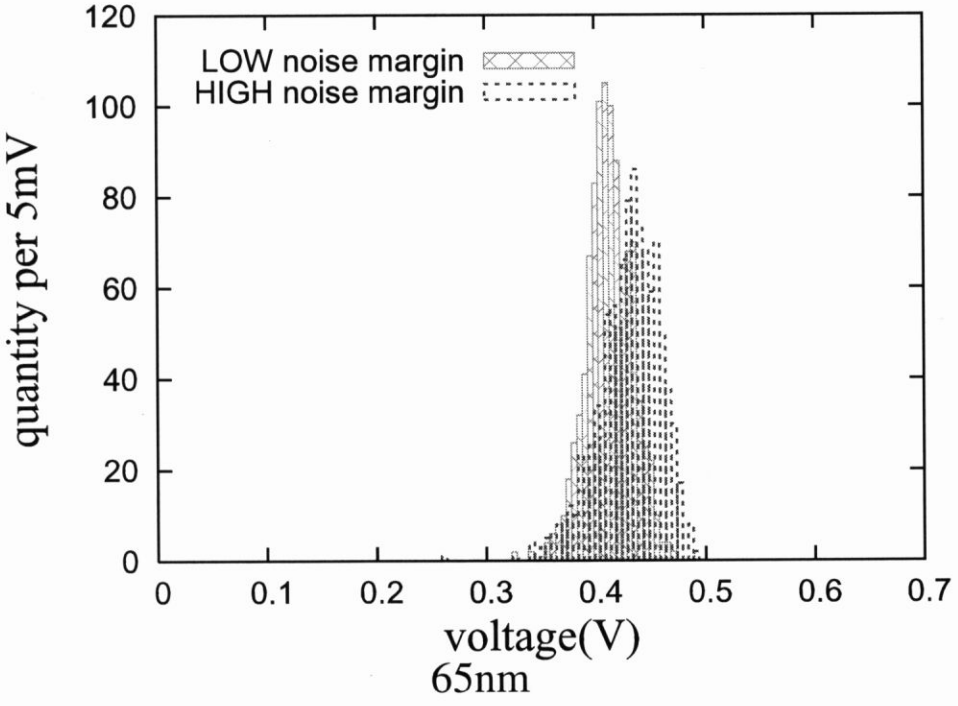


Figure 3.9 Voltage transfer characteristic of an inverter. A pair of maximum rectangles can no longer be embedded inside the loop when the process shrinks to 32nm.


LOW noise margin
HIGH noise margin
quantity per 5mV
voltage(V)
0
20
40
60
80
100
120
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

90nm


LOW noise margin
HIGH noise margin
quantity per 5mV
voltage(V)
0
20
40
60
80
100
120
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

65nm

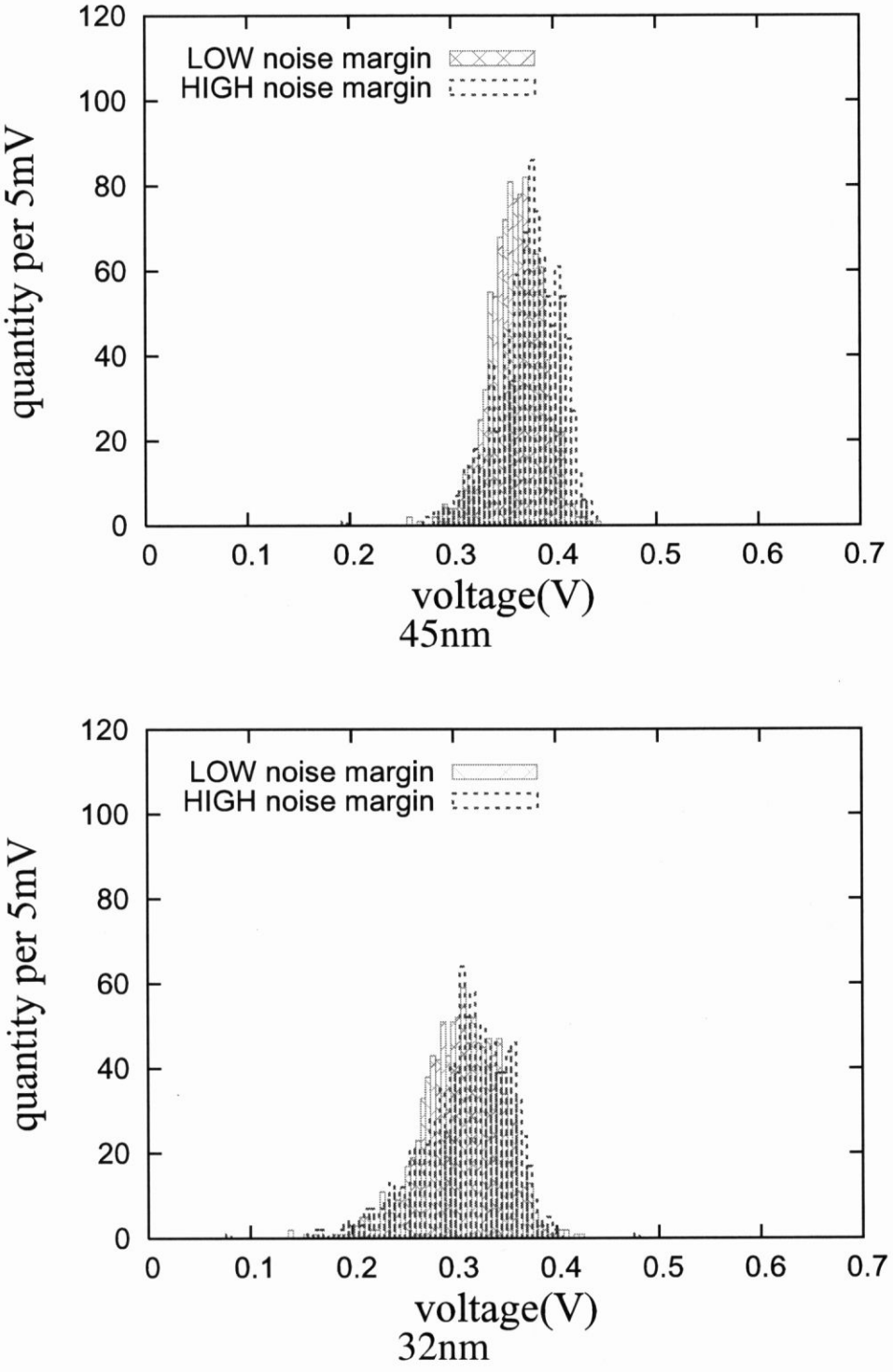


Figure 3.10 Noise margin profiles of an inverter defined in Figure 3.3 along with scaling. They are histograms with an interval of 5mV.

