Circuit Design for Low-Power High-Speed VLSI Processor in 0.5V Generation

(0.5V 世代の低電力・高速 VLSI プロセッサを志向した回路設計)

by

Koichi Nose

野瀬　浩一

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Supervisor:　Professor Takayasu Sakurai

The University of Tokyo
December 14, 2001

Abstract

In the past 30 years, the semiconductor industry has been expanded drastically by the downsizing of the transistors. The progress of the downsizing, however, has increased the chip power. As the battery-powered products, like mobile computers and mobile phones, become popular, the low-power design becomes the one of the most important issues of the LSI design. On the other hand, high-performance design is also a crucial issue since main-stream supply voltage, V_{DD}, will be scaled down to below 0.5V in the coming years. In order to achieve the low-power and high-performance at a same time, not only the device improvement but also the new low-power circuit schemes and new low-power architecture are needed.

First, the short-circuit power component and the voltage dependence on the capacitance, which have not been fully investigated, are discussed. Then, the closed-form formulas are presented for optimum supply voltage and threshold voltage that minimize the power dissipation when technology parameters and required speed are given. The formulas take into account the short-channel effects, the temperature variation and V_{TH} fluctuation. From the calculation using these formulas, it is shown that a simple guideline for power optimization is to set the ratio of the maximum leakage power to the total power around 30%. Extending the analysis, the future VLSI design trend is discussed based on ITRS. The optimum target V_{TH} is almost constant at 0.2V over generations. The proposed scenario shows that more number of MOSFET is consumed in the memory blocks than the logic blocks in the future.

Long wires, like signal buses, become dominant performance limiter in high-performance VLSI's. In this study, closed-form formulas for optimum buffer insertion where the junction capacitance is taken into account are proposed. Using these formulas, the optimum interconnect delay and power comparison among bulk, silicon-on-insulator (SOI) and the double-gate structure are discussed. MOSFET with small junction capacitance, like SOI, can suppress both the interconnect delay and power by 15% compared with MOSFET where the junction capacitance is equal to the gate capacitance, like conventional bulk MOSFET. Based on the above-mentioned analysis, a new buffer insertion scheme for bi-directional buses, namely dual-rail bus (DRB) scheme, which does not have noise problems, and a high-speed buffer insertion scheme for uni-directional buses, namely staggered firing bus (SFB) scheme, are proposed and measured. When 0.07μm design rule is used, DRB scheme can improve the performance

of bi-directional buses by an order of magnitude and SFB scheme can suppress the delay of uni-directional buses by about 20% at 0.18μm generation and beyond. If we use SFB scheme instead of conventional uni-directional buses, 27% power reduction can be achieved while the performance of SFB is the same as that of conventional buses.

Finally, new active leakage power reduction schemes are proposed. In order to suppress the leakage power, it is effective to increase the threshold voltage. In the proposed schemes, the threshold voltage, V_{TH}, is dynamically controlled through software depending on a workload. The dynamical control system consists of the cooperation between software and hardware. There are two techniques to control the threshold voltage dynamically. The first one is controlling the back-gate bias of the transistors, which is called V_{TH}-hopping. The V_{TH}-hopping scheme can achieve 82% power saving compared with the fixed low-V_{TH} circuits in 0.5V supply voltage regime for multimedia applications. A small-scale RISC processor with V_{TH}-hopping and the positive back-gate biased scheme is fabricated. Based on the measured data, performance evaluation is conducted using MPEG-4 video coding. The result shows that 86% power saving can be achieved by using V_{TH}-hopping compared with the fixed positive back-bias scheme. The other technique to control the threshold voltage is controlling V_{DD}. This technique utilizes the Drain Induced Barrier Lowering (DIBL). If the drain-source voltage, that is, the supply voltage is lowered, the subthreshold leakage current can be suppressed since the threshold voltage increases by the DIBL effect. In order to verify the effectiveness of DIBL-hopping, MPEG-4 encoding is simulated based on the measured results. The result shows that 75% power reduction can be achieved compared with the fixed V_{DD} scheme.

These schemes are effective for the design of the future low-voltage, low-power CMOS VLSI's.

Acknowledgement

I would like to express my sincere gratitude to the dissertation supervisor, Prof. Takayasu Sakurai, for much more than the continuous and heartful advices but for providing the opportunities and encouragement on processing my research for years.

I also appreciate Professor Koichiro Hoh, Professor Yoichi Okabe, Professor Tadashi Shibata, Associate Professor Toshiro Hiramoto and Associate Professor Miroru Fujishima, all of the University of Tokyo, for their advice as the qualifying examination committee members.

I would like to thank all the members of Sakurai laboratory. Especially, I would like to give my thanks to Mr. Hiroshi Kawaguchi for giving me a comfortable research environment and many advices in all my life in our laboratory.

I gratefully acknowledge the support of Hitachi, Ltd and Toshiba Corporation. I would like to thank the members of Semiconductor Technology Academic Research Center for discussing our research, and Professor Tadahiro Kuroda of Keio University for his advice and powerful encouragements.

I would like to express my appreciation to all the companies and institutions, such as Toshiba Corporation, Semiconductor Technology Academic Research Center (STARC), Rohm Corporation, Toppan Printing Corporation, NTT Electronics Corporation and Dai Nippon Printing Corporation, for giving me valuable opportunities to fabricate the test chip.

Table of Contents

Chapter 1. Introduction

1.1 Background

In the past 30 years, the semiconductor industry has been expanded drastically by the progress of high-performance and cost-down of a chip. The growth of the semiconductor is based on the "Moore's Law [1]", which is a postulate that number of transistors in a processor or other device will double every 18 to 24 months. In order to increase the number of transistors on a chip, the feature size of the transistor has to be scaled down. The guideline of the transistor scaling has been called as the "scaling theory [2]", which is proposed by R. H. Dennard et al. Table 1.1 shows device parameter sets of a typical scaling theory [3][4]. It can be seen that transistor delay decreases by $1/\kappa$ each generation, yielding faster device with each technology shrinking. Therefore, the scaling theory becomes the target of the manufacture of the LSI (Large Scale Integration) and is widely used as a reasonable theory which can achieve the high-speed and downsizing at a same time.

Table 1.1 Influence of scaling on MOS device characteristics

Parameters		Scaling model		
		Constant field	Constant voltage	$1/\kappa^{0.5}$ voltage
Device size		$1/\kappa$	$1/\kappa$	$1/\kappa$
Gate-oxide thickness	t_{OX}	$1/\kappa$	$1/\kappa$	$1/\kappa^{0.5}$
Substrate doping		κ	κ^2	$\kappa^{1.5}$
Supply voltage	V	$1/\kappa$	1	$1/\kappa^{0.5}$
Electric field	E	1	κ	1
Current	I	$1/\kappa$	κ	$1/\kappa$
Area	A	$1/\kappa^2$	$1/\kappa^2$	$1/\kappa^2$
Capacitance	$C \propto A/t_{OX}$	$1/\kappa$	$1/\kappa$	$1/\kappa^{1.5}$
Gate delay	CV/I	$1/\kappa$	$1/\kappa^2$	$1/\kappa$
Power consumption	IV	$1/\kappa^2$	κ	$1/\kappa^{1.5}$
Power density	IV/A	1	κ^3	$\kappa^{0.5}$

The progress of the downsizing, however, causes the increase of the chip power. Fig. 1.1 shows the plot of the power of MPU and DSP which is shown in ISSCC technical digest of papers. In the 1980's, the mainstream of the LSI technology is shifted from bipolar and NMOS process to CMOS process, which is superior in the downsizing and the low power. At that time, the main design target is not the low power but the high-speed and downsizing. However, as the battery-powered products, like mobile computers and mobile phones, become popular, the low-power design becomes one of the most important targets of the VLSI design. Then, not only the researches on the high-performance designs but also the researches on the low-power designs became active in 1990's.

The power issue is not limited for the portable devices. As is shown in Fig. 1.1, the power dissipation is much more than 10W for high-performance processors. In this case, the heat problems (thermal runaway and noisy fans), and the package cost become crucial issues.

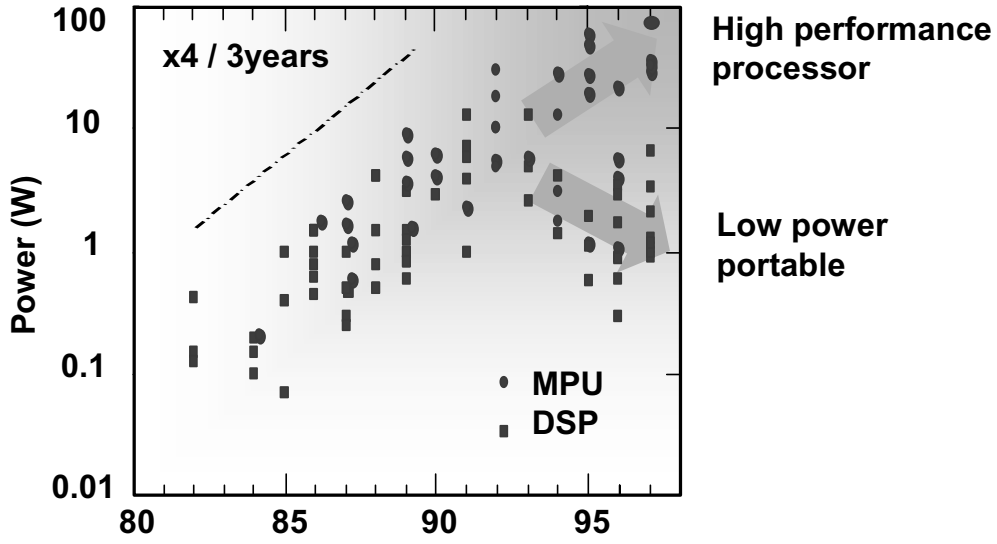


Fig. 1.1 Plot of power of MPU and DSP which is shown in ISSCC technical digest of papers

Another issue of VLSI design is an increase of the interconnect delay. Table 1.2 shows scaling scenarios of the interconnect delay [6]. Although the gate delay decreases by the downsizing of the transistors, the interconnect delay does not decrease since the interconnect resistance and the interconnect capacitance cannot be suppressed at a same time. In particular, the global wiring becomes a more serious problem. The interconnect length scaling for global wiring is set by the chip-size length, which is not shrinking, as are gate dimensions. As a result, the global RC delay scales as κ^3. Some new technologies, copper interconnects [7], new lower dielectric materials [8][9] and hierarchy interconnect [10][11] have been developed to suppress the interconnect resistance and the interconnect capacitance. Even if these technologies are used, however, there is a limit in the improvement of the delay. In order to suppress the delay of a long wiring, the buffer insertion is the most effective. Combining the field solver and the device simulator could be useful to calculate the interconnect delay and optimum buffer design more accurately. This method, however, cannot be used for the present million gates LSI since it takes time too much. Then, a simple but accurate analysis of the

Table 1.2 Scaling scenario of interconnect

Parameters	Local wiring	Global wiring
Line width & spacing	$1/\kappa$	$1/\kappa$
Wire thickness	$1/\kappa$	$1/\kappa$
ILD thickness	$1/\kappa$	$1/\kappa$
Wire length	$1/\kappa$	$1/\kappa^{0.5}$
Resistance (per unit length)	κ^2	κ^2
Capacitance (per unit length)	1	1
RC delay	1	κ^3

interconnect delay and the buffer insertion methodology are indispensable techniques for deep submicron VLSI's.

1.2 Design Issues in 0.5V CMOS VLSI's

As is mentioned in the previous section, low power design is getting one of the key design issues. The power and delay of CMOS gate are simply approximated as

$$POWER \cong afCV_{DD}^2 + I_0 \cdot 10^{-\frac{V_{TH}}{S}} V_{DD}, \qquad (1.1)$$

$$DELAY \cong K\frac{CV_{DD}}{(V_{DD}-V_{TH})^{\alpha}}, \qquad (1.2)$$

where a is switching activity of the gate, f is the operating frequency, C is the load capacitance, V_{DD} is the supply voltage, I_0 is the drain current when the threshold voltage is equal to zero, S and K are the device parameters, V_{TH} is the threshold voltage of the transistor and α is the velocity saturation index [12] whose value is about 1.3~1.5 in the advanced short-channel devices. The details of these formulas and parameters are written in Chapter 2.

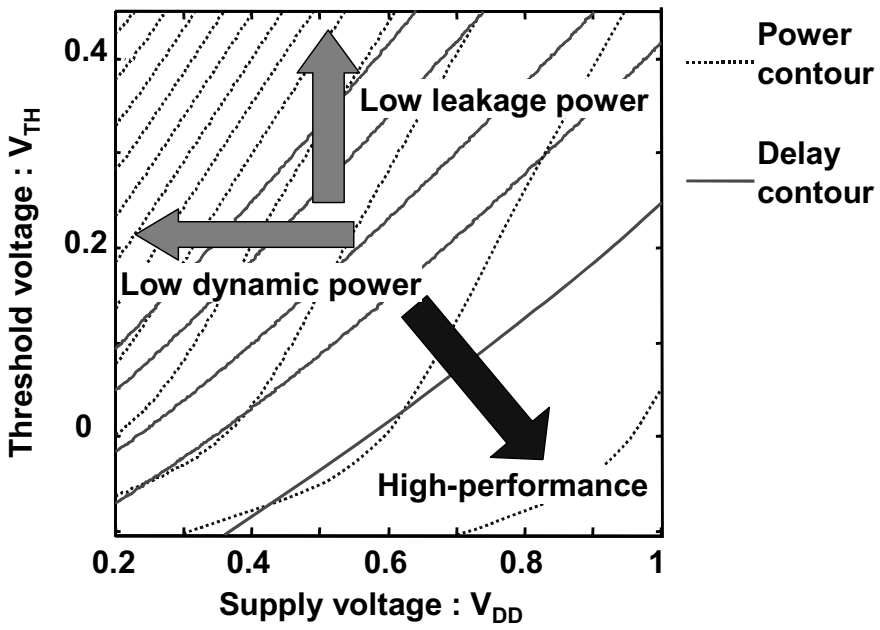


Fig. 1.2 Delay and power dependence on V_{DD} and V_{TH}

The first term of (1.1) is a dynamic power which corresponds to the charging and discharging of a load capacitance. The second term is a subthreshold leakage power component. Based on the formulas, the delay and power dependence on V_{DD} and V_{TH} are depicted in Fig. 1.2. In recent years, V_{DD} has gradually decreased due to the gate oxide reliability and reduction of the dynamic power. Fig. 1.3 shows the future trend of the supply voltage and the chip power, which is predicted by International Technology Roadmap for Semiconductor (ITRS) [5]. This figure shows that main-stream V_{DD} will be scaled down to below 0.5V in the coming years. Lowering V_{DD}, however, causes an increase in gate delay. In order to achieve the high-performance, V_{TH} has to be decreased. Reducing V_{TH}, however, could cause a significant increase in the static leakage power component. Therefore, there is an optimum design where the power is minimized while maintaining the gate delay.

Especially, when V_{TH} is lower than 0.1V, the leakage power becomes a dominant component in the total power consumption. In order to suppress the power consumption in low-voltage processors, it is necessary to reduce the leakage power component.

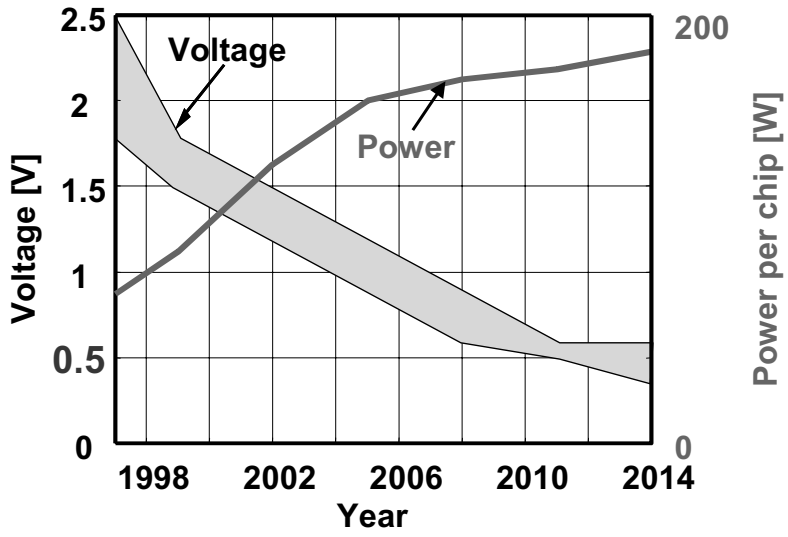


Fig. 1.3 Future trend of V_{DD} and chip power

As for the interconnect delay optimization, H. B. Bakoglu [13] proposed a buffer insertion methodology based on the method of replacing the buffer with a linear resistor. In this context, the interconnect delay optimization by buffer insertion has been investigated [14] but the existing theories are lacking in the detailed consideration on the non-linear feature of buffers and the influence of junction capacitances. Moreover, the existing theories are lacking in the trade-off between the delay and the power consumption although the power is one of the most important index in future giga-scale integration.