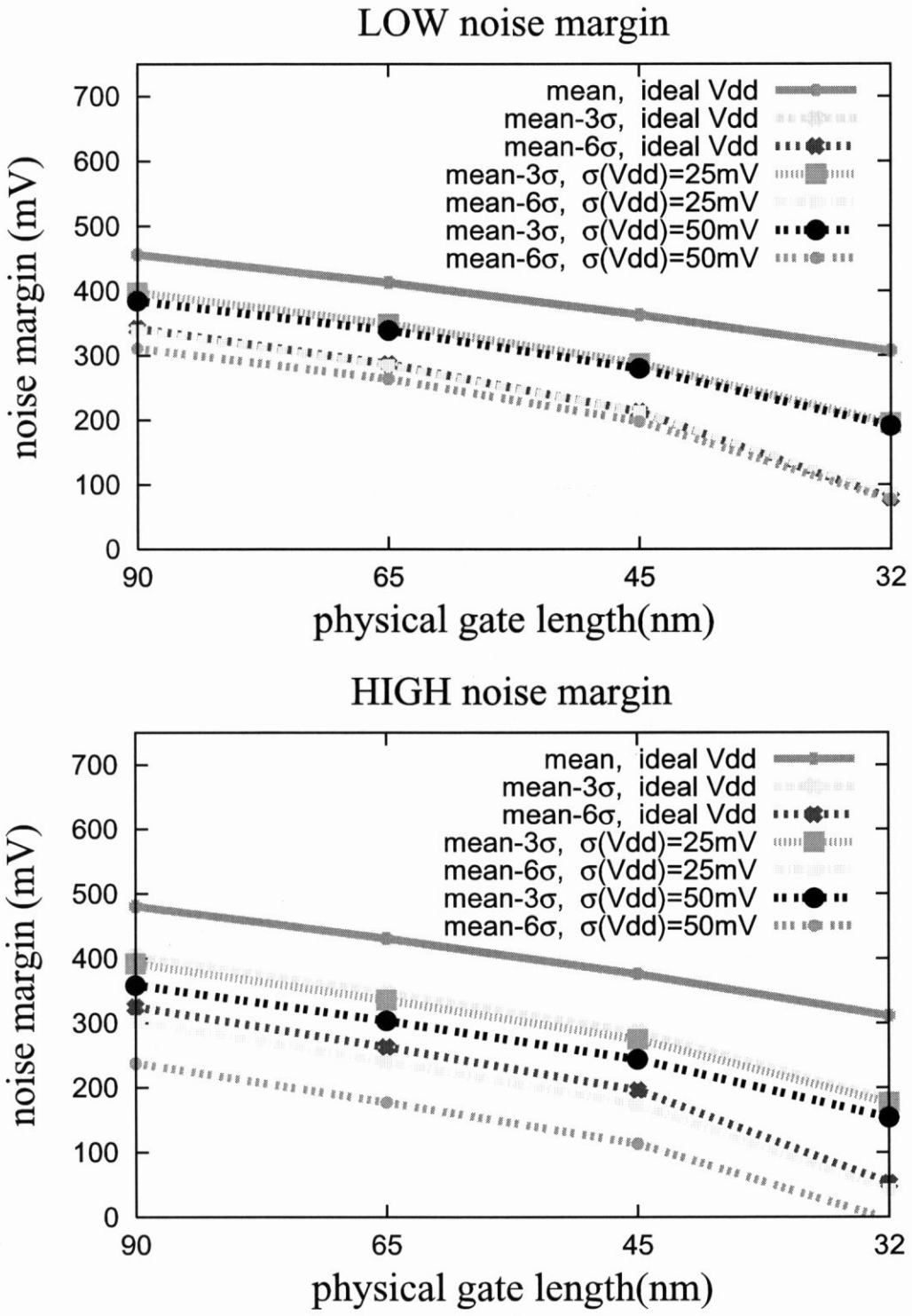


Figure 3.11 The trend of noise margins of an inverter according to the unity gain definition of Figure 3.3 along with scaling. When $\sigma_{Vdd} = 50mV$, there is no HIGH noise margin left for 6σ assurance for an inverter at 32nm. thus we predict the static CMOS can not work reliably at 32nm process.

3.5 Parameter Sensitivity Analysis of Random CMOS

Instead of varying all the parameters simultaneously, we vary one parameter for one time to estimate the sensitivity of each parameter to the random CMOS. Results are shown in Figure 3.12. It shows how sensitive the σ of noise margins of the random CMOS is to individual parameter. V_{dd}, L_g, L_{ovs} and L_{ovd}, V_{th} affect the σ of noise margins obviously.

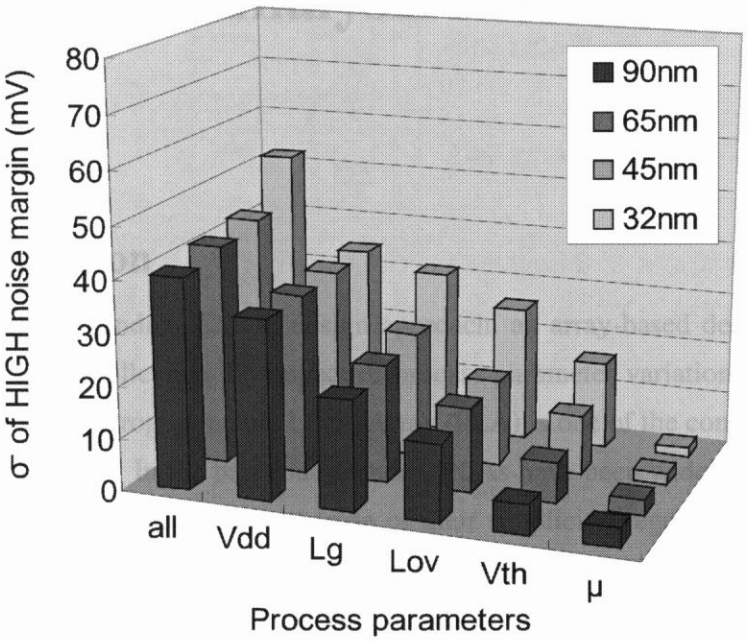


Figure 3.12 Parameter sensitivity analysis of the random CMOS. V_{dd}, L_g, L_{ovs} and L_{ovd}, V_{th} affect the σ of noise margins of the random CMOS obviously.

Chapter 4

The Variation Analysis of the Dual-Rail PLA

4.1 Introduction

Compared with the random CMOS design approach, an array-based design approach is supposed to be able to alleviate the impact of process parameter variations and with more manufacturability. The Programmable Logic Array (PLA) is one of the conventional CMOS array logic architectures. In the past four decades, PLAs have been widely used for combinational and sequential logic circuits because of their simplicity, regularity, and flexibility. However, PLAs generally require larger chip area than random logic implementations realized by standard-cell-based design. To overcome this drawback, lots of new PLA structures have been proposed [2]-[5]. A dynamic dual-rail PLA with latch sense amplifiers and logic cells [5] is one example. In this chapter, we analyze the robustness of this kind of PLA in future sub-100nm regime with respect to process parameter variations.