1.3 Research Objectives and Chapter Organization

In order to design a low-power and high-performance processor, optimization methodologies of CMOS circuits and new low-power circuit schemes are proposed in this thesis. Fig. 1.4 is the schematic of the proposed low-power and high-performance circuit design.

In Chapter 2, a drain current model and a simple power consumption model are introduced to analyze the power and delay of CMOS circuits accurately.

In Chapter 3, new analytical formulas for power and delay calculation are proposed. In order to calculate the power more accurately, a short-circuit power component and a voltage dependence on capacitances, which have not been fully investigated, are discussed. An optimum design for low-power processor considering the temperature variation and the threshold voltage fluctuation is analyzed. Extending the analysis, the future VLSI design trend is discussed based on ITRS [5].

Long wires, like signal buses, become dominant performance limiter in high-performance VLSI's. In Chapter 4, new interconnect methodologies are proposed to improve the chip performance and total power dissipation. Closed-form formulas for optimum buffer insertion where the junction capacitance is taken into account are proposed. In order to use the derived formulas, an appropriate choice of the effective linear resistance of the driving transistor is also clarified. The result have been applied to bulk and SOI technologies and implications of buffered interconnect on technology are proposed. In order to alleviate the noise problems and delay fluctuation problems, new buffer insertion schemes for bi-directional and uni-directional buses are implemented and measured.

In Chapter 5, new active leakage power reduction schemes are proposed where the threshold voltage is dynamically controlled through software depending on a workload.

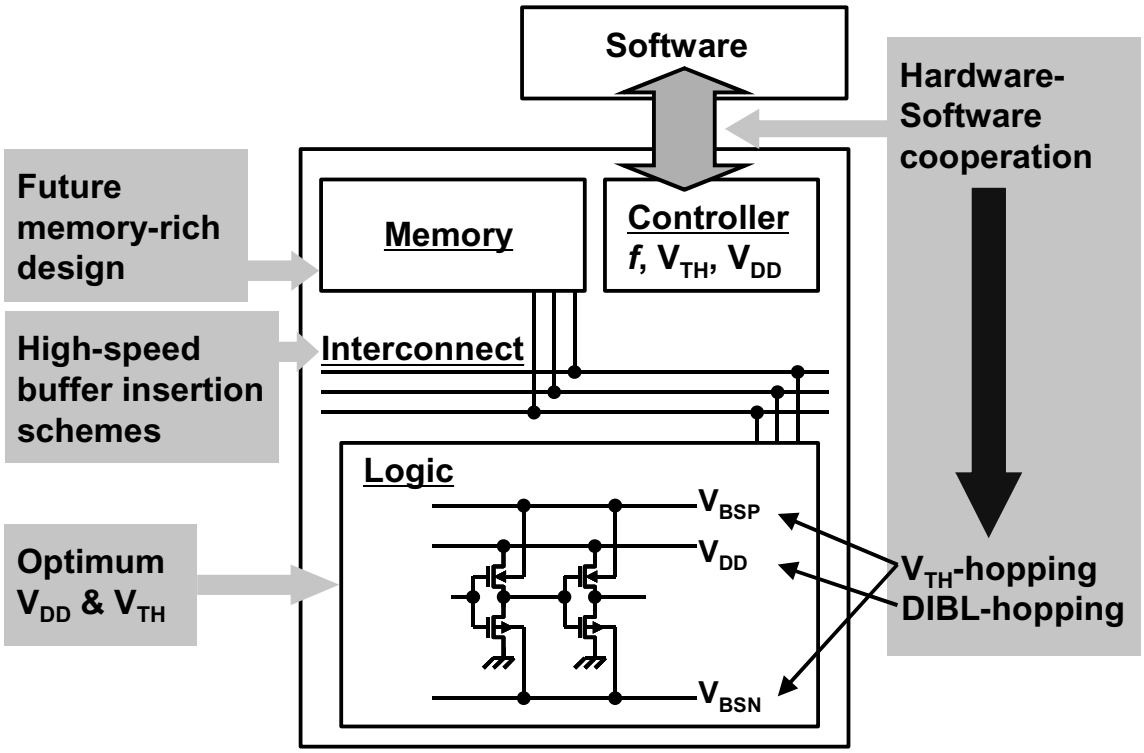


Fig. 1.4 Schematic of low-power and high-performance circuit design

In the proposed schemes, the threshold voltage is dynamically controlled through software depending on a workload. The dynamical control system consists of the cooperation between software and hardware. We propose two techniques to control the threshold voltage dynamically, one is controlling the back-gate bias of the transistors, and the other is controlling the supply voltage. The latter technique utilizes the Drain Induced Barrier Lowering (DIBL) [15].

Finally, the conclusion of this thesis is given in Chapter 6.

References

[1] G. M. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, vol.3, no.8, Apr., 1965.

[2] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous and A. R. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol.9, pp.256-268, 1974.

[3] T. Kuroda and T. Sakurai, "Overview of low-power ULSI circuit techniques," *IEICE Trans. Electron.*, vol.E78-C, no.4, pp.334-343, Apr., 1995.

[4] M. Kakumu, "Process and device technologies of CMOS devices for low-voltage operation," *IEICE Trans. Electron*., vol.E76-C, no.5, pp.672-680, May, 1993.

[5] The International Technology Roadmap for Semiconductors, SIA Handbook, 1998.

[6] D. Sylvester and C. Hu, "Analytical modeling and characterization of deep-submicrometer interconnect," *Proceeding of the IEEE*, vol.89, no.5, pp.634-664, May, 2001.

[7] D. Edelstein et al, "Full copper wiring in a sub-0.25μm CMOS VLSI technology," *International Electron Device Meeting (IEDM) Tech. Dig.*, pp.773-776, 1997.

[8] S. P. Jeng, K. Taylor, T. Seha, M. C. Chang, J. Fattaruso and R. H. Havemann, "Highly porous interlayer dielectric for interconnect capacitance reduction," *Symp. on VLSI Tech. Dig. of Papers*, pp.61-62, 1995.

[9] A. R. K .Ralston et al, "Integration of thermally stable, low-k AF4 polymer for 0.18μm interconnects and beyond," *Symp. on VLSI Tech. Dig. of Papers*, pp.81-82, 1997.

[10] J. A. Davis, V. K. De and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) – part II: applications to clock frequency, power

dissipation and chip size estimation," *IEEE Trans. Electron Devices*, vol. 45, no.3, pp.590-597, Mar., 1998.

[11] K. Yamashita and S. Odanaka, "Interconnect scaling scenario using a chip level interconnect model," *Symp. on VLSI Tech. Dig. of Papers*, pp.53-54, 1997.

[12] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol.SC-25, pp.584-594, Apr., 1990.

[13] H. B. Bakoglu, *Circuits, interconnections, and packaging for VLSI*, Reading MA: Addison Wesley, 1990.

[14] Y. I. Ismail and E. G. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits," *IEEE Trans. VLSI systems*, vol. 8, no. 2, pp.195-206, Apr., 2000.

[15] R. R. Troutman, "VLSI limitations from drain induced barrier lowering," *IEEE Trans. Electron Devices*, vol.ED-26, no.4, pp.461, Apr., 1979.

Chapter 2. Principles of MOSFET Models

2.1 Short-Channel MOSFET Model

The most famous MOSFET current model is Shockley model [1], which is proposed by W. Schockley in 1952. In the Shockley model, the drain current, I_D, is expressed as

$$I_D = \begin{cases} \beta\left\{(V_{GS} - V_{TH})V_{DS} - \frac{1}{2}V_{DS}^2\right\} & (\mathrm{V_{DS}} < \mathrm{V}_{GS} - \mathrm{V_{TH}} : \text{linear region}) \\ \frac{\beta}{2}(V_{GS} - V_{TH})^2 & (\mathrm{V_{DS}} \geq \mathrm{V}_{GS} - \mathrm{V_{TH}} : \text{saturation region}) \end{cases}, \quad (2.1)$$

where V_{GS} is the gate-source voltage, V_{DS} is the drain-source voltage, V_{TH} is the threshold voltage and β is the coefficient which is determined by the device parameters. The Shockley model is widely known as the drain current model of conventional long-channel MOSFETs. This model, however, is not suitable for short-channel MOSFETs since the short-channel effects are not taken into account. In order to calculate the delay and the power of the short-channel MOSFETs, the simple short-channel MOSFET model has been proposed as "alpha-power law model" by T. Sakurai et al [2][3]. In this model, I_D is

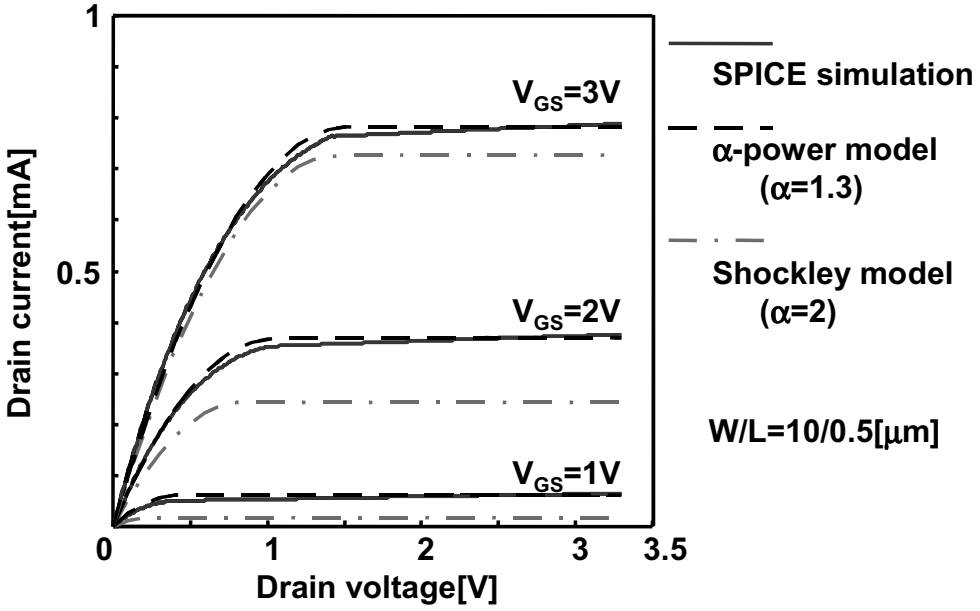


Fig. 2.1 Comparison among SPICE simulation, Shockley model and alpha-power law model

expressed as

$$I_D = \begin{cases} I'_{D0}\left(2 - \dfrac{V_{DS}}{V'_{DS}}\right)\dfrac{V_{DS}}{V'_{DS}} & (V_{DS} < V'_{D0} \text{ linear region}) \\ I'_{D0} & (V_{DS} \geq V'_{D0} \text{ saturation region}) \end{cases}, \qquad (2.2)$$

where

$$I'_{D0} = I_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\alpha}, \qquad (2.3)$$

$$V'_{D0} = V_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\frac{\alpha}{2}}. \qquad (2.4)$$

V_{D0} is defined as the drain saturation voltage when $V_{GS}=V_{DD}$, and I_{D0} (=drain current when $V_{GS}=V_{DS}=V_{DD}$) is a good index of the drivability of a MOSFET. α is the velocity saturation index. α is about 1.3~1.5 for an advanced short-channel MOSFET. When

α=2, the model corresponds to the Shockley model. Fig. 2.1 shows the comparison among SPICE simulation result, the Shockley model and the alpha-power law model. The process of the SPICE model is 0.5μm CMOS technology. The result shows the alpha-power law model is better approximation than the Shockley model. Then, in this thesis, the alpha-power law model is used as the transistor current model.

The inverter delay can be approximately derived from the alpha-power law model. In [3], the delay formula is approximated as

$$T_{pd} \cong \frac{CV_{DD}}{I_{D0}}\left\{\left(\frac{0.9}{0.8}+\frac{1}{0.8}\frac{V_{D0}}{V_{DD}}\ln\frac{10V_{D0}}{eV_{DD}}\right)\left(\frac{V_{TH}/V_{DD}+\alpha}{1+\alpha}-\frac{1}{2}\right)+\frac{1}{2}\right\}. \quad (2.5)$$

This formula, however, is too complicated to use as a simple model. In this study, simple formula which is written as (2.6) is used.

$$DELAY \cong K\frac{CV_{DD}}{(V_{DD}-V_{TH})^{\alpha}} \quad (2.6)$$

K is the delay coefficient which does not depend on V_{DD} and V_{TH}. The comparison between (2.6) and SPICE simulation result is shown in Fig. 2.2. In this simulation, 0.3μm CMOS process parameters are used. The maximum discrepancy between the simulation result and the analytical formula is 6% when $V_{TH} \geq$ -0.1V. Generally, V_{TH} is not set to below –0.1V since the excessive leakage current flows. Thus, the approximate formula is effective as the simple delay formula.

On the other hand, the leakage power becomes one of the most important components of the power consumption. In order to calculate the leakage power, a simple transistor leakage current model is used. The model is expressed as

$$I_{LEAK} = I_0 \cdot 10^{-\frac{V_{GS}-V_{TH}}{S}}, \quad (2.7)$$

where I_0 is the drain current of MOSFET when V_{GS}=V_{DS}=V_{DD} and V_{TH}=0, S is the subthreshold slope, called S-factor, which is determined by the device structure.

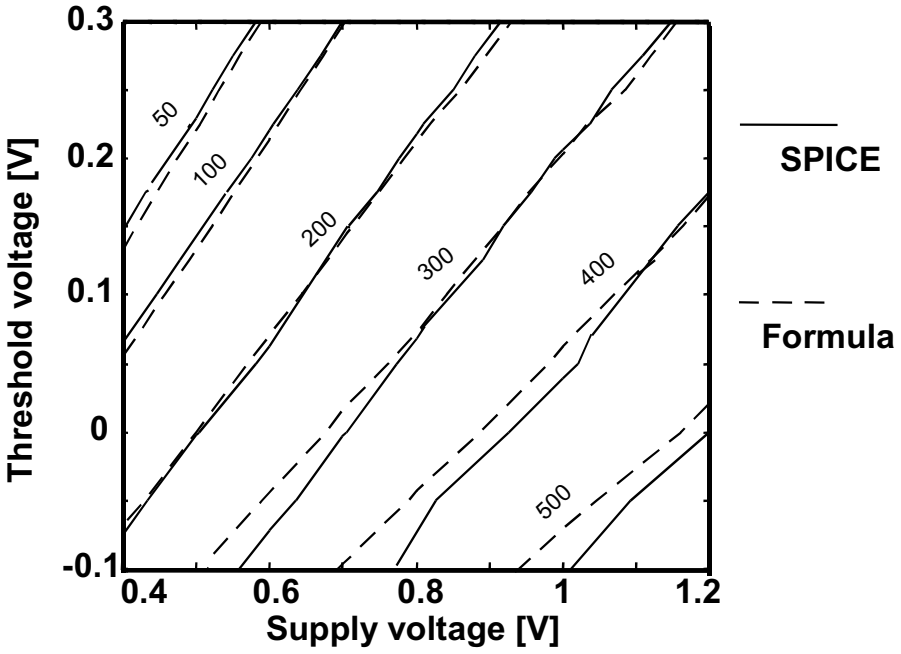


Fig. 2.2 Delay comparison between SPICE simulation and simple formula

In order to suppress the leakage power, lowering S-factor is effective. The smallest S-factor, however, is 60mV/decace at room temperature and this value is not scaled by the reduction of feature size of MOSFET but only by the temperature. Another method of lowering leakage power is to increase V_{TH} but the delay of the gate, which is written as (2.6), increases. The trade-off between V_{DD} and V_{TH} and the guideline of the optimum design are discussed in Section 3.4.

2.2 Power Consumption Model of CMOS Circuit

The total power of CMOS gate is shown as the following formula.

$$POWER = P_D + P_{LEAK} + P_S = afCV_{DD}^2 + I_0 \cdot 10^{-\frac{V_{TH}}{S}} \cdot V_{DD} + aI_S V_{DD} \qquad (2.8)$$

a is switching activity of the gate, f is the operating frequency, C is the load capacitance and I_S is a short circuits current.

In this formula, the first term, P_D, is the dynamic power which corresponds to the charging and discharging of the load capacitance. The second term, P_{LEAK}, is the subthreshold leakage power component. The last term, P_S, is the short-circuit power, which flows through a turning-off MOSFET. Although the first and the second terms are well characterized, the short-circuit power component has not been fully studied. In order to analyze the power dissipation more accurately, studying the short-circuit power is crucial for the future VLSI design. The short-circuit power is discussed in Section 3.2.

References

[1] W. Shockley, “A unipolar field effect transistor,” *Proceedings of IRE*, vol.40, pp.1365-1376, Nov., 1952.

[2] T. Sakurai and A. R. Newton, “Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas,” *IEEE Journal of Solid-State Circuits*, vol. SC-25, pp.584-594, Apr. 1990.

[3] T. Sakurai and A. R. Newton, “Delay analysis of series-connected MOSFET circuits,” *IEEE Journal of Solid-State Circuits*, vol. SC-26, pp.122-131, Feb. 1991.

Chapter 3. Analysis and Future Trend of Low-Power and High-Performance Circuits

3.1 Introduction

As power dissipation becomes the more serious problem, the more accurate estimation of the power dissipation is needed. In order to analyze the power dissipation accurately, it is necessary to discuss two effects which have not been fully investigated in low-voltage, low-power CMOS circuits.

One is a short-circuit power component, P_S, which is mentioned in Chapter 2. Veedrick [1] first reported an expression for P_S but it did not take into account the P_S dependence on the load capacitance, C_{OUT}, although P_S is a strong function of C_{OUT}. In [2], P_S dependence on C_{OUT} was first introduced but it neglected the short-channel effects on P_S. Papers [3] and [4] then introduced the short-channel effects in P_S through the use of α-power law MOS model [5], but their expressions diverge to infinity when C_{OUT}=0 which is not true in reality, and hence loses reliability when the load capacitance is small. One

more drawback is that the expressions include the solution of quadratic or cubic equations so that the expressions are complicated. In Section 3.2, a closed-form expression is presented which resolves the above-mentioned problems and the future trend of the P_S / (P_D + P_S) is discussed, which answers a long-standing question if the P_S is getting more and more serious or not in the future. This study also handles the short-circuit power of series-connected MOSFET structures which appear in NAND and other complex gates.

The other effect is a gate capacitance dependence on the terminal voltages. Load capacitance of CMOS circuits, C_{LOAD}, which determines the power and delay is expressed as follows.

$$C_{LOAD} = \sum C_G + \sum C_J + \sum C_{INT} , \qquad (3.1)$$

where C_G, C_J and C_{INT} denote gate, junction and interconnection capacitance, respectively. In these capacitances, C_G and C_J have complex voltage dependency on terminal voltages but the impact of this voltage dependency of C_G and C_J on power and delay has not been fully investigated, especially, in low-voltage, low-power designs. In Section 3.3, the effect of the voltage dependent gate capacitance on circuit behaviors is analyzed and an appropriate choice of the effective constant gate capacitance is discussed. The impact of the voltage dependent nature is investigated for low-voltage, low-power designs.

In Section 3.4, closed-form formulas are presented for optimum V_{DD} and V_{TH} that minimize power dissipation when the technology and required speed are given. $V_{TH,min}$ is considered in this formula to incorporate V_{TH} fluctuation effects. The resultant formulas have been applied to the technology roadmap to discuss the future VLSI design trend.

3.2 Analysis and Future Trend of Short-Circuit Power

List of Parameters Used

α	velocity saturation index
α_N	velocity saturation index of NMOS
α_P	velocity saturation index of PMOS
α_{NNJ}	effective velocity saturation index of N series-connected NMOS structure with J-th NMOS gate from output as an input
α_{PNJ}	effective velocity saturation index of N series-connected PMOS structure with J-th PMOS gate from output as an input
β_r	beta ratio ($=I_{D0P}/I_{D0N}$)
C_{IN}	input node capacitance
C_{OUT}	output load capacitance
C_G	gate capacitance of inverter
η_P	power ratio ($=P_S/(P_D+P_S)$)
f	frequency
FO	fanout ($=C_{OUT}/C_{IN}$)
fo	transistor drivability ratio of succeeding gates
I_D	drain current
I_{DN}	drain current of NMOS
I_{DP}	drain current of PMOS
I_{D0N}	saturated drain current of NMOS at $V_{GSN}=V_{DSN}=V_{DD}$
I_{D0P}	saturated drain current of PMOS at $\|V_{GSP}\|=\|V_{DSP}\|=V_{DD}$
I_{D0NIN}	saturated drain current of NMOS of previous gate stage
I_{D0PIN}	saturated drain current of PMOS of previous gate stage
L_N	NMOS channel length

L_P PMOS channel length

P_D dynamic power dissipation per switching ($=C_{OUT}V_{DD}^2/2$)

P_S short-circuit power dissipation per switching

t time

t_T transition time of input voltage

τ_N $=C_{OUT}V_{DD}/I_{D0N}$, transition time of output voltage

V_{DD} supply voltage

V_{D0} drain saturated voltage at $V_{GSP}=V_{DD}$

V_{D0P} drain saturated voltage of PMOS at $V_{GSP}=V_{DD}$

V_{DS} drain-source voltage

V_{GS} gate-source voltage

V_{GSP} gate-source voltage of PMOS

V_{OUT} output voltage

V_{TH} threshold voltage

V_{THN} threshold voltage of NMOS

V_{THP} threshold voltage of PMOS

v_{D0P} normalized drain saturated voltage of PMOS at $V_{GSP}=V_{DD}$ ($=V_{D0P}/V_{DD}$)

v_{OUT} normalized output voltage ($=V_{OUT}/V_{DD}$)

v_{TN} normalized threshold voltage of NMOS ($=V_{THN}/V_{DD}$)

v_{TP} normalized threshold voltage of PMOS ($=V_{THP}/V_{DD}$)

3.2.1 Short-Circuit Power Dissipation Formula

Fig. 3.1 shows the typical input and output voltage waveforms of a CMOS inverter discharging the load capacitance. Although discharging case is described here, the charging case can be treated similarly. t_T is a transient time of the input voltage, t_0 is the time when the input voltage reaches the threshold voltage of NMOS, and t_1 is the time

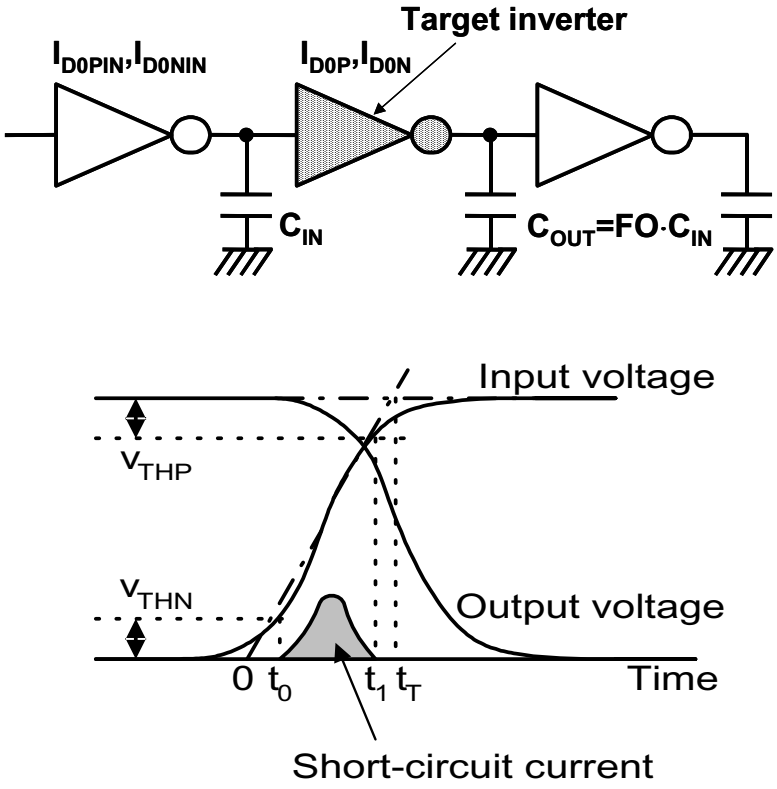


Fig. 3.1 Voltage waveform of CMOS inverter operation

when the input voltage reaches the threshold voltage of PMOS. The short-circuit current flows between t_0 and t_1. When C_{OUT} is sufficiently large, it can be assumed that NMOS operates in the saturated region and PMOS operates in the linear region between t_0 and t_1. With these assumptions, an expression for short-circuit power when the input is very fast, $P_S(t_T \ll \tau_N)$, can be derived as follows.

$$P_S(t_T << \tau_N) = 2\frac{I_{D0P}I_{D0N}}{v_{D0P}C_{OUT}}t_T^2\frac{(1-v_{TN}-v_{TP})^{\frac{\alpha_P}{2}+\alpha_N+1}}{(1-v_{TN})^{\alpha_N}(1-v_{TP})^{\alpha_P/2}}\frac{f(\alpha)}{\alpha_N+1} \qquad (3.2)$$

where

$$f(\alpha) = \left\{\frac{1}{\alpha_N+2} - \frac{\alpha_p}{2(\alpha_N+3)} + \frac{\alpha_P}{\alpha_N+4}\left(\frac{\alpha_P}{2}-1\right)\right\}. \qquad (3.3)$$

The detailed derivation of $f(\alpha)$ can be found in Appendix A. The expression for a charging case of the load capacitance can be obtained by exchanging *N* and *P* suffixes.

This formula, however, suffers from the above-mentioned problem that the P_S diverges to infinity when C_{OUT}=0. On the other hand, P_S expression for C_{OUT}=0 case, which means that the input rump is slower than the output transition, has been obtained ($P_S(t_T \gg \tau_N)$) as follows [5]

$$P_S(t_T >> \tau_N) = V_{DD} t_T I_{D0P} \frac{1}{\alpha_P + 1} \frac{1}{2^{\alpha_P}} \frac{(1 - v_{TN} - v_{TP})^{\alpha_P + 1}}{(1 - v_{TP})^{\alpha_P}}. \quad (3.4)$$

Now, (3.2) and (3.4) are combined by taking a harmonic average of the two quantities to build the general formula, P_S, which covers both of the slow and fast input case. The resultant expression for P_S is free from the above-mentioned divergence problem.

$$P_S = \frac{1}{\frac{1}{P_S(t_T << \tau_N)} + \frac{1}{P_S(t_T >> \tau_N)}}. \quad (3.5)$$

Substituting (3.2) and (3.4) into (3.5), the short-circuit power dissipation is obtained as follows.

$$P_S = \frac{1}{\frac{v_{D0P} C_{OUT} g(v_T, \alpha)}{2 I_{D0P} I_{D0N} t_T^2} + \frac{h(v_T, \alpha)}{V_{DD} t_T I_{D0P}}} \quad (3.6)$$

where

$$g(v_T, \alpha) = \frac{\alpha_N + 1}{f(\alpha)} \frac{(1 - v_{TN})^{\alpha_N} (1 - v_{TP})^{\alpha_P / 2}}{(1 - v_{TN} - v_{TP})^{\alpha_P / 2 + \alpha_N + 2}} \quad (3.7)$$

$$h(v_T, \alpha) = 2^{\alpha_P} (\alpha_P + 1) \frac{(1 - v_{TP})^{\alpha_P}}{(1 - v_{TN} - v_{TP})^{\alpha_P + 1}} \quad (3.8)$$

This formula expresses the P_S in terms of t_T and can be used to estimate the short-circuit

power when input transition time is given. In discussing the scaling characteristics of the short-circuit power dissipation, however, it is better to eliminate t_T by replacing t_T with a function of the saturated drain current of the previous gate stage, I_{D0NIN}, I_{D0PIN}, and the input node capacitance, C_{IN} [5].

Since the input voltage is the output voltage of another CMOS logic gate, the transient time, t_T, can be expressed as below [5].

$$t_T = \frac{C_{IN}V_{DD}}{I_{D0PIN}}\left(\frac{0.9}{0.8} + \frac{v_{D0P}}{0.8}\ln\frac{10v_{D0P}}{e}\right)\left(= \frac{C_{IN}V_{DD}}{I_{D0PIN}}k(v_{D0P})\right) \quad (3.9)$$

Substituting (3.9) into (3.6), the short-circuit power dissipation without using t_T is readily obtained as follows.

$$P_S = \frac{k(v_{D0P}){V_{DD}}^2 C_{IN} {fo}^2}{\frac{v_{D0P}g(v_T,\alpha)}{2k(v_{D0P})}FO\beta_r + h(v_T,\alpha)fo} \quad (3.10)$$

$$k(v_{D0P}) = \frac{0.9}{0.8} + \frac{v_{D0P}}{0.8}\ln\frac{10v_{D0P}}{e} \quad (3.11)$$

$$g(v_T,\alpha) = \frac{\alpha_N + 1}{f(\alpha)}\frac{(1-v_{TN})^{\alpha_N}(1-v_{TP})^{\alpha_P/2}}{(1-v_{TN}-v_{TP})^{\alpha_P/2+\alpha_N+2}} \quad (3.12)$$

$$h(v_T,\alpha) = 2^{\alpha_P}(\alpha_P+1)\frac{(1-v_{TP})^{\alpha_P}}{(1-v_{TN}-v_{TP})^{\alpha_P+1}} \quad (3.13)$$

$$FO = \frac{C_{OUT}}{C_{IN}}, \qquad fo = \frac{I_{D0P}}{I_{D0PIN}}, \qquad \beta_r = \frac{I_{D0P}}{I_{D0N}} \quad (3.14)$$

3.2.2 Comparison Between Calculated and SPICE Simulation Results

The calculation results by the proposed formula (3.10) agree well with the SPICE

Table 3.1 SPICE level3 MOS parameter sets

	Tech. A	Tech. B
V_{THN} (V_{BS}=0) [V]	0.55	0.57
V_{THP} (V_{BS}=0) [V]	0.61	0.56
I_{D0}(W_N=10μm) [mA]	0.92	1.8
V_{D0}	0.5	0.5
α_N	1.38	1.6
α_P	1.3	1.6

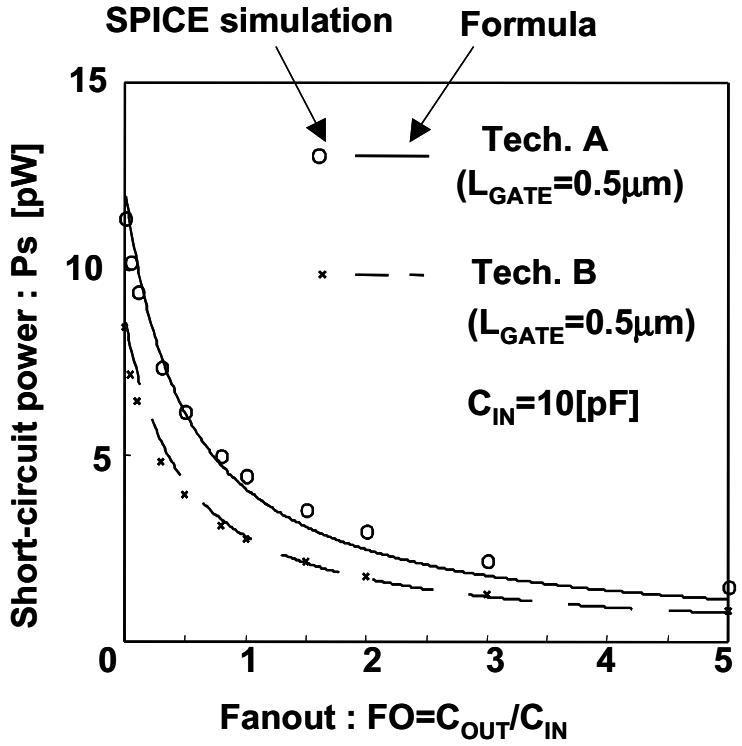


Fig. 3.2 Short-circuit power dependence on fanout

simulation results as shown in Fig. 3.2. Two completely different MOS model parameter sets are used to show the validity of the formula. The MOS parameter sets are listed in Table 3.1. A CMOS inverter chain shown in Fig. 3.1 is used for the comparison. In order to confirm the validity of the proposed formulas when the typical load capacitance (fF order) is used, the short-circuit power dependence on the load capacitance (C_{IN} and

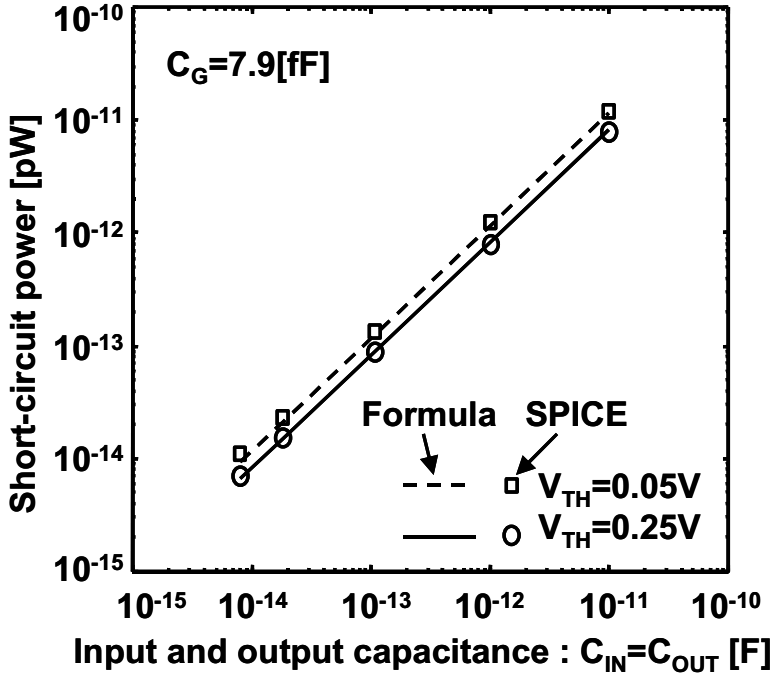


Fig. 3.3 Short-circuit power dependence on input and output node capacitance

C_{OUT}) is calculated. Fig. 3.3 shows the result. FO is set to 1 and C_{IN} and C_{OUT} changes from 7.9fF(the gate capacitance of the inverter, C_G) to 10pF. It is seen that the proposed formulas are in good accordance with the SPICE simulation.