4.2 A Dynamic Dual-Rail PLA with Latch Sense Amplifiers

4.2.1 The Column Circuit

Figure 4.1 shows a column circuit of the dynamic dual-rail PLA with latch sense amplifiers (sense-amp). The circuit is a dual-rail configuration and consists of a stack of basic logic cells, a reference cell, a virtual ground (VG) controller, a pre-charge and equalization circuit, and a sense amplifier. Logical OR and NOR of the outputs of the basic logic cells can be obtained from the output signals, OUT and $\overline{OUT}$, respectively. A logical AND is also obtained by performing a logical NOR of complement input signals. Thus, an AND-plane and an OR-plane for a PLA can be realized by arranging the column circuits. By using a

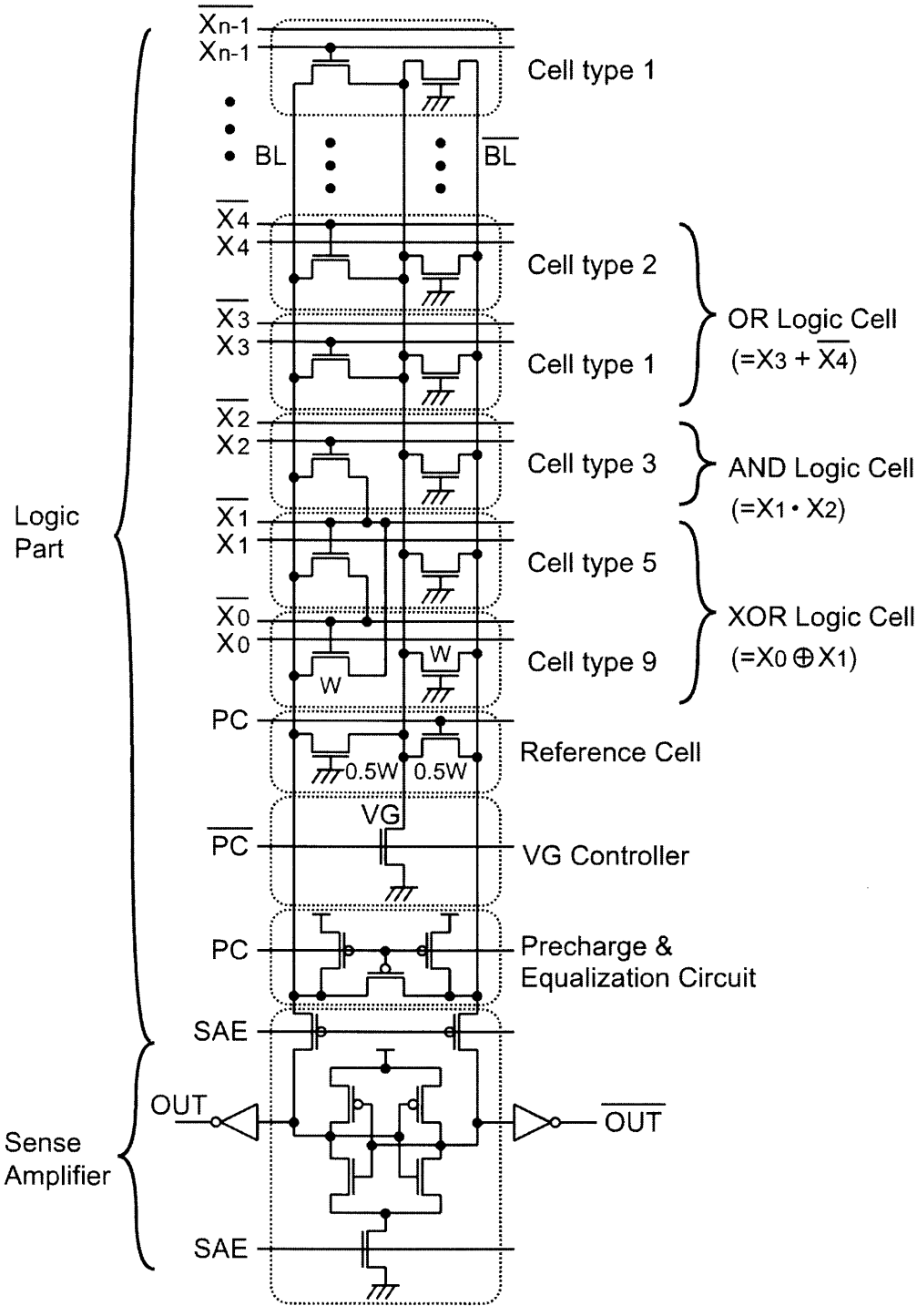


Figure 4.1 A column circuit of the dual-rail PLA [5]

sense amplifier, the output signals are activated by sensing the differential voltage between the bit-lines, BL and $\overline{BL}$. VG is provided to reduce the voltage swings of the bit-lines.

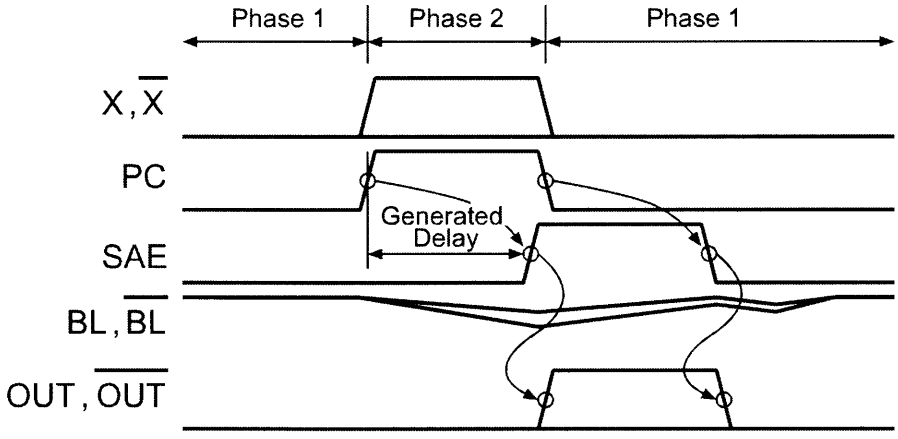


Figure 4.2 A timing diagram of signals [5]

Figure 4.2 shows a timing diagram of control signals. Input and output signals, and bit-line potentials in a column circuit. The circuit operates in two phases. In phase 1, the PC signal and all the input signals, ($X_0 - X_{n-1}$ and $\overline{X_0} - \overline{X_{n-1}}$,) are low. Thus, the bit-lines are pre-charged high and equalized. At the same time, the VG node is discharged low. When the PC signal becomes high and the input signals are activated, the circuit enters phase 2.

In phase 2, $\overline{BL}$ is pulled down by charge sharing with VG through a reference cell. When at least one of the basic logic cells pulls BL down, the voltage potential of BL becomes lower than that of $\overline{BL}$. Otherwise, BL stays high. This is because the device size of the basic logic cells is twice of the reference cell, as shown in Figure 4.1. This half-size device is provided to avoid the meta-stable condition, which may be caused when there is no pull-down path on BL. The SAE signal is activated when the developed voltage difference between the bit-lines becomes larger than the designed sense voltage, which takes the worst case of considerable noise margin and process parameter variations into account (0.35μm technology). By activating the sense amplifier, one of the output signals, i.e., OUT or $\overline{OUT}$ becomes high depending on the developed voltage difference. After the activation of the sense amplifier, the PC signal becomes low and the circuit starts to pre-charge the bit-lines.

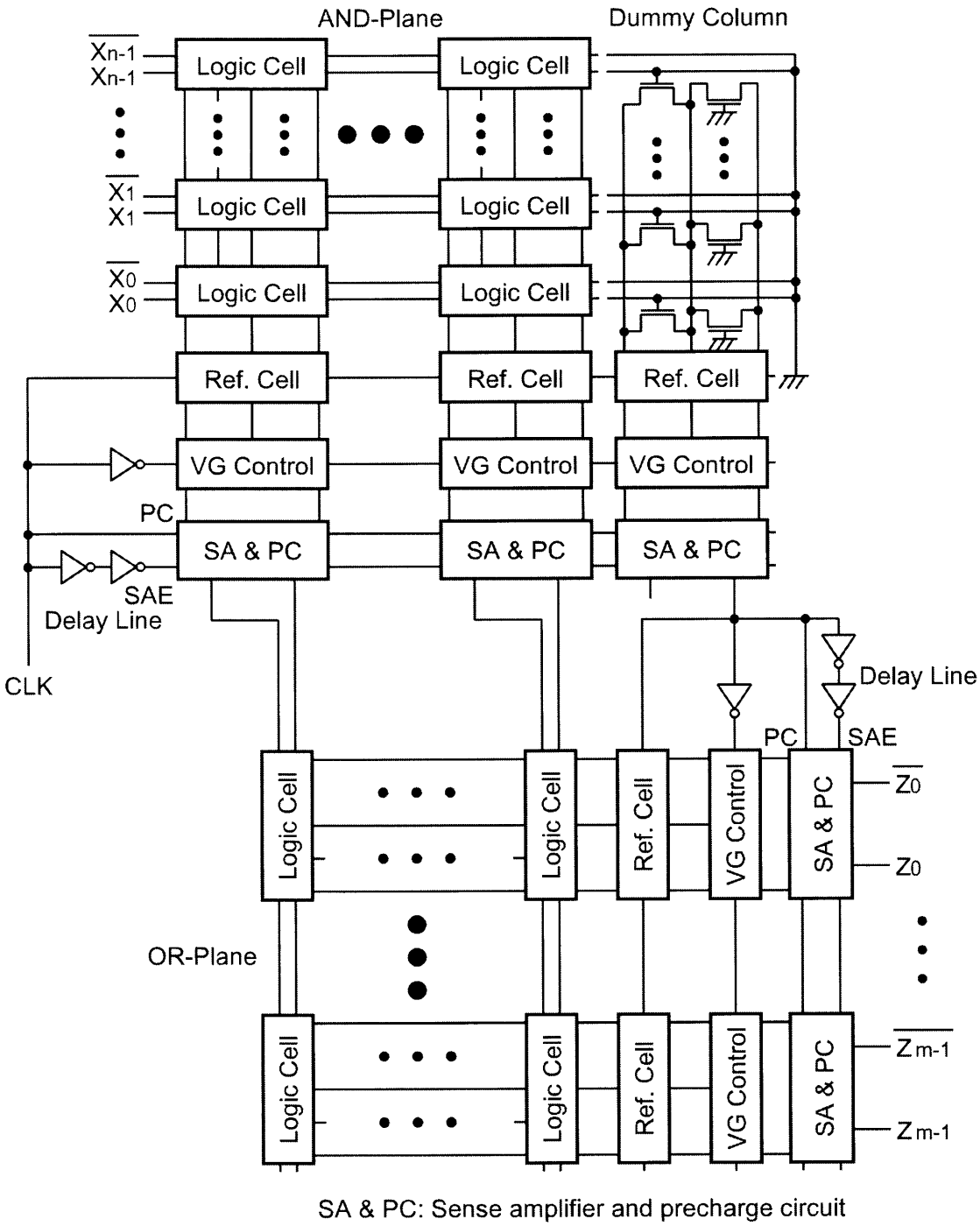


Figure 4.3 The overall configuration of the dual-rail PLA [5]

4.2.2 The Dual-Rail PLA Configuration

The overall configuration of the dual-rail PLA is shown in Figure 4.3. An array of column circuits is used as an AND-plane and an OR-plane. Control signals of the AND-plane are generated from the *CLK* signal with a delay line of a chain of sized inverters. On the other hand, control signals of the OR-plane are generated from a dummy column and a delay line of a chain of sized inverters. The dummy column is designed so that its output signal arrives last in the AND-plane. For this purpose, it has the largest number of basic logic cells in the AND-plane, and gate terminals of the basic logic cells are connected to ground. Its output signal, i.e., the *PC* signal of the OR-plane is generated every cycle and follows operating conditions, such as temperature and supply voltage variations, as well as process parameter variations in sync with column circuits in the AND-plane.

4.3 The Variation Analysis of the Dual-Rail PLA

4.3.1 The Input Offset Voltage Variation of the Sense Amplifier

We analyze the trend of noise margins of the described PLA above. We divide a column circuit of the dual-rail PLA into 2 parts - the logic part and the sense-amp as shown in Figure 4.1. First change the closed loop sense-amp to open loop as Figure 4.4 to estimate the input offset voltage (Voff) variations, which is plotted in Figure 4.5. Voff variations increase along with process scaling due to increasing process parameter variations.

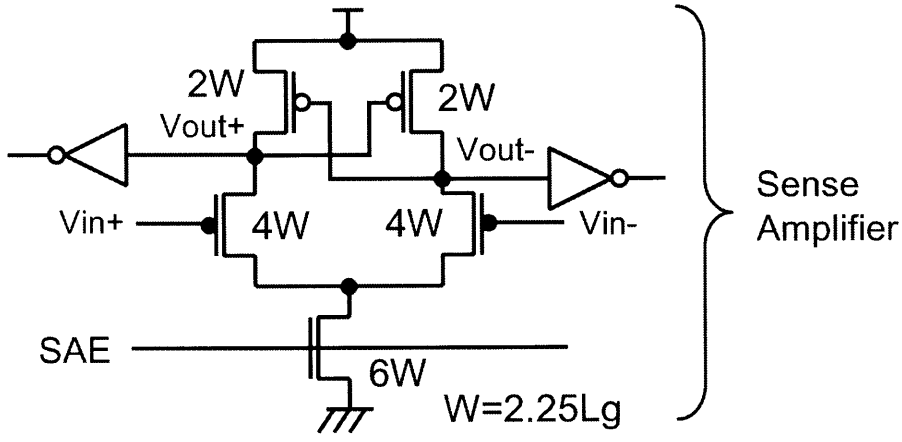
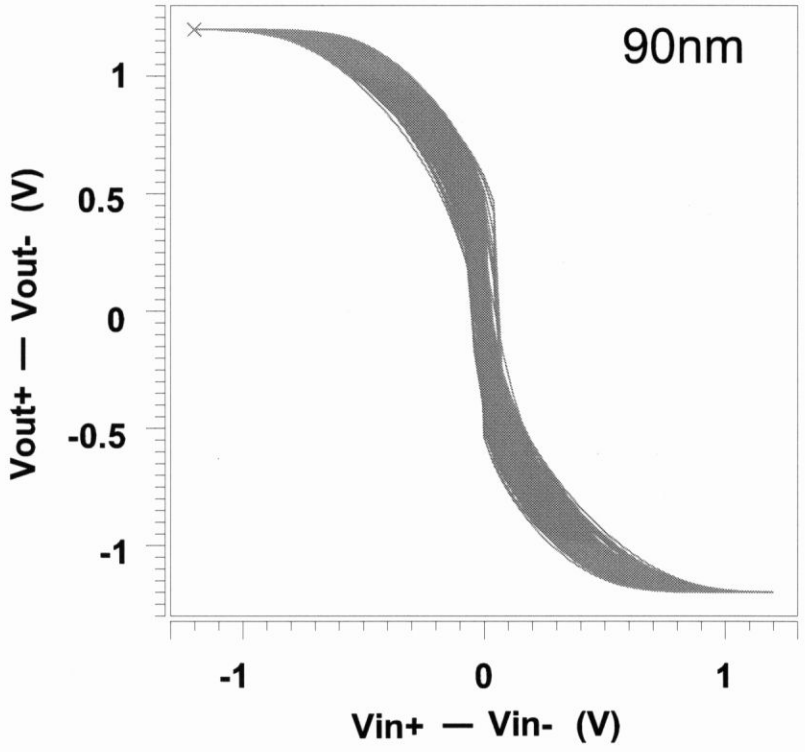
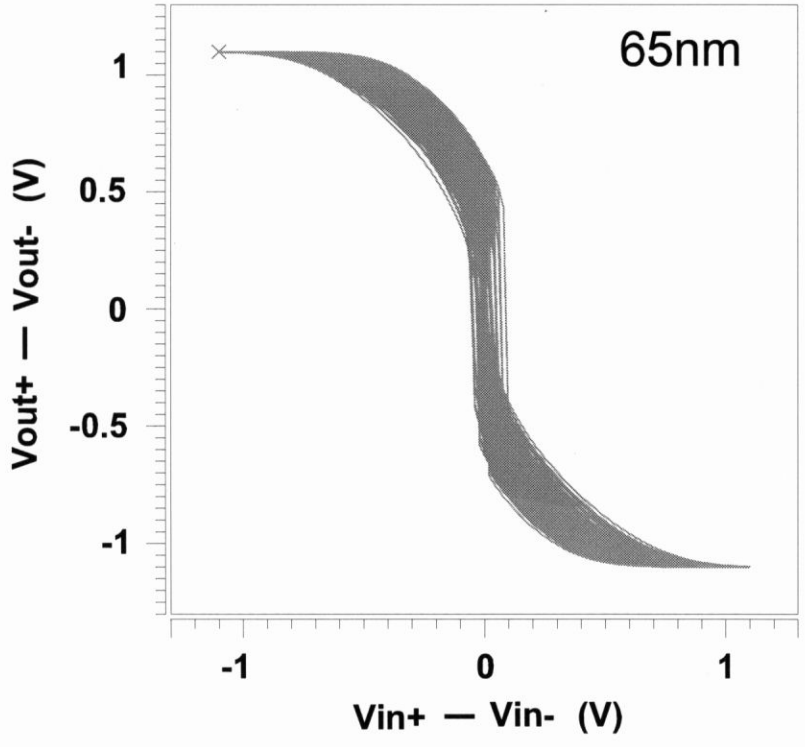