In Fig. 3.4, the SPICE simulation results for the dependence of P_S on FO are compared with the calculation results by the present formula (3.10) and the previously published Vemuru et al's formula in [3]. Vemuru et al's formula deviates from the simulation results when the fanout is very small and when the fanout is greater than 3. On the other hand, the proposed formula reproduces the simulation results well.

The dependence of P_S on I_{D0N}, I_{DOP} and α is also compared between SPICE simulation and the present expression. Fig. 3.5 shows the short-circuit power dependence on PMOS and NMOS drivability ratio, β_r. Again the present formula reproduces the simulation results well. Fig. 3.6 shows the dependence on the MOSFET channel length, L_N and L_P. Since α is changed when the channel length is changed, Fig. 3.6 indicates the validity of the short-circuit power dependence on α_N and α_P of the current formula.

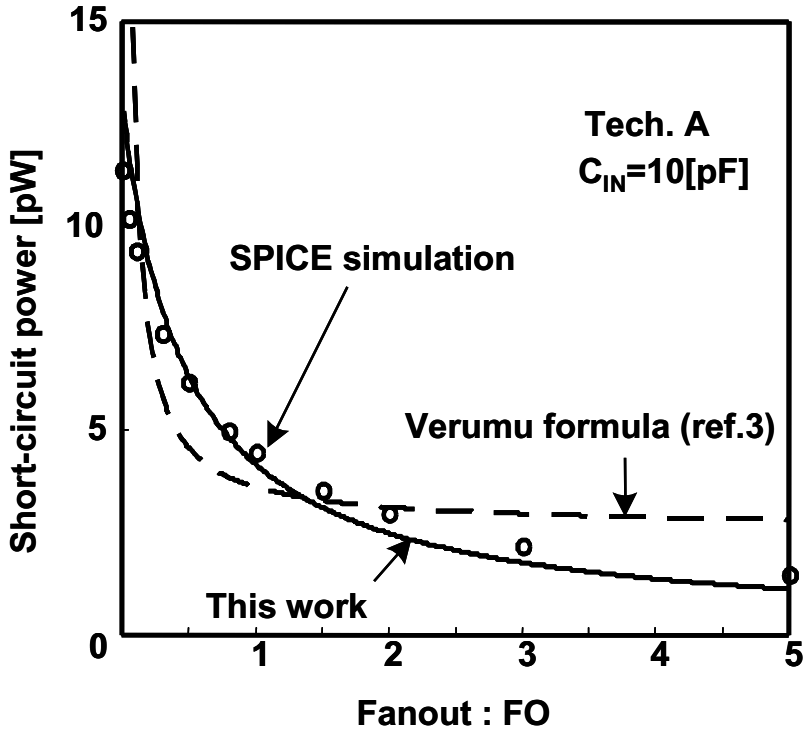


Fig. 3.4 Comparison between this work and previous published formula [3]

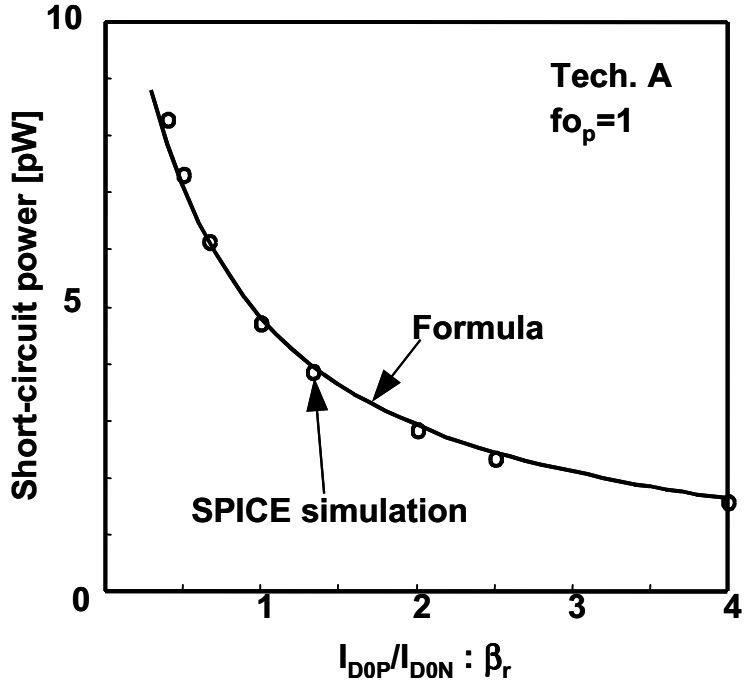


Fig. 3.5 Short-circuit power dependence on β_r.

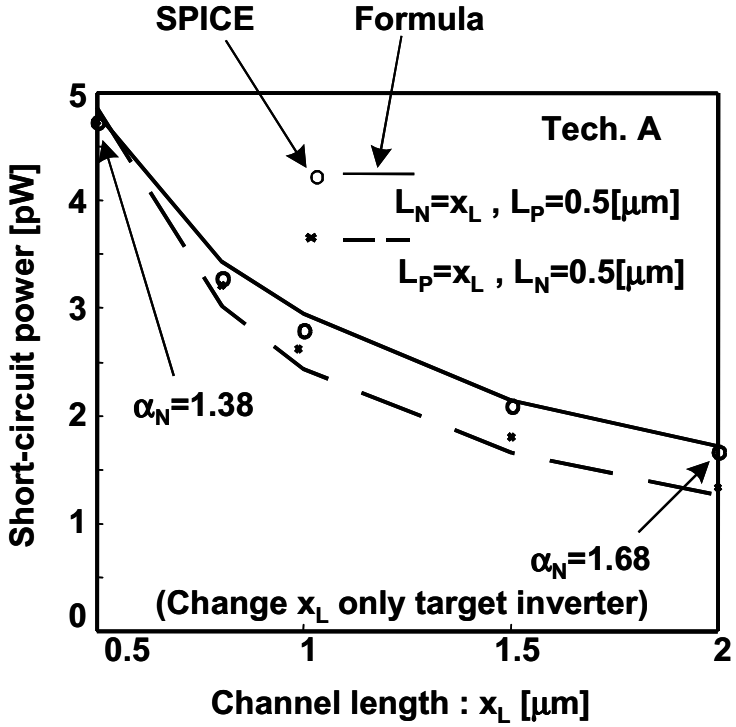


Fig. 3.6 Short-circuit power dependence on channel length

3.2.3 Short-Circuit Power Dissipation of Series-Connected MOSFET Structure

So far, the short-circuit power of only a CMOS inverter is considered. In this section, however, the more complicated structure, series-connected MOSFET structure, SCMS, which appears in NAND/NOR gates (see Fig. 3.7) is investigated. Here, in order to handle the SCMS, the idea in [6] is employed. In [6], in order to derive the delay of the SCMS, the *N* series-connected MOSFET is replaced by a single MOSFET structure, SMS (see Fig. 3.8). A method has been proposed to extract effective parameters, I_{D0NN} (effective I_{D0N} of the SMS), V_{D0}, and α for the SMS. This study follows the proposed method in [5] to extract I_{D0NN}, and V_{D0} for the SMS but the method to extract the effective α is modified.

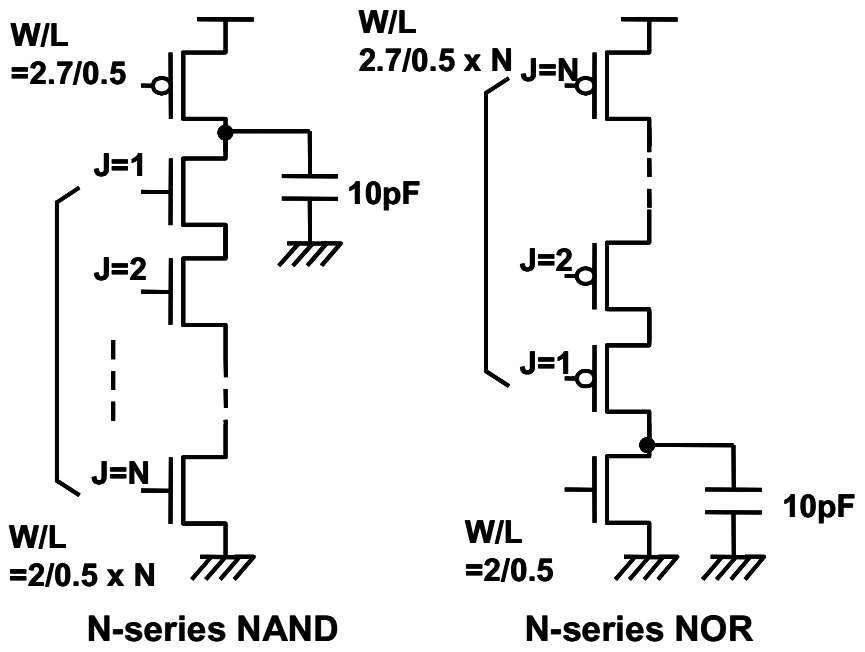


Fig. 3.7 *N* series-connected model structure (SCMS)

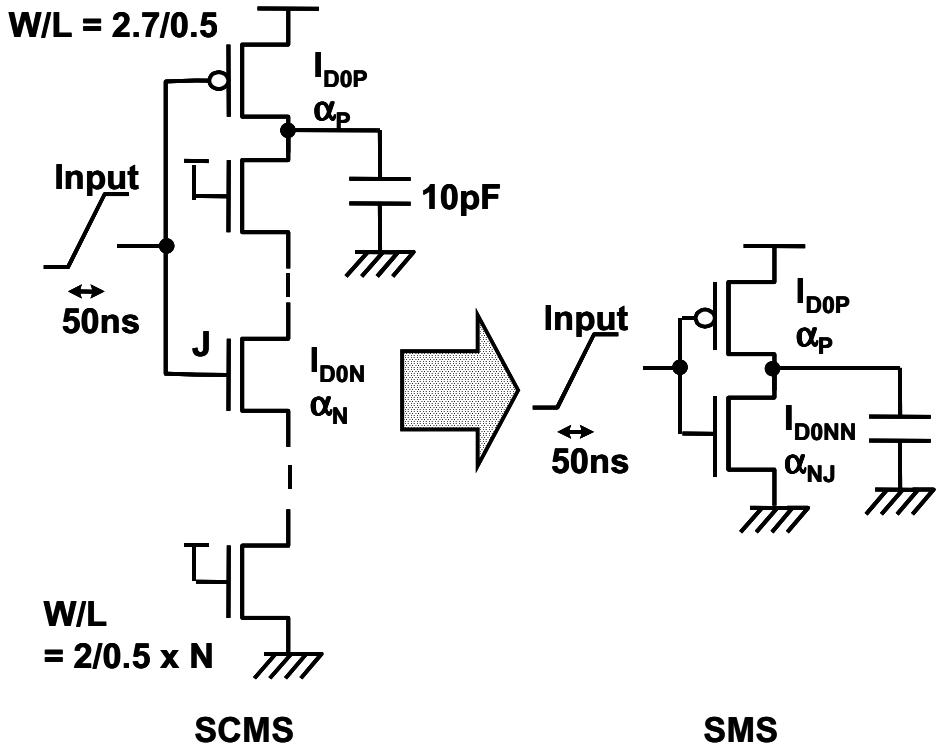


Fig. 3.8 Approximate *N* series-connected MOSFETs (SCMS) with single MOSFET structure (SMS)

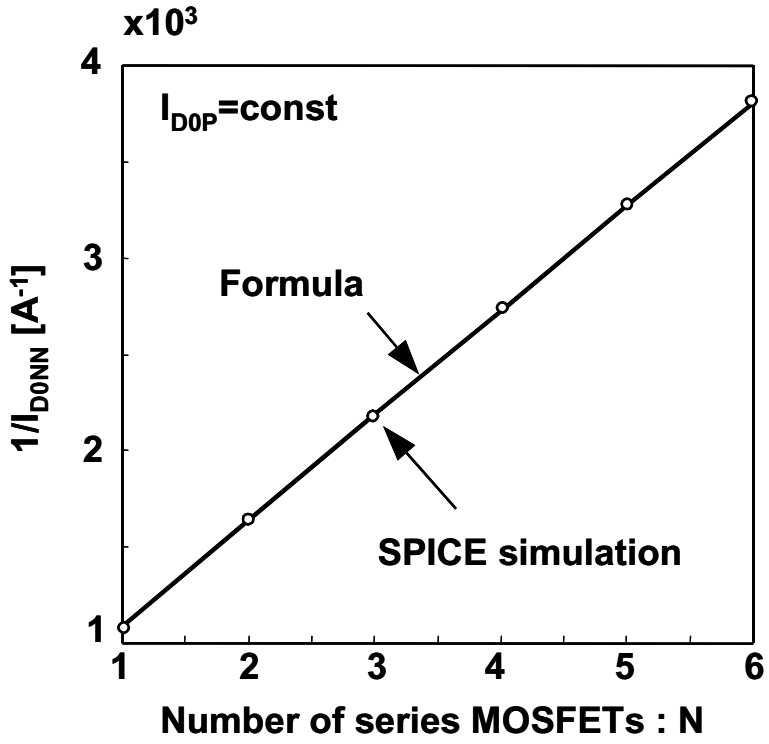


Fig. 3.9 $1/I_{D0NN}$ dependence on number of series MOSFETs

As is shown in [6], V_{D0} of the SMS is unchanged from the V_{D0} of one MOSFET in the SCMS, and I_{D0N} is calculated from I_{D0N1} and I_{D0N2} as follows:

$$I_{D0NN} = \frac{I_{D0N1} I_{D0N2}}{(I_{D0N1} - I_{D0N2})(N-1) + I_{D0N2}}. \qquad (3.15)$$

The calculated I_{DONN} is shown in Fig. 3.9 which shows good agreement with the simulation results. On the other hand, α is not so easy to approximate. In this study, a method to calculate $\alpha_{N(P)NJ}$ formula is proposed using simulated $\alpha_{N(P)11}$, $\alpha_{N(P)21}$, and $\alpha_{N(P)22}$. $\alpha_{N(P)NJ}$ is effective velocity saturation index of N series-connected N(P)MOS structure with J-th N(P)MOS gate from output as an input.