Figure 4.4 The closed loop sense amplifier used in Figure 4.1 is changed to open loop to estimate the input offset voltage variations along with process scaling.


90nm
1
0.5
0
-0.5
-1
Vout+ — Vout- (V)
-1
0
1
Vin+ — Vin- (V)


65nm
1
0.5
0
-0.5
-1
Vout+ — Vout- (V)
-1
0
1
Vin+ — Vin- (V)

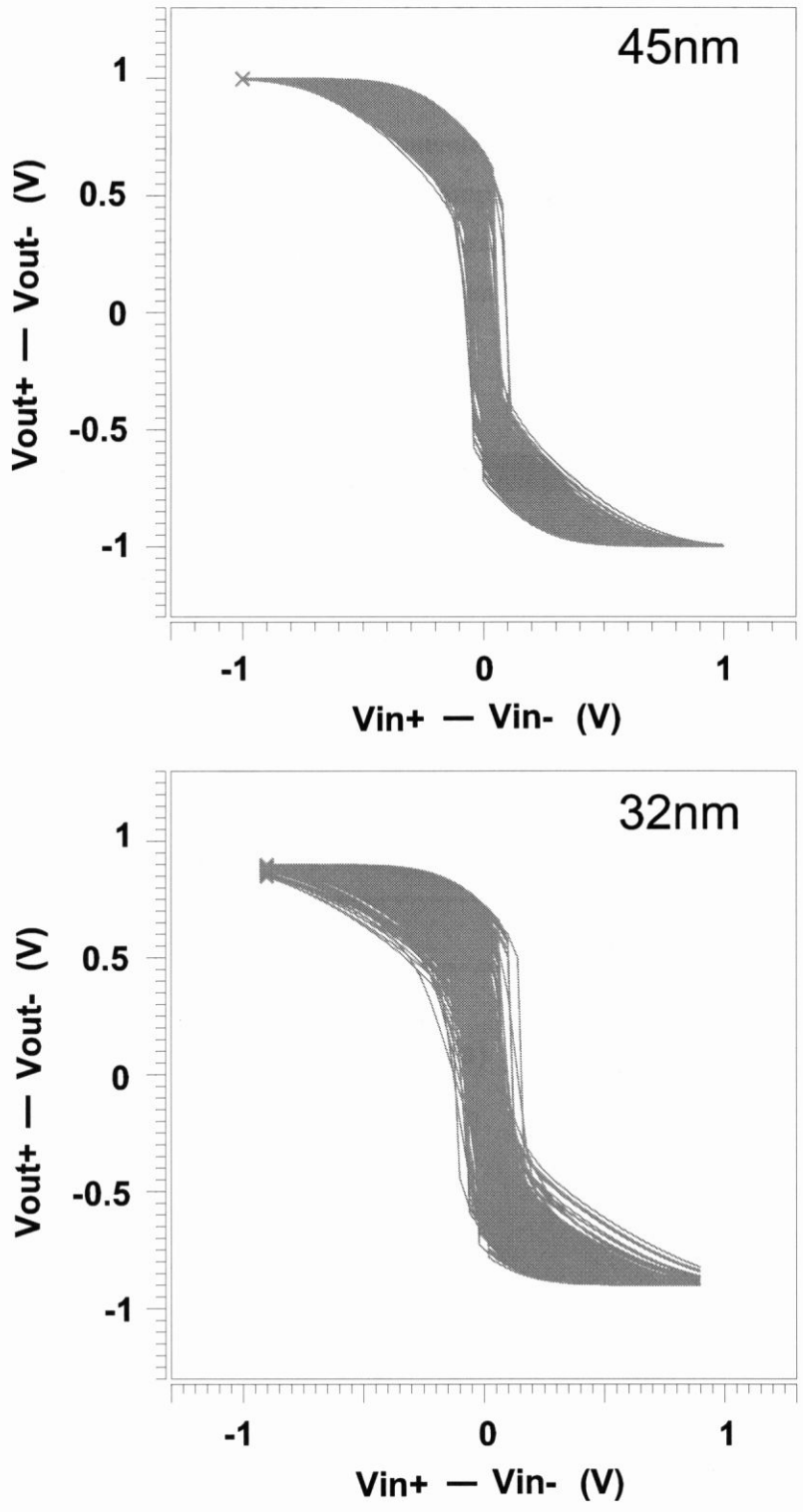
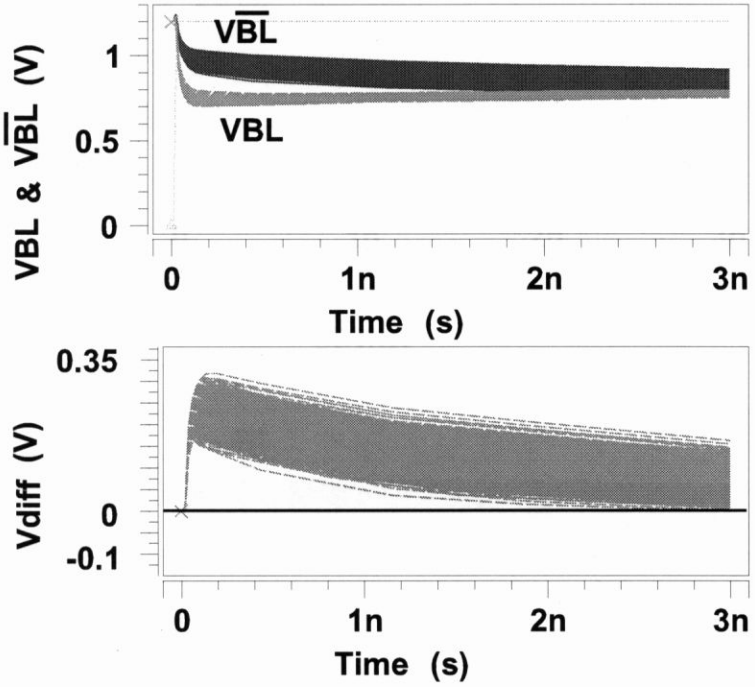


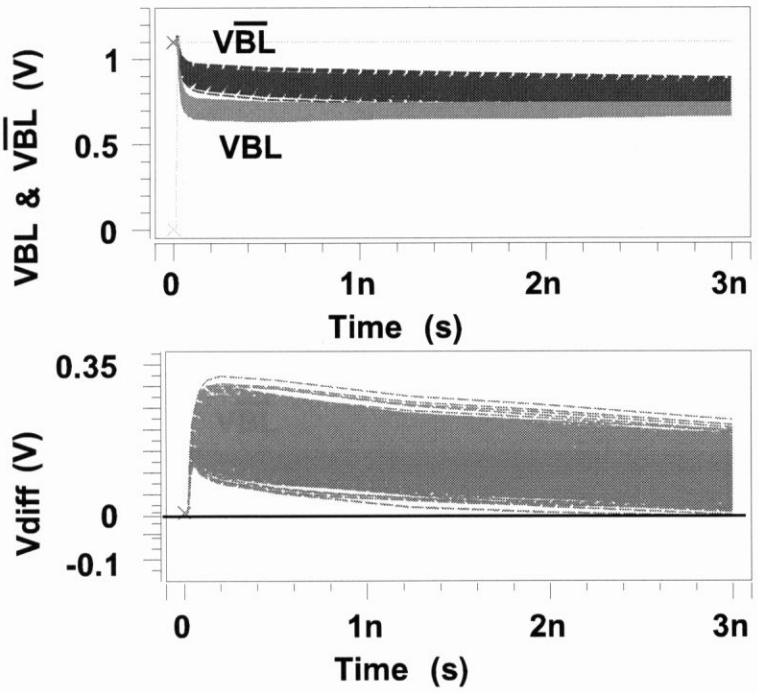
Figure 4.5 Input offset voltage variations of the sense-amp of Figure 4.4

4.3.2 The Variation of the Differential Voltage Generated Between Bit Lines

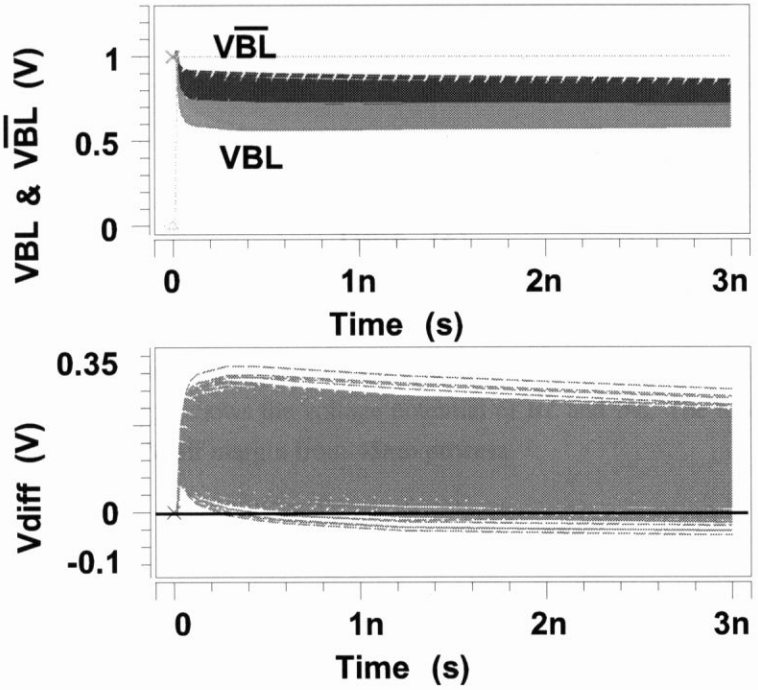
To work properly, the differential voltage generated between bit lines of the logic part (Vdiff) must be larger than the Voff of the sense-amp. Figure 4.6 shows the variations of the Vdiff when only one input is high. The upper graph shows the voltage of $\overline{BL}$ and BL. The lower graph shows Vdiff: $V_{\overline{BL}} - V_{BL}$. It shows that there is no Vdiff margin from 45nm process.



90nm


VBL & VBL (V)
1
0.5
0
VBL
VBL
0
1n
2n
3n
Time (s)
Vdiff (V)
0.35
0
-0.1
0
1n
2n
3n
Time (s)

65nm


VBL & VBL (V)
1
0.5
0
VBL
VBL
0
1n
2n
3n
Time (s)
Vdiff (V)
0.35
0
-0.1
0
1n
2n
3n
Time (s)