Scrutinizing the SPICE simulation results, the following empirical formulas can be used for the case of J=1 and J=N,

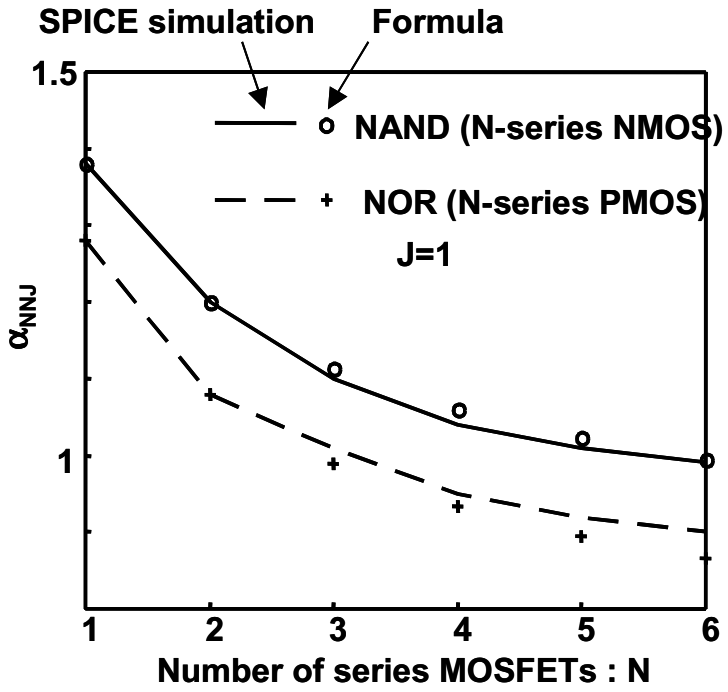
$$\alpha_{NN1} = \frac{\alpha_{N11}\alpha_{N21}}{(\alpha_{N11} - \alpha_{N21})\dfrac{\ln N}{\ln 2} + \alpha_{N21}}, \quad \alpha_{PN1} = \frac{\alpha_{P11}\alpha_{P21}}{(\alpha_{P11} - \alpha_{P21})\dfrac{\ln N}{\ln 2} + \alpha_{P21}} \qquad (3.16)$$

$$\alpha_{NNN} = \frac{\alpha_{N11}\alpha_{N22}}{(\alpha_{N11} - \alpha_{N22})(N-1) + \alpha_{N22}}, \quad \alpha_{PNN} = \frac{\alpha_{P11}\alpha_{P22}}{(\alpha_{P11} - \alpha_{P22})\dfrac{\ln N}{\ln 2} + \alpha_{P22}} . \qquad (3.17)$$

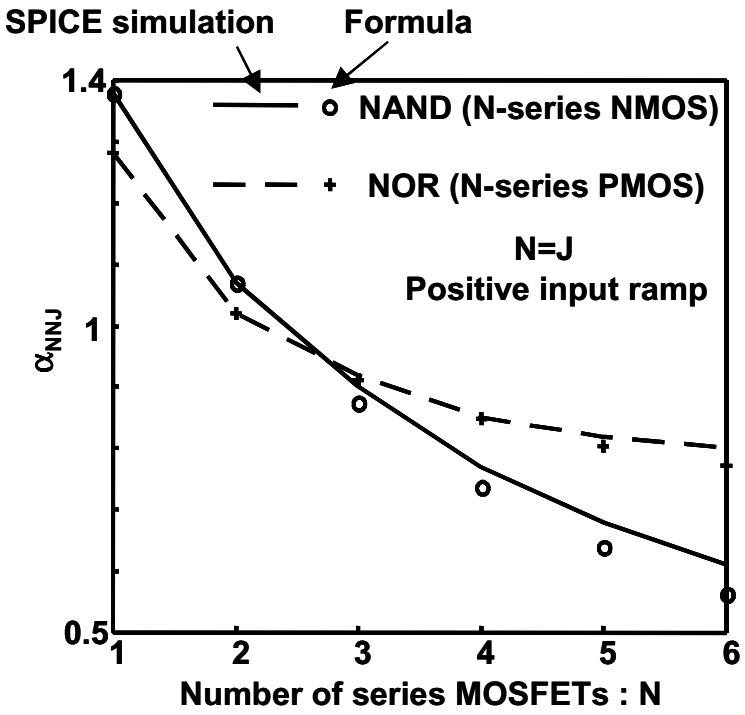
A comparison of the calculated α's with the simulation results is shown in Fig. 3.10(a) and (b).

Fig. 3.11 shows the short-circuit power comparison of the SCMS between the calculation and simulation. The calculation results can be favorably compared with the simulation. Once α_{NNJ} and α_{PNJ} are obtained, the general N and J can be obtained using the following formula.

$$\alpha_{NNJ} = \alpha_{NN1} + \frac{(\alpha_{NNN} - \alpha_{NN1})(J-1)}{(N-1)}, \quad \alpha_{PNJ} = \alpha_{PN1} + \frac{(\alpha_{PNN} - \alpha_{PN1})(J-1)}{(N-1)} . \qquad (3.18)$$

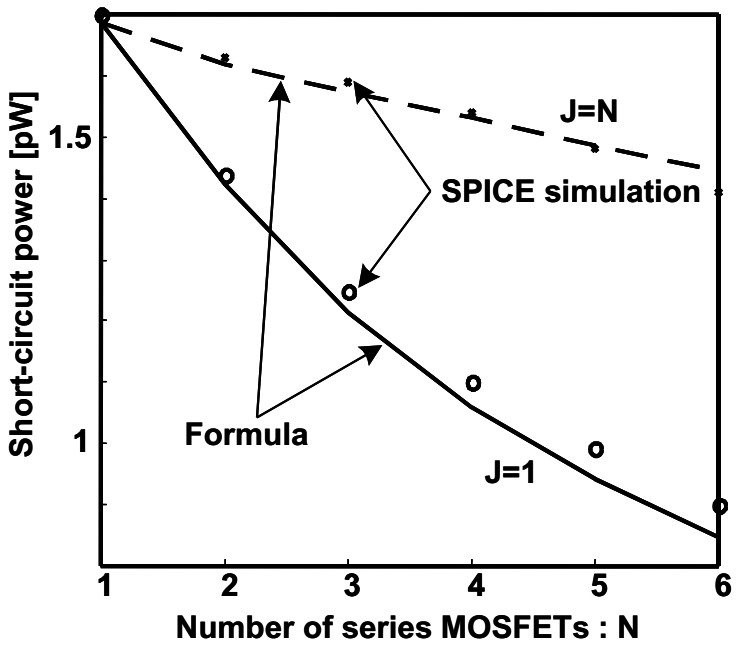


(a)

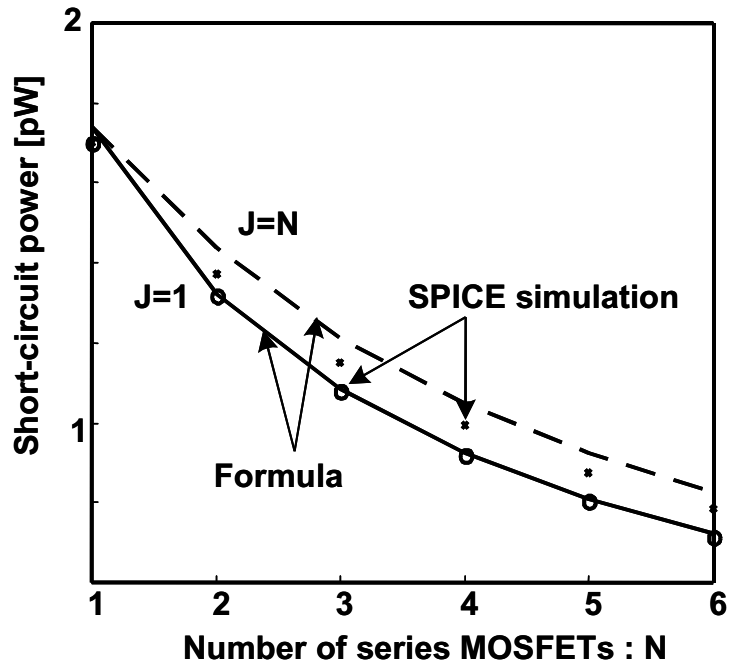


(b)

Fig. 3.10 α_{NNJ} dependence on number of series MOSFETs (a) J=1 (b) J=N



(a)



(b)

Fig. 3.11 Short-circuit power dependence on number of series MOSFETs

(a) *N*-series NAND (b) *N*-series NOR

3.2.4 Change of Short-Circuit Power Dissipation with Scaling

Now, let us consider the power ratio, $\eta_P = P_S / (P_D + P_S)$, to investigate the impact of the short-circuit power. It is straightforward to obtain the power ratio η_P knowing that P_D is expressed as $C_{OUT}V_{DD}^2/2$. Fig. 3.12 and Fig. 3.13 show comparisons of η_P between calculation and simulation. The dependence of η_P on the threshold voltage and the supply voltage is well reproduced over a wide range of V_{TH} and V_{DD} by the present formula. It is seen from the figures that $P_S / (P_D + P_S)$ is about 10% for a typical design. This means that the contribution of the short-circuit power to the total active power is about 10%.

Can η_P be changed over time? η_P is a function of α, fanout, V_{TH}/V_{DD} and I_{D0OUT}/I_{D0IN} as shown in the following formula.

$$\eta_P = \frac{P_S}{P_D + P_S} = \frac{1}{\dfrac{v_{D0P}\, g(v_T, \alpha)}{4k^2(v_{D0P})}\left(\dfrac{FO}{fo}\right)^2 \beta_r + \dfrac{h(v_T, \alpha)}{2k(v_{D0P})}\dfrac{FO}{fo} + 1}. \qquad (3.19)$$

The fanout and I_{D0OUT}/I_{D0IN} are essentially unchanged if the design style is unchanged even if the device is shrunk. α is not a strong function of a device scaling (see Fig. 3.14). It is shown from (3.19) that if V_{TH}/V_{DD} is constant, η_P remains constant even though the V_{DD} is scaled. In order to confirm the validity of this result, η_P dependence on V_{DD} scaling is shown in Fig. 3.15. Considering the tendency that V_{TH}/V_{DD} will be slightly increasing to keep the standby power in a tolerant level when the supply voltage is decreased as device miniaturization proceeds, the importance of the short-circuit power will not be increased (see Fig. 3.16).