45nm

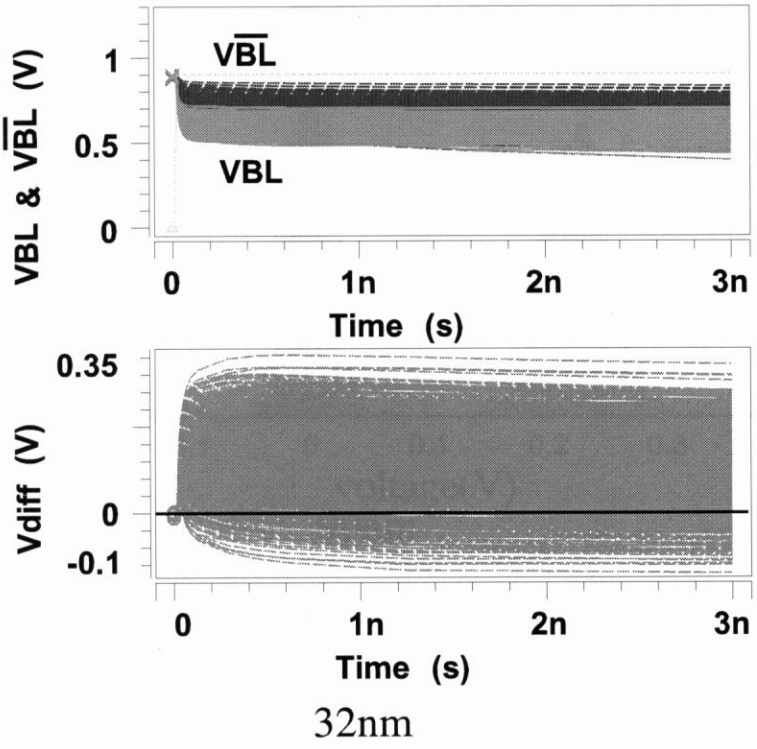
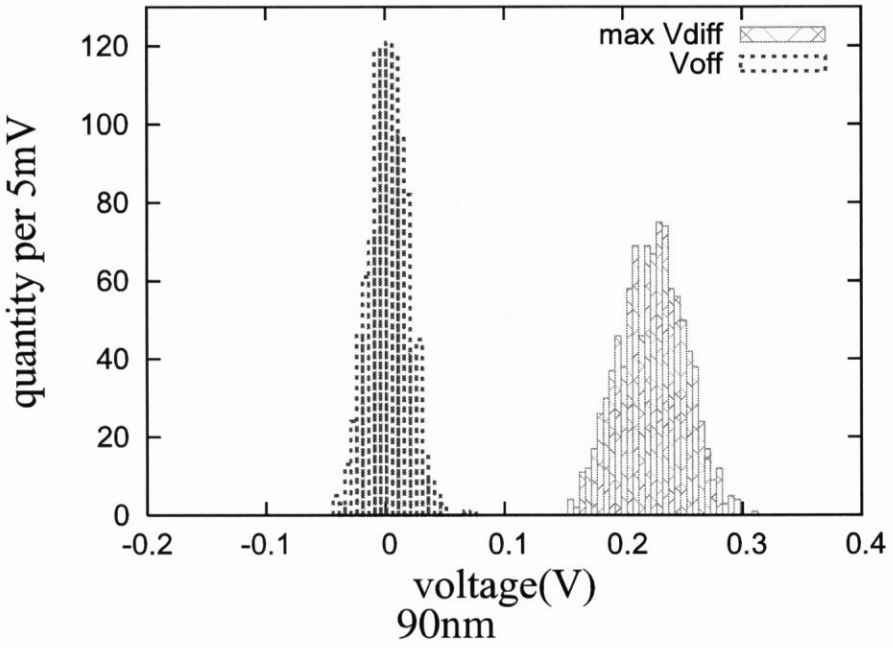
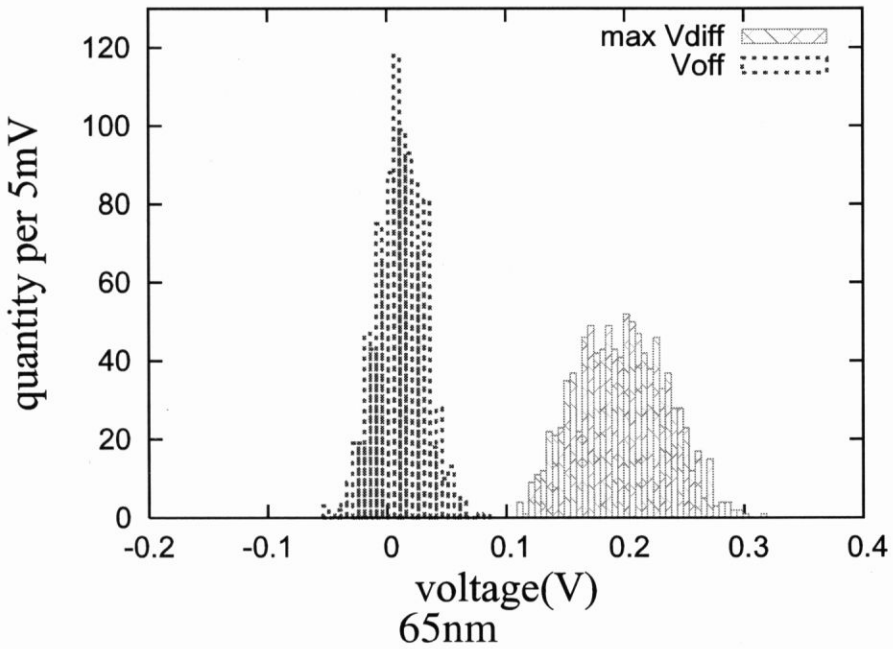


Figure 4.6 Variations of differential voltage between bit lines (Vdiff) when only one input is high. Of each process, the upper graph shows the voltage potential of $\overline{BL}$ and BL. The lower graph shows Vdiff: $V_{\overline{BL}} - V_{BL}$. There is no Vdiff margin from 45nm process.

4.3.3 The Noise Margin of the Dual-Rail PLA

Figure 4.7 gives the profiles of Voff and the maximum Vdiff. They are histograms with an interval of 5mV. In this thesis the noise margin of the dual-rail PLA is defined as the space between the Voff and the maximum Vdiff profiles, which is insufficient and disappears from 65nm process.





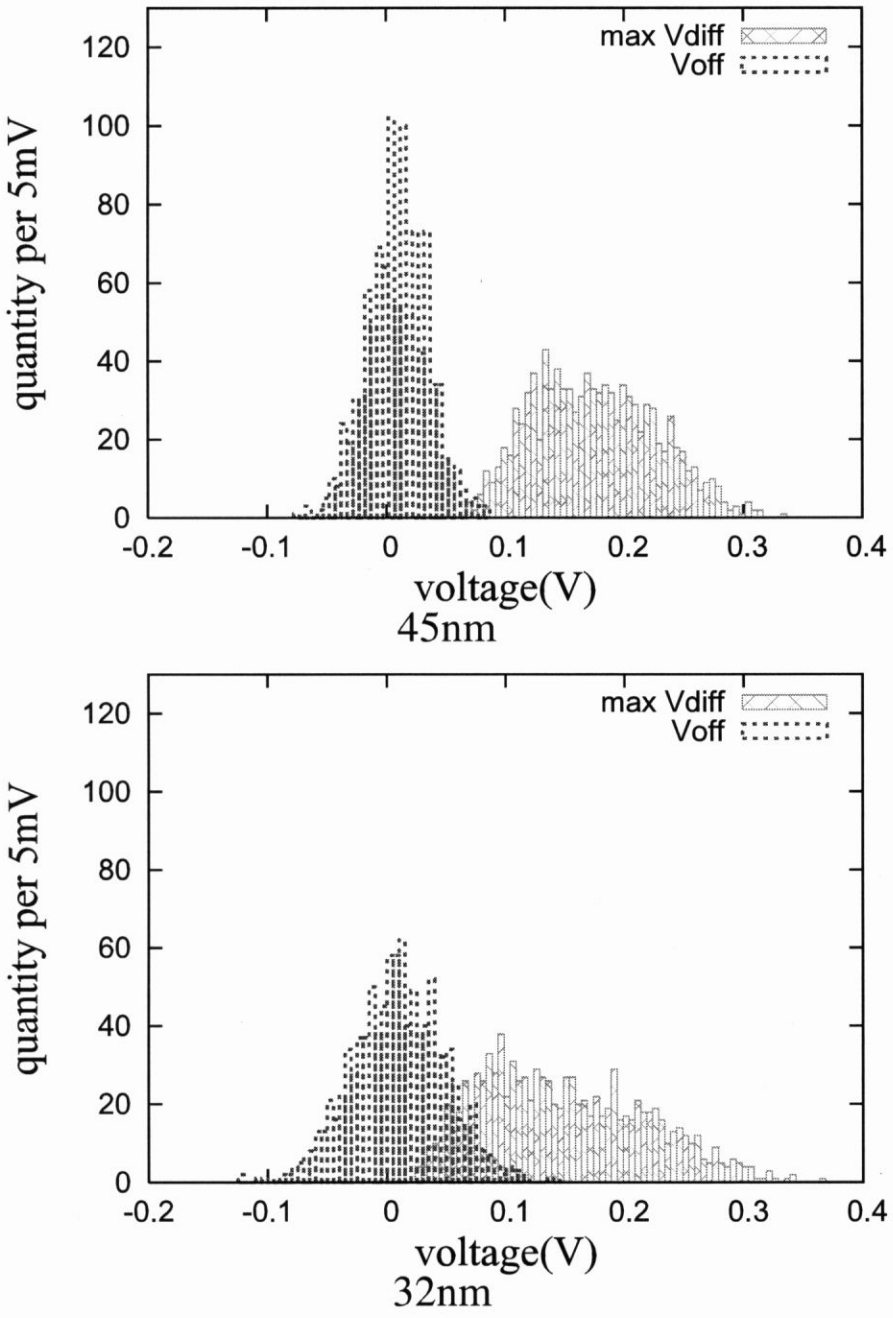


Figure 4.7 Profiles of Voff and the maximum Vdiff. They are histograms with an interval of 5mV. In this thesis the noise margin of the dual-rail PLA is defined as the space between the Voff and the maximum Vdiff profiles, which is insufficient and disappears from 65nm process. (Voff: input offset voltage of sense-amp (Figure 4.5), max Vdiff: maximum differential voltage generated between bit lines (Figure 4.6)

Figure 4.8 shows the trend of noise margins of the dual-rail PLA when only one input is high, with and without Vdd noise. It shows there is no noise margin left for 6σ assurance from 90nm process when $\sigma_{Vdd} = 50mV$. It means this kind of dual-rail PLA can not work reliably in future sub-100nm process.

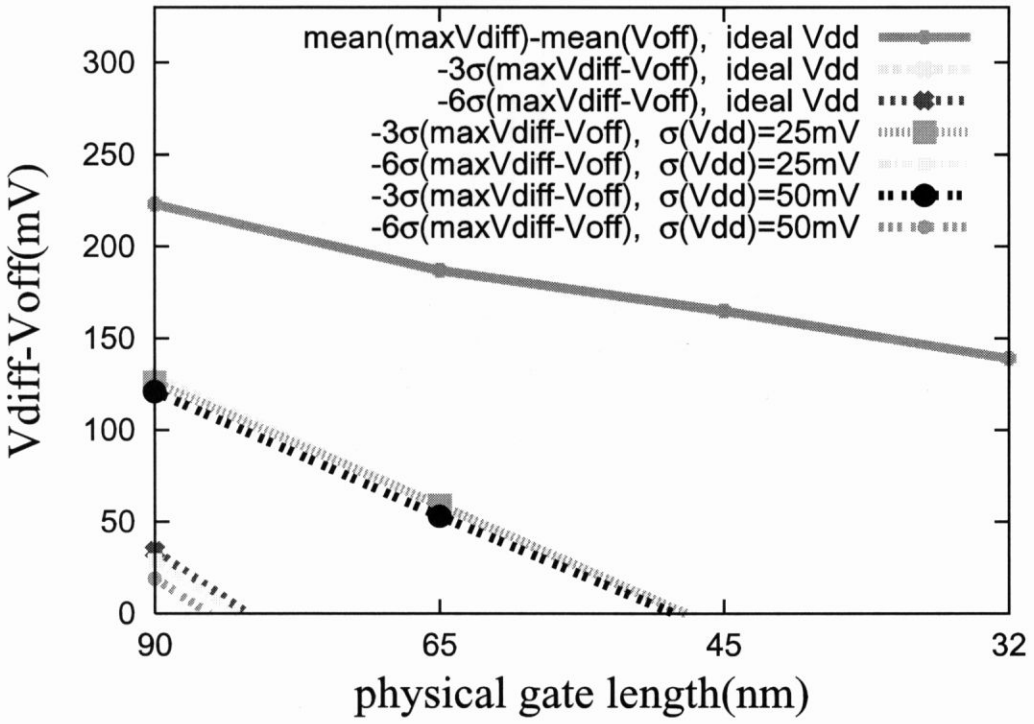


Figure 4.8 The trend of noise margins of the dual-rail PLA when only one input is high, with and without Vdd noise. There is no noise margin left for 6σ assurance from 90nm process when $\sigma_{Vdd} = 50mV$. It means this kind of dual-rail PLA can not work reliably in future sub-100nm process. ($\sigma^2_{maxVdiff-Voff} = \sigma^2_{maxVdiff} + \sigma^2_{Voff}$)

4.4 Parameter Sensitivity Analysis of the Dual-Rail PLA

Instead of varying all the parameters simultaneously, we vary one parameter for one time to estimate the sensitivity of each parameter to the dual-rail PLA. Results are shown in Figure 4.9. It shows how sensitive the σ of noise margins of the dual-rail PLA is to individual parameter. V_{th}, L_g, L_{ovs} and L_{ovd}, V_{dd}, μ affect the σ of noise margins obviously. Compared to random CMOS (refer to Figure 3.12), the sensitivity of V_{dd} is smaller due to the differential voltage operation and the sensitivity of V_{th} is bigger due to the minimum size transistor used in the logic part, and the mismatch problem inside the logic part and the sense-amp (refer to

Eq. 2.6 and note that ΔV_{th} is totally random).

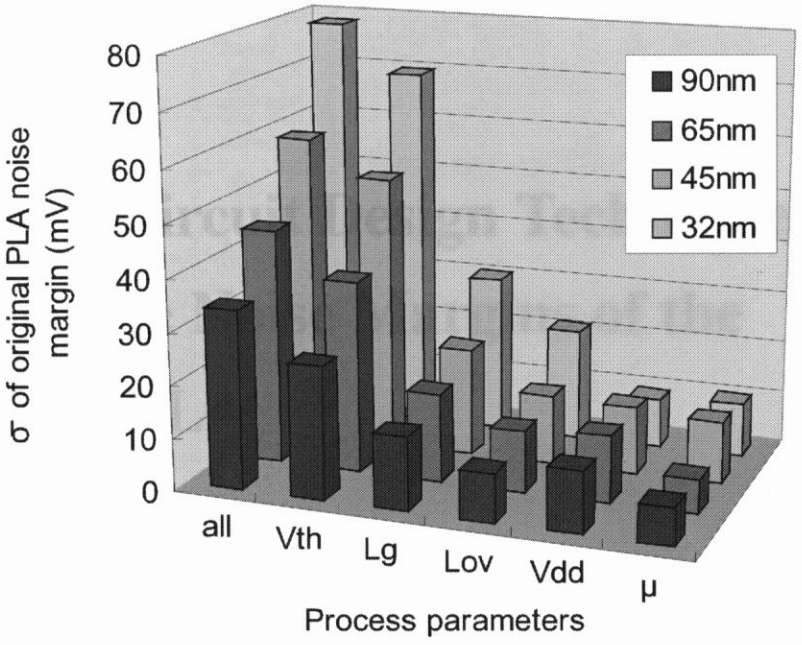


Figure 4.9 Parameter sensitivity analysis of the dual-rail PLA. V_{th}, L_g, L_{ovs} and L_{ovd}, V_{dd}, μ affect the σ of noise margins of the dual-rail PLA obviously. Compared to random CMOS (refer to Figure 3.12), the sensitivity of V_{dd} is smaller due to the differential voltage operation and the sensitivity of V_{th} is bigger due to the minimum size transistor used in the logic part, and the mismatch problem inside the logic part and the sense-amp (refer to Eq. 2.6 and note that ΔV_{th} is totally random).