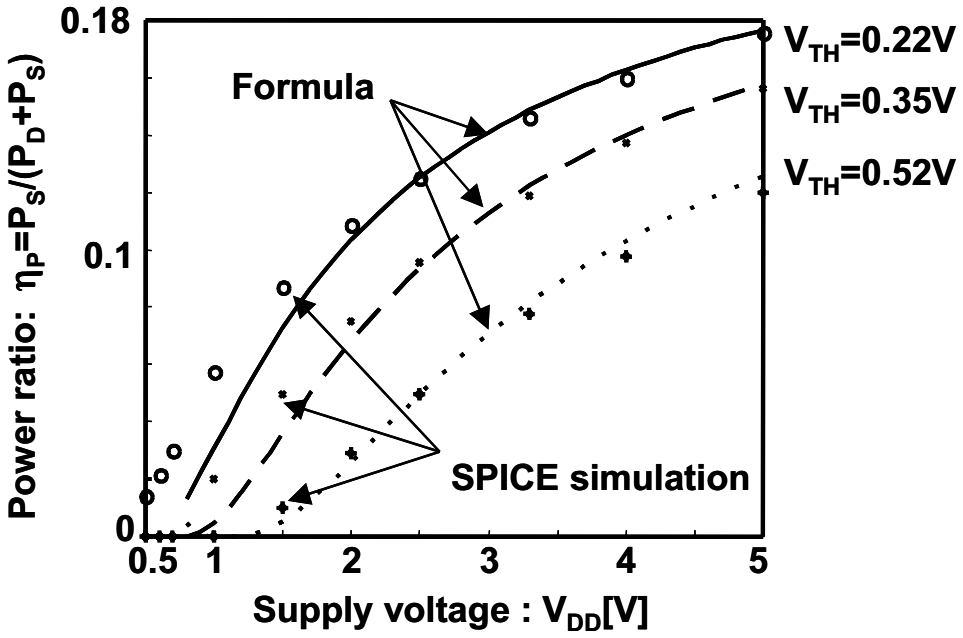


Fig. 3.12 Power ratio dependence on supply voltage

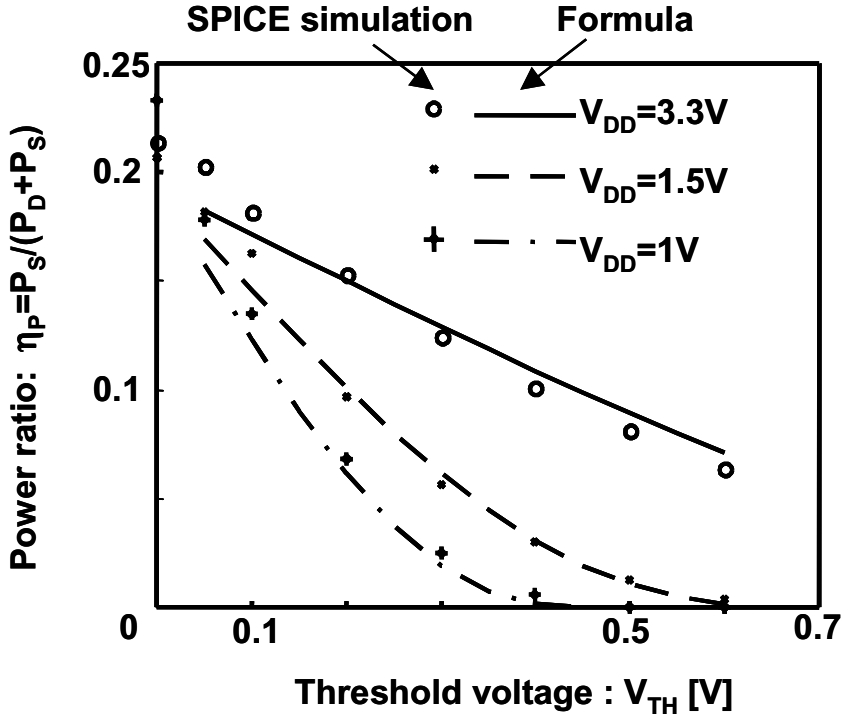


Fig. 3.13 Power ratio dependence on threshold voltage

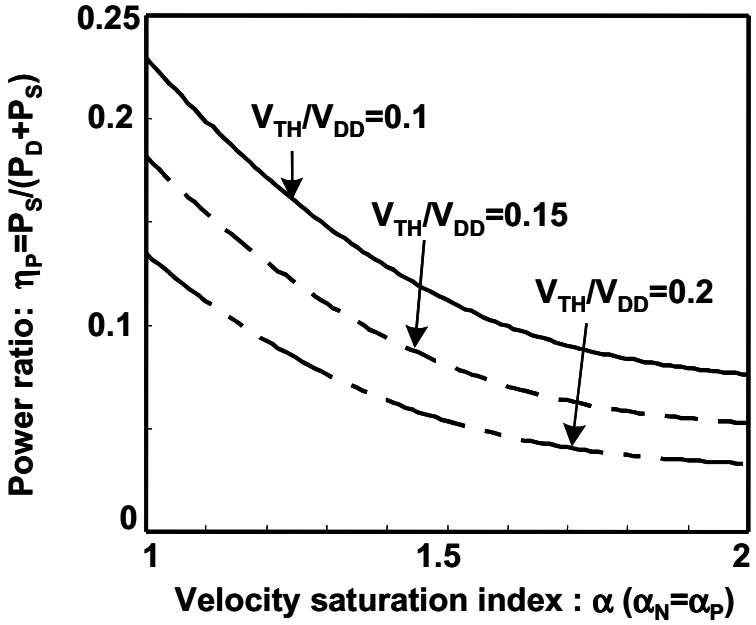


Fig. 3.14 Power ratio dependence on α

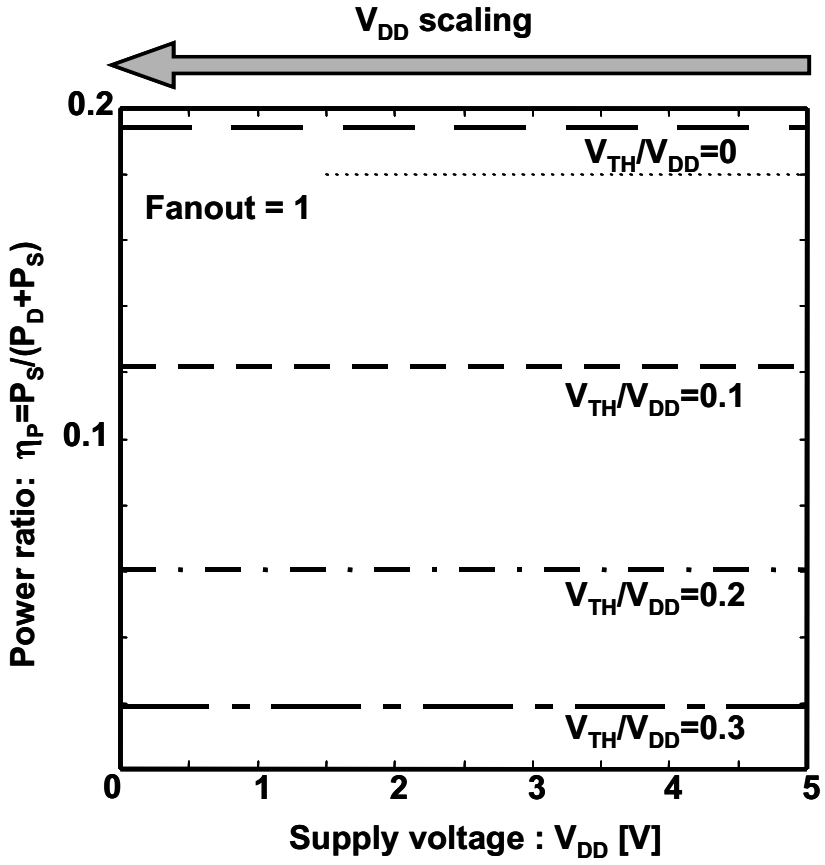


Fig. 3.15 Power ratio dependence on V_{DD} scaling

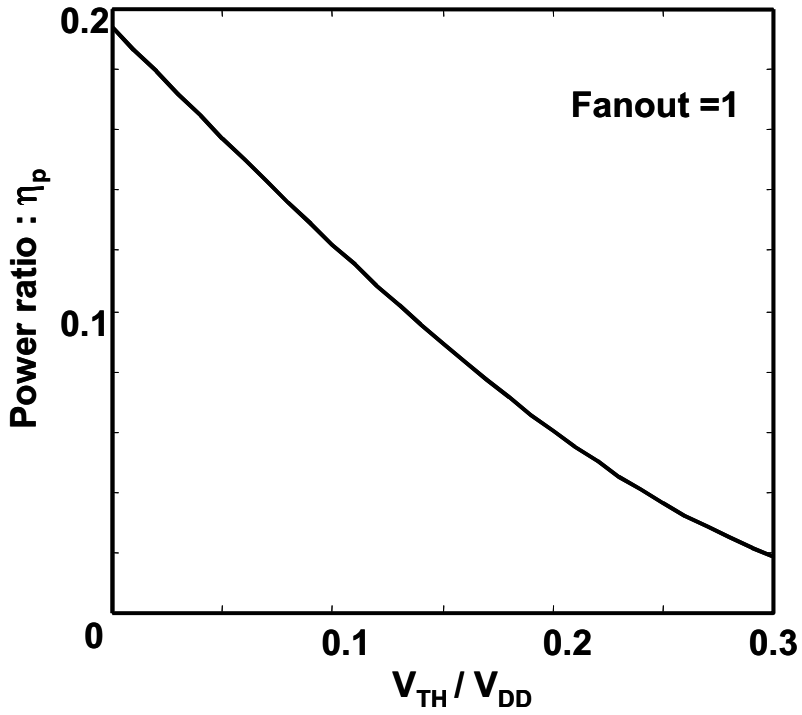


Fig. 3.16 Power ratio dependence on V_{TH}/V_{DD}

3.2.5 Simplified Formula for Short-Circuit Power

If the precision is of importance in estimating the short-circuit power, (3.10) is to be used but if the dependence on various parameters is of interest, the simpler expression is of use. In this Section, the simpler but less accurate formula is presented so as to give insight in the parametric dependence of the short-circuit power.

First, v_{D0} can be fixed at 0.5, v_{TN} and v_{TP} are both set equal to v_T and α_N and α_P are both set to α without much degradation in accuracy. Then, the following expressions are obtained.

$$P_S(t_T << \tau_N) = \frac{10}{3}\frac{(0.5-v_T)^3}{\alpha^2 2^{3v_T^2}} C_{IN} V_{DD}^2 \frac{fo}{FO\beta_r} \qquad (3.20)$$

$$P_S(t_T >> \tau_N) = \frac{(0.5 - v_T)^{3/2}}{10 \cdot 2^{3v_T + 2\alpha}} C_{IN} V_{DD}^2 fo \,. \qquad (3.21)$$

As is seen from Fig. 3.17, the relative error of the expression compared with the (3.2) and (3.4) is less than 20% in the range of $0 \le v_T \le 0.4$ and $1 \le \alpha \le 2$. When $v_T \ge 0.5$, the short-circuit current does not flow at all. It is easily seen from these formulas that the short-circuit power monotonically increases as α decreases, as fanout decreases and as the ratio of the threshold voltage over V_{DD} decreases.

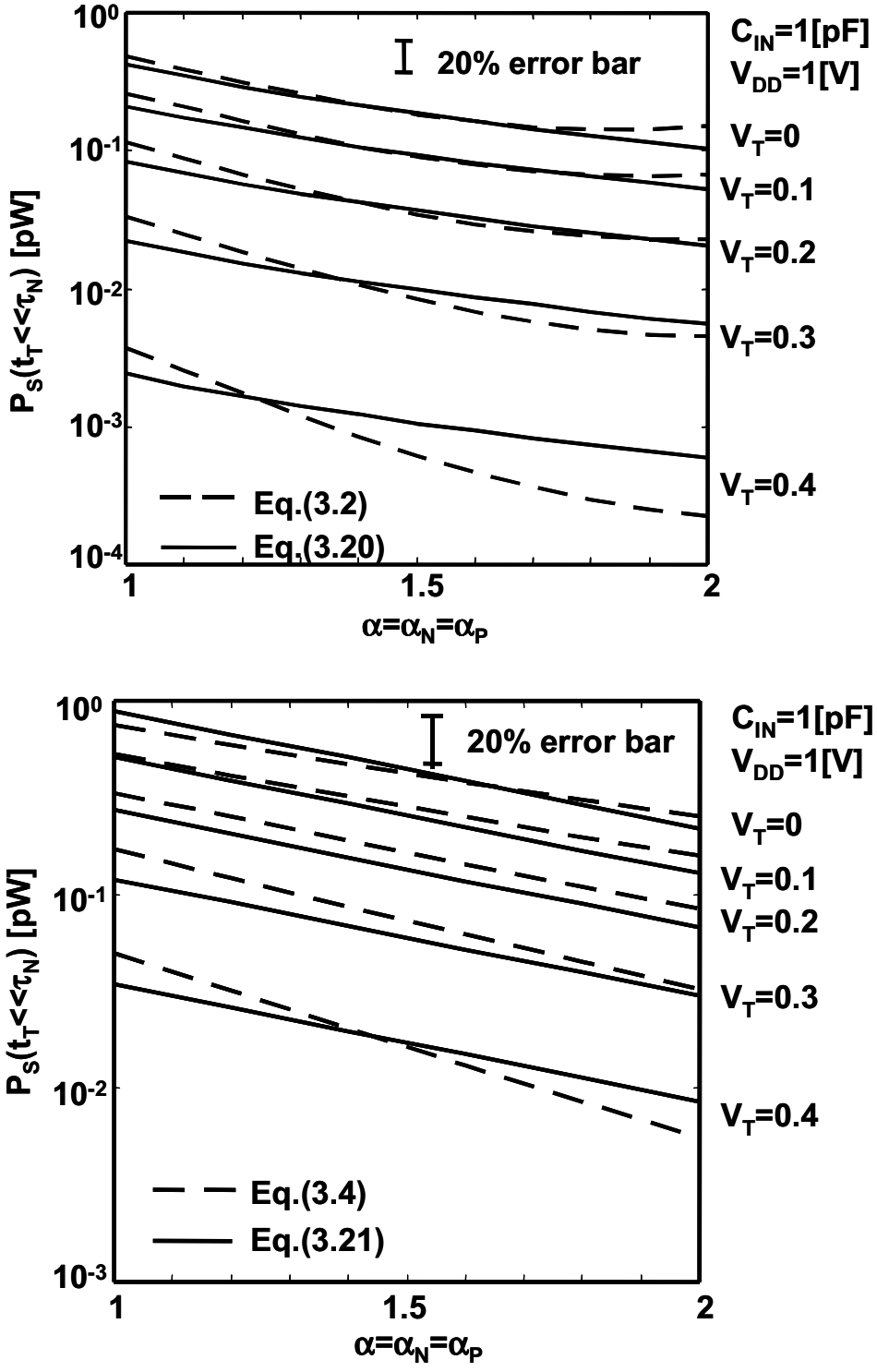


Fig. 3.17 Comparison between (3.2),(3.4) and simple formulas (3.20),(3.21)

3.3 Voltage Dependent Gate Capacitance and its Impact in Estimating Power and Delay of CMOS Digital Circuits with Low Supply Voltage

3.3.1 Voltage Dependent Capacitance of MOSFET

Gate capacitance seen from the input, C_G, is a function of terminal voltages as is shown in Fig. 3.18. C_G is not equal to C_{OX}, which is calculated from oxide thickness and is constant. In a subthreshold region, C_G is much smaller than C_{OX} and in an on-state, C_G is different between a linear region and a saturation region. If a CMOS inverter is formed, the input capacitance changes as in Fig. 3.19. In calculating the capacitance, the current flown into a gate terminal is integrated over time. It is obvious that the behavior of C_G changes depending on the threshold voltage. Since C_G is always smaller than C_{OX} and shows the minimum just before the threshold voltage, the effect of C_G is expected to decrease when V_{TH}/V_{DD} gets larger.

There is also a gate-drain overlap capacitance, C_{OV}, associated with a MOSFET. Since the overlap capacitance is not voltage dependent, it is not considered in this study. The overlap capacitance effect can be considered by just adding $2C_{OV}$ in an estimation process.

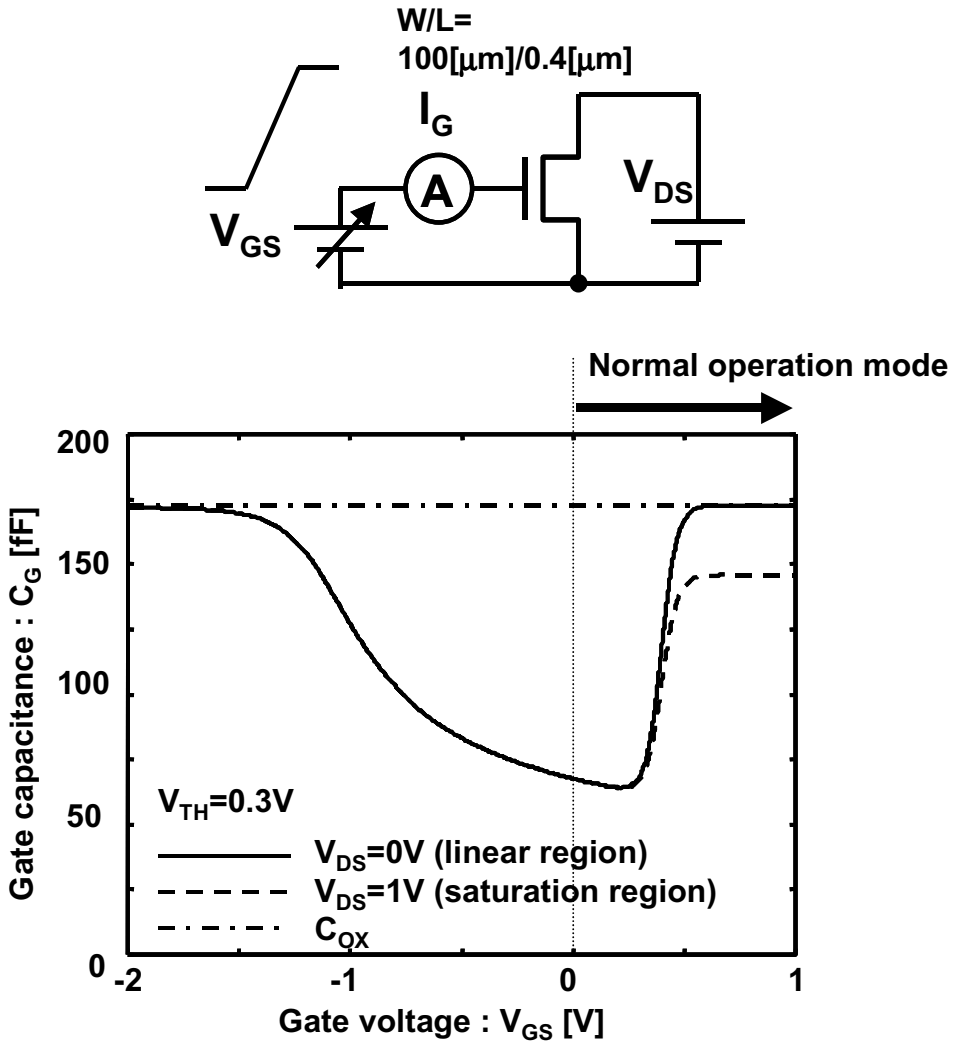


Fig. 3.18 Dependence of gate capacitance on gate and drain voltage

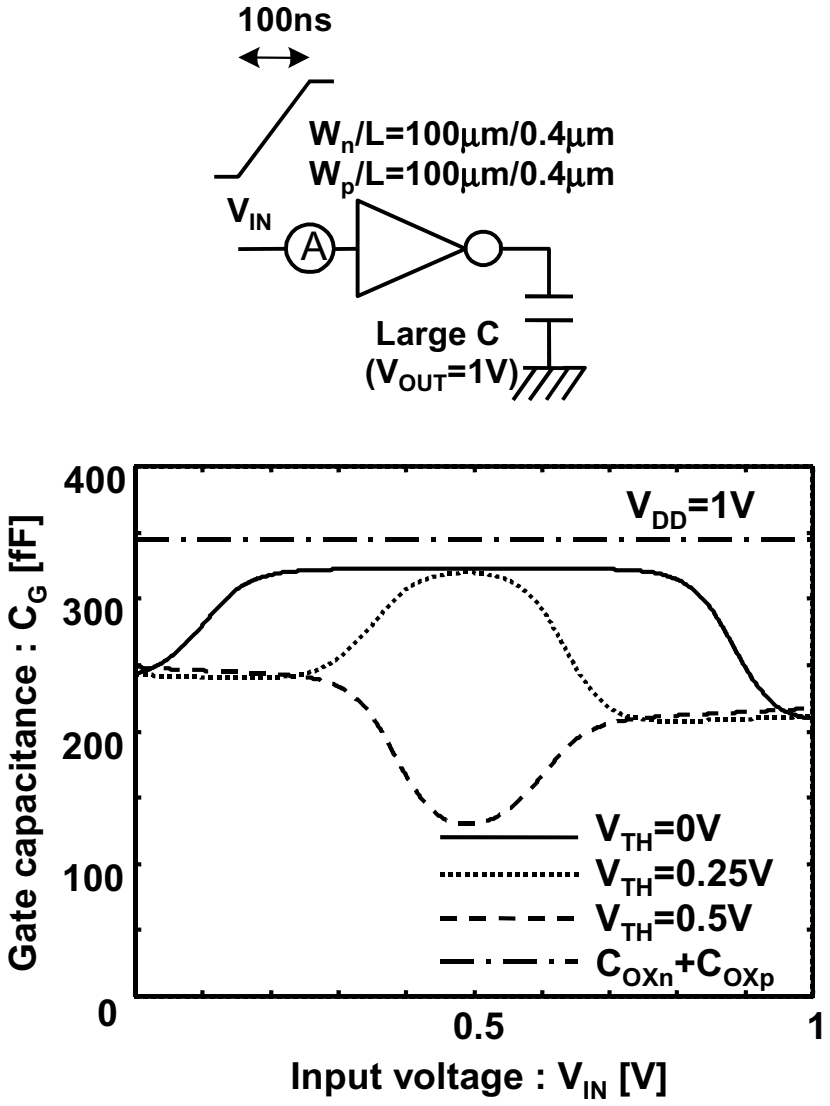


Fig. 3.19 Dependence of inverter input capacitance on gate voltage and threshold voltage

3.3.2 Definition of Effective Gate Capacitance

Let us consider an NMOS case for simplicity. An extension to a PMOS case is straightforward. Considering an inverter turning on, in an initial state, V_{GS} is 0 and V_{DS} is V_{DD} and V_{GS} reaches V_{DD} and V_{DS} reaches 0 at a final state. Considering this situation, let us define an effective gate capacitance, $C_{G,eff}$, as follows.

$$C_{G,eff} = \frac{\Delta Q_G}{V_{DD}} = \frac{1}{V_{DD}}\{Q_G(V_{GS}=V_{DD}, V_{DS}=0, V_{BS}=0) - Q_G(V_{GS}=0, V_{DS}=V_{DD}, V_{BS}=0)\} \quad (3.22)$$

Q_G is charge stored on a gate and ΔQ_G is gate charge difference between the final state and the initial state. This amount of charge should be poured into a gate terminal in circuit operation, which determines power and delay of digital circuits.

In calculating ΔQ_G, the current flown into a gate terminal can be integrated over time as is shown in Fig. 3.20. As is seen from the same figure, ΔQ_G is not path dependent so that any waveforms for V_{GS} and V_{DS} can be used to obtain ΔQ_G. An example of the extracted $C_{G,eff}$ is shown in Fig. 3.21. It is seen that $C_{G,eff}/C_{OX}$ becomes smaller in high V_{TH}/V_{DD} region.

It should be noted that $C_{G,eff}$ is defined for an NMOS and a PMOS transistor. Thus, the number of simulations needed to extract $C_{G,eff}$ for an LSI is limited to the number of kinds of transistors in a design, which is usually two or a little more for most digital designs. Input gate capacitance of a complex gate can be calculated by adding $C_{G,eff}$ of MOSFET's.

When $C_{G,eff}$ is to be calculated from device parameters, $C_{G,eff}$, can be calculated as $\Delta Q_G/V_{DD}$, using the following expressions [7].

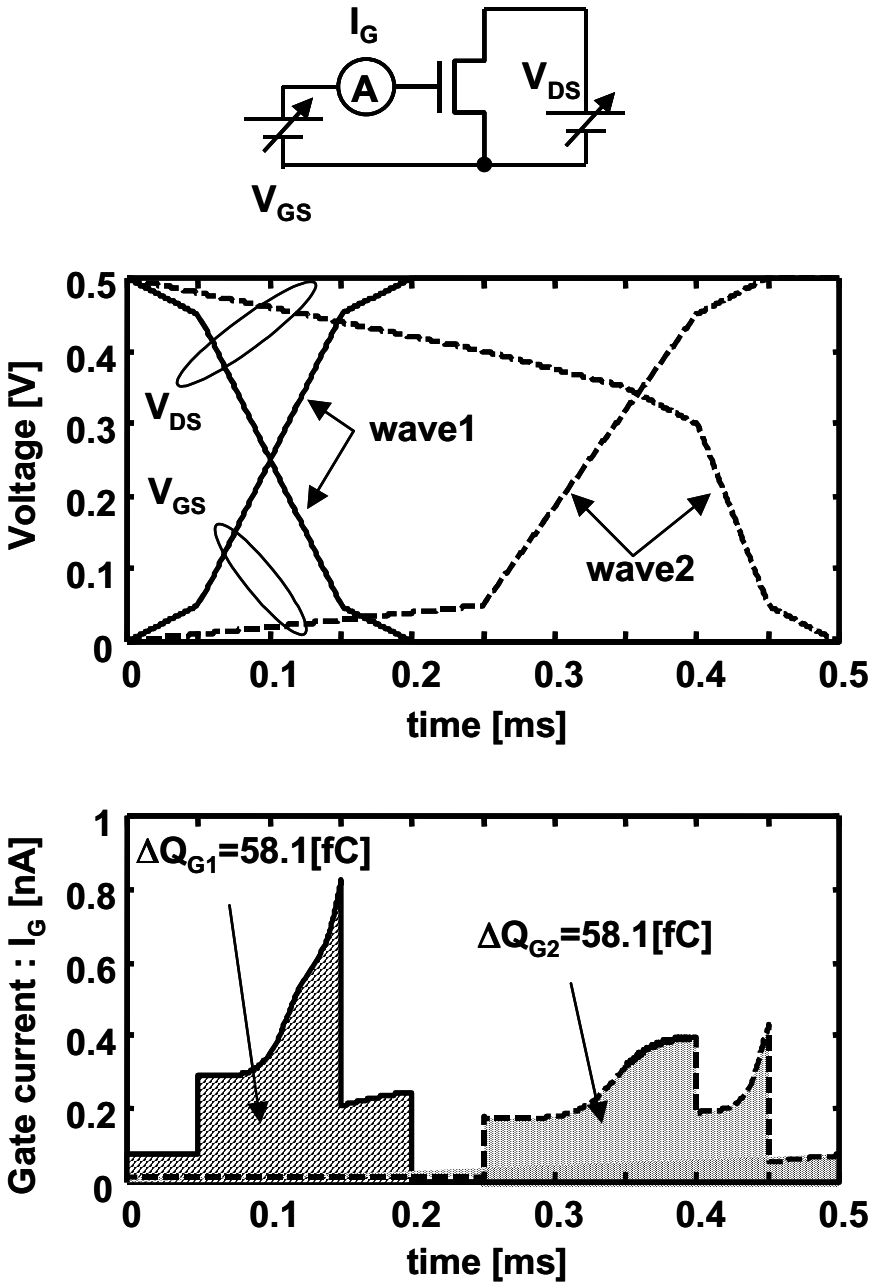


Fig. 3.20 Method to obtain $C_{G,eff}$ and ΔQ_G dependence on waveforms of gate voltage and drain voltage

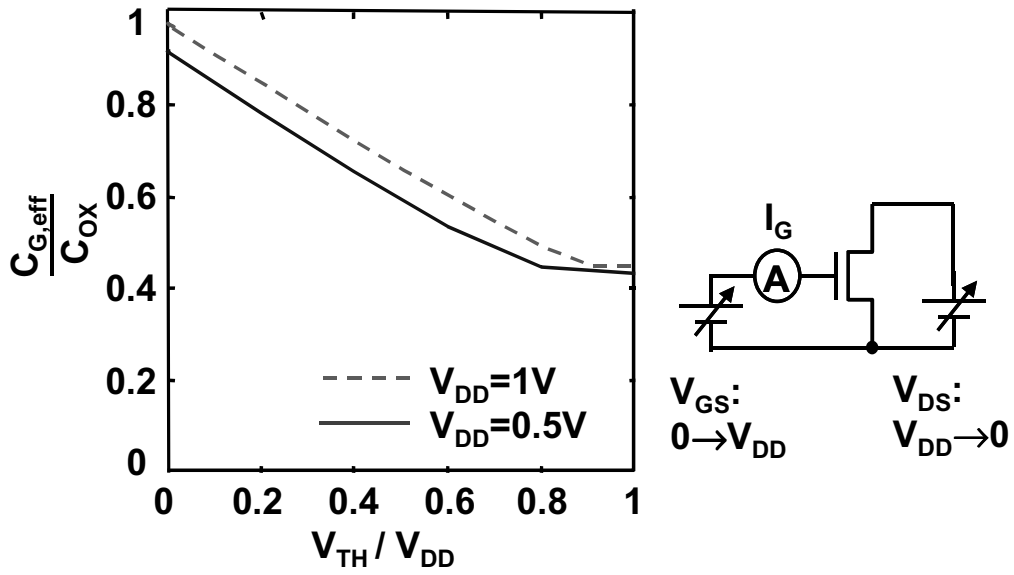


Fig. 3.21 Extracted effective gate capacitance, $C_{G,eff}$

$$\begin{aligned}\Delta Q_G &= \Delta Q_{GS} + \Delta Q_{GD} + \Delta Q_{GB} \\ &= LWC_{OX}\left\{V_{DD} - V_{FB} - 2\phi_F \right. \\ &\quad \left. - \frac{BE^2}{2}\left(-1 + \sqrt{1 + \frac{4(-V_{FB} - V_{BS})}{BE^2}}\right)\right\} \\ &\qquad (0 \le V_{TH} \le V_{DD})\end{aligned} \tag{3.23}$$

where

$$V_{FB} = V_{TH} - 2\phi_F - BE\sqrt{2\phi_F - V_{BS}} \quad : Flat\ band\ \ voltage \tag{3.24}$$

$$BE = t_{OX}\frac{\sqrt{2qN_A}}{\varepsilon\varepsilon_0}\cdot F_{CS} \quad : \ Body\ \ effect \tag{3.25}$$

$$\phi_F = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \quad : Fermi\ \ potential, \tag{3.26}$$

where F_{CS} is a charge-share factor and other notations follow a convention in device physics [7][8].

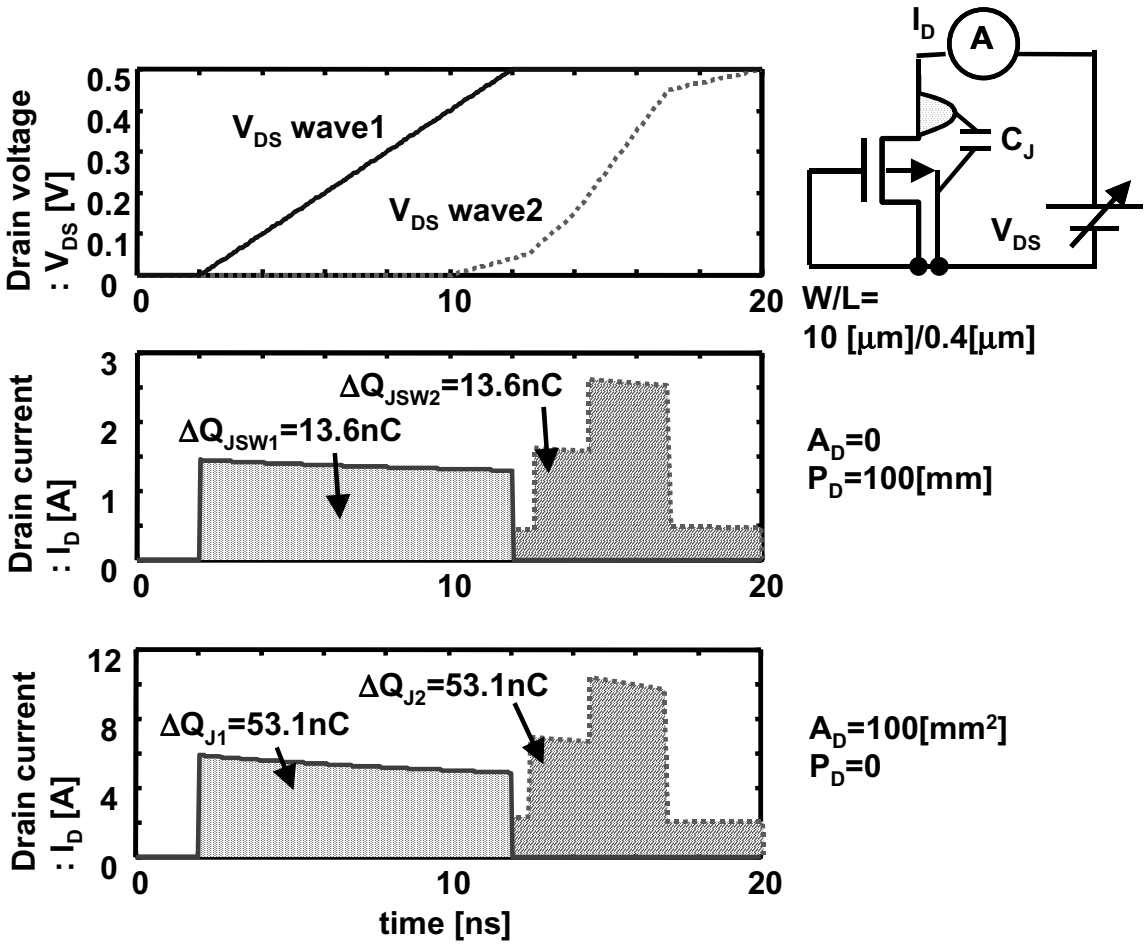


Fig. 3.22 Circuit to obtain $C_{J,eff}$ and ΔQ_J dependence on waveforms of gate voltage and drain voltage

Junction capacitance is also voltage dependent but it is a two-terminal device and the definition of the effective capacitance, $C_{J,eff}$, is trivial as follows.

$$C_{J,eff} = \frac{1}{V_{DD}} \int_0^{V_{DD}} C_J(V) dV = \frac{Q_J(V_{DD}) - Q_J(0)}{V_{DD}} = \frac{\Delta Q_J}{V_{DD}} \qquad (3.27)$$

A circuit in Fig. 3.22 can be used to obtain ΔQ_J by integrating the current flown into the junction terminal over time. It is not dependent on voltage wave shape (Fig. 3.22) and well-defined. In order to exclude C_{DG} effect which exists in parallel with the drain junction capacitance, very large drain junction (mm order) should be used compared with

the transistor size. Since junction capacitance has two components, that is, area dependent and periphery dependent component, two calculations should be carried out changing the area of drain and the periphery of drain [9].

3.3.3 Application of Effective Gate Capacitance

The effective gate capacitance, $C_{G,eff}$, is applied to estimate power and delay of a CMOS inverter in Fig. 3.23 and Fig. 3.24. Power and delay simulated by using constant C_{OX} and $C_{G,eff}$ as gate capacitance are denoted as $P(C_{OX})$, $t_d(C_{OX})$, $P(C_{G,eff})$, and $t_d(C_{G,eff})$, respectively. Power and delay simulated by using real MOS gate is denoted as P(MOS) and t_d(MOS), which are supposed to be true. Two different device models are used to check the effectiveness of the proposed $C_{G,eff}$. Both models are based on BSIM model and charge conservation in capacitance models is observed [8]. In order to concentrate on the gate capacitance effect, C_J and C_{INT} are set zero in the simulations.

P(MOS)/$P(C_{OX})$ and t_d(MOS)/$t_d(C_{OX})$ are less than 0.5 when V_{TH}/V_{DD} is above 0.6. This means that constant C_{OX} approximation for a gate capacitance becomes poor when V_{TH}/V_{DD} increases. The discrepancy is mainly due to the smaller capacitance in the subthreshold region. If we use $C_{G,eff}$ instead of C_{OX}, $P(C_{G,eff})$ and $t_d(C_{G,eff})$ can reproduce P(MOS) and t_d(MOS) well.

In order to check the validity of the $C_{G,eff}$ approximation, a more complex circuit, 4-bit counter, is analyzed. Again, simulations are carried out using $C_{G,eff}$, C_{OX} and real MOS gate for gate capacitances. Circuits shown in Fig. 3.25 are adopted to represent three cases. Each gate in a counter is substituted by one of the three types of gates. The results are shown in Fig. 3.26. In both power and delay comparison, $C_{G,eff}$ reproduce well the real gate for gate capacitance, while C_{OX} approximation gives larger power and delay by a factor of more than two.

Slight disagreement in power and delay between $C_{G,eff}$ approximation and the MOS gate simulation is due to the fact that the operation of MOSFET does not always start with V_{GS} =0 and $V_{DS} = V_{DD}$ and end with $V_{GS} = V_{DD}$ and $V_{DS} = 0$. This situation is observed in series connected MOS structures in NAND and other complex gates. The disagreement is also due to the substrate bias effect in the stacked structure. It can be said, however, that the disagreement is small and using $C_{G,eff}$ is much more accurate than to use C_{OX} as a constant capacitance in estimating power and delay.

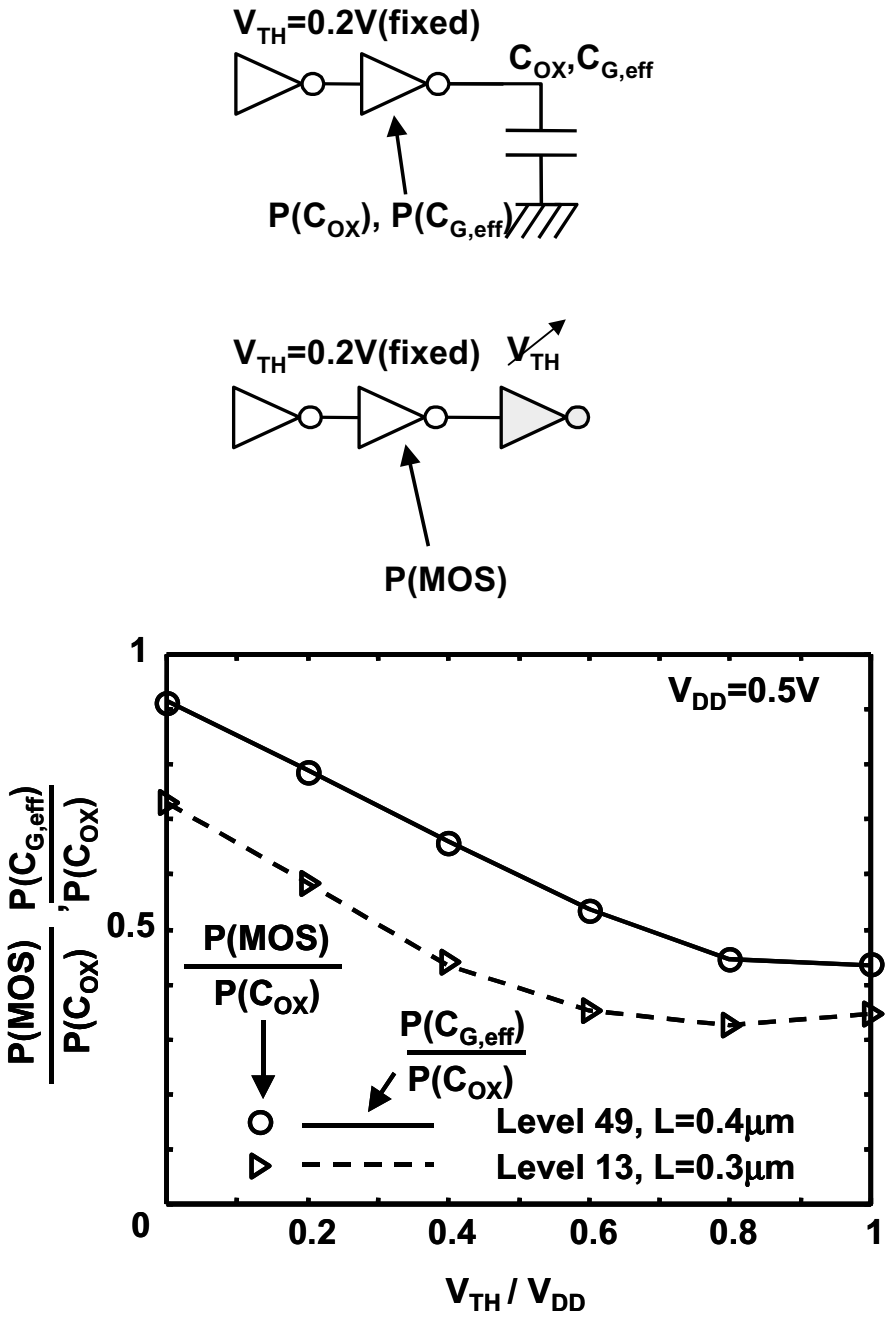


Fig. 3.23 Comparison of power estimated by using $C_{G,eff}$ ($P(C_{G,eff})$), C_{OX} ($P(C_{OX})$) and real MOS gate (P(MOS))

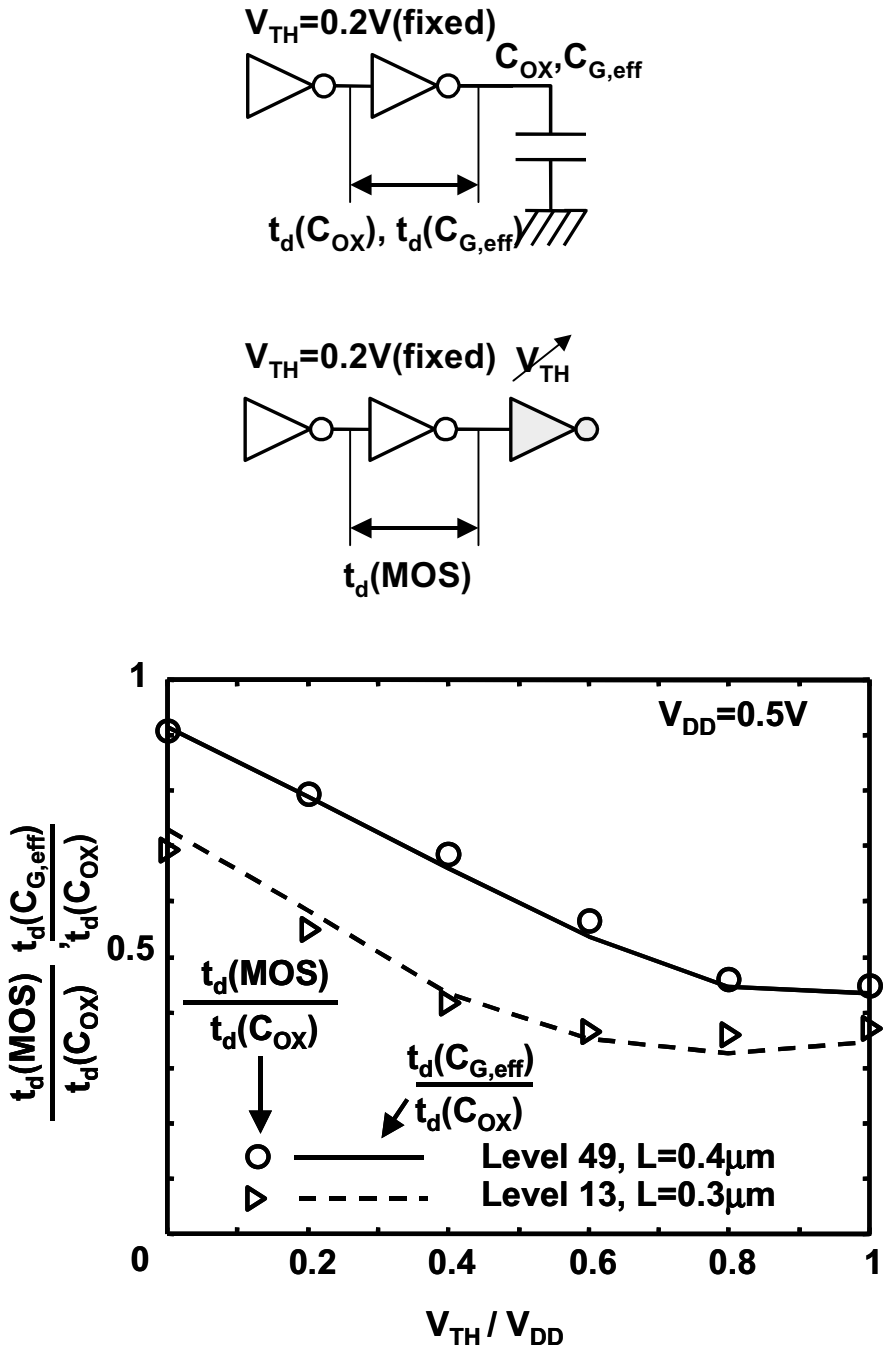


Fig. 3.24 Comparison of delay estimated by using $C_{G,eff}$ ($t_d(C_{G,eff})$), C_{OX} ($t_d(C_{OX})$) and real MOS gate (t_d(MOS))

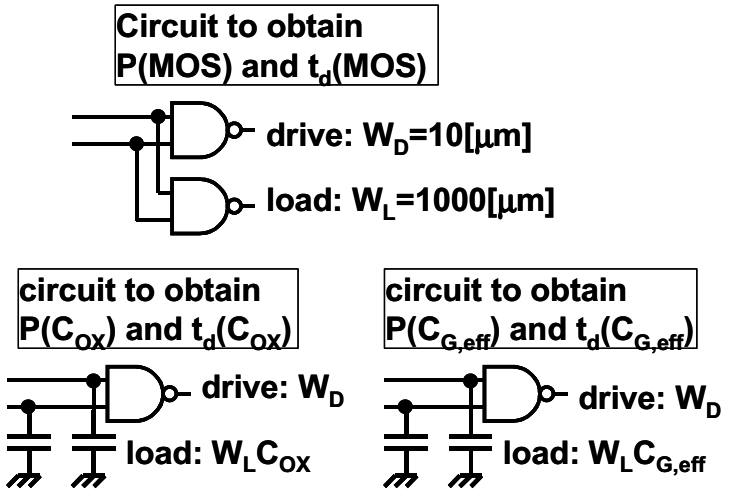


Fig. 3.25 Circuit to simulate effect of substituting real MOS gate capacitance by C_{OX} and C_{Geff}

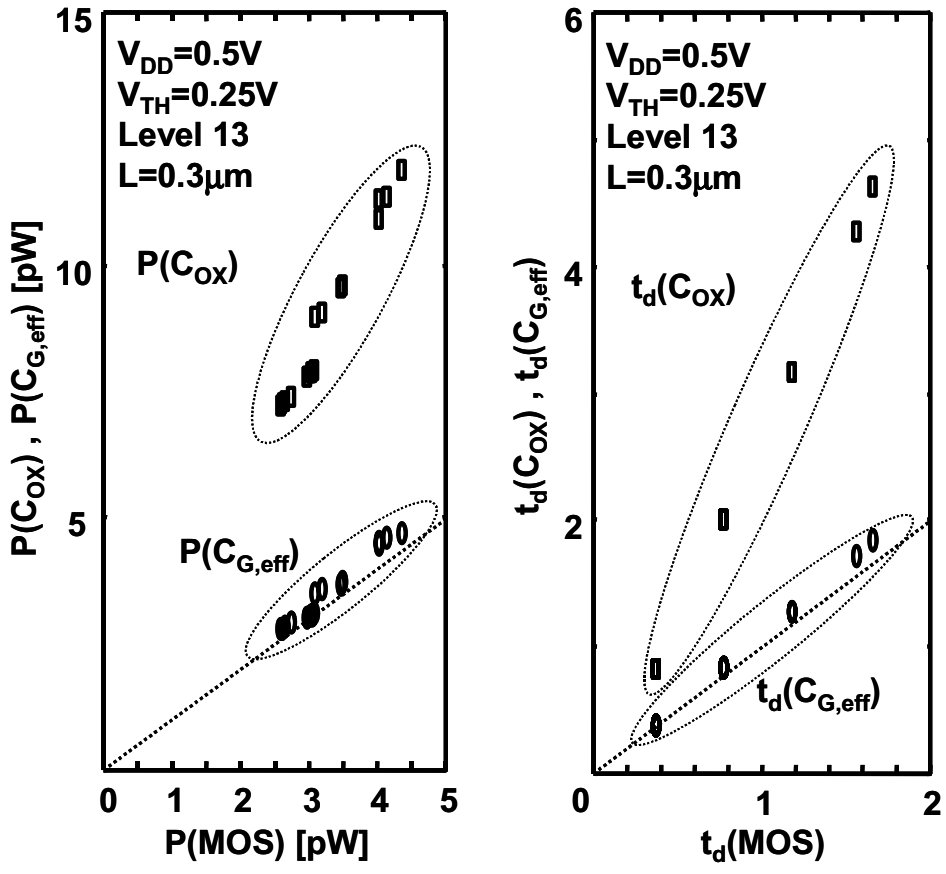


Fig. 3.26 Dynamic power dissipation and delay of 4-bit counter

3.3.4 Discussion

Supply voltage, V_{DD}, will be decreased in the future to cope with the power increase problem and to guarantee sufficient reliability. Low V_{DD} is also used for achieving low-power CMOS VLSI's. The threshold voltage, however, cannot be decreased with the same rate as V_{DD} decreases due to the exponential increase of subthreshold leakage. As a result, V_{TH}/V_{DD} tends to increase in the future and the discrepancy between $C_{G,eff}$ and C_{OX} gets bigger.

Although CAD tools take the voltage dependent capacitance effect correctly, designers use C_{OX} instead of $C_{G,eff}$ as an effective gate capacitance from time to time and it seems working well at present. This is because V_{TH}/V_{DD} is about 0.15 and the discrepancy between $C_{G,eff}$ and C_{OX} is about 10%, that is, small.

Moreover, although the power and delay are estimated a little larger than reality, this effect is being canceled out by neglecting short-circuit current component which tends to increase the delay and the power by about 10% (see Section 3.2). In low-voltage designs, however, V_{TH}/V_{DD} becomes larger and the short-circuit current tends to diminish while the discrepancy between $C_{G,eff}$ and C_{OX} tends to increase. Then the cancellation does not take place. Consequently, the constant capacitance approximation using C_{OX} becomes less and less accurate and $C_{G,eff}$ should be used instead in the future.

C_{INT} is dominant in C_{LOAD} in many cases, and in that situation, the accuracy of the gate capacitance approximation is less important but there are cases where C_{INT} is small and gate capacitance affects the circuit behavior much like in some hand crafted data-path circuits.

3.4 Optimization of V_{DD} and V_{TH} for Low-Power and High-Speed Applications

3.4.1 Problems of Previous V_{DD}-V_{TH} Optimization Methods

In order to minimize the power dissipation, V_{DD}-V_{TH} optimization has been investigated extensively but previous publications on V_{DD}-V_{TH} optimization have following three problems.

First, Energy-Delay product (ED product) has been often used as an object function in optimizing CMOS circuit power consumption [10][11][12]. In practice, however, the objective of the optimization is to minimize the power consumption while satisfying a speed constraint. When we take the ED product as an object function, we get only one pair of the optimized V_{DD} and optimized V_{TH} if the technology is fixed. This is not what we want, since the optimized V_{DD} and V_{TH} should be different if the target circuit speed is different. In this section, the optimization is carried out taking the power as an object function and the speed as a constraint to make the optimization results more practical.

The second issue is on the drain current modeling of MOSFET's. Fig. 3.27 shows a comparison between the present model and the previous model that has been used in power optimization papers [10][11]. It is seen that the previous drain current model has discontinuity around the V_{TH} while the present model rectifies the issue, details of which is discussed in the text.

The last problem is that the previous calculation has not considered the effects of both V_{TH} fluctuation and temperature variation. Since these effects are getting more important in the deep submicron region, the analysis should take these effects into account.

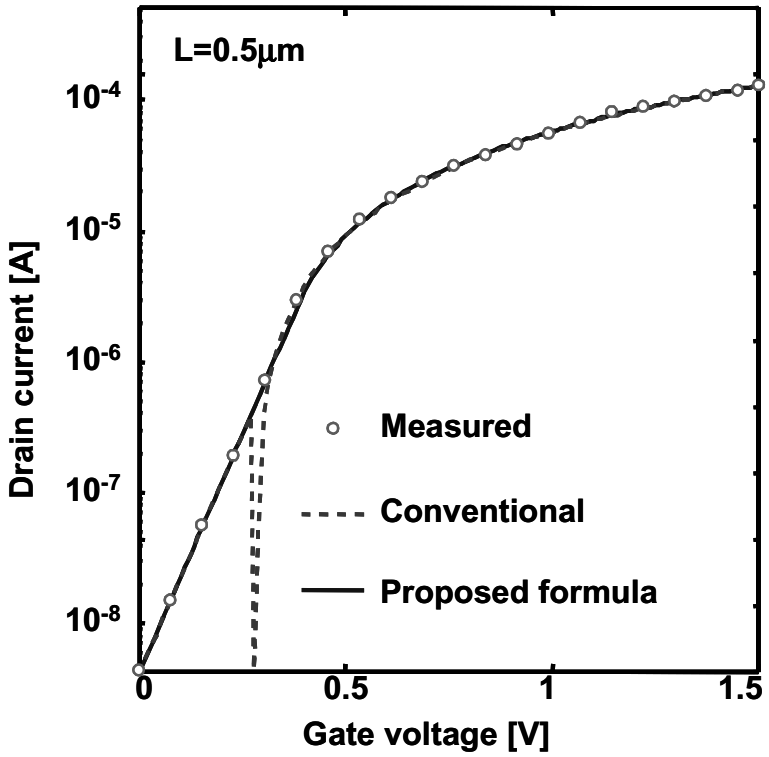


Fig. 3.27 Drain current models used in power optimization

In this section, closed-form formulas are presented for optimum V_{DD} and V_{TH} that minimize power dissipation when the technology and required speed are given. Above-mentioned problems are eliminated in the analysis. $V_{TH,min}$ is considered in this study to incorporate V_{TH} fluctuation effects. The resultant formulas have been applied to the technology roadmap to discuss the future VLSI design trend.

3.4.2 Closed-Form Formulas for Optimum V_{DD} and V_{TH}

A new drain current model for short-channel MOSFET's is proposed that provides smooth transition across subthreshold region and above-threshold region. By using the model, accurate calculation of power and delay near the threshold is possible. The model is described as the following expressions.

Table 3.2 Notations used in Section 3.4

notation	meaning
a	switching activity
L_d	logic depth of critical path
f	given clock frequency
C_L	load capacitance
α	velocity saturation index [5]
I_0	drain current when V_{GS}=V_{TH} at lowest temperature
T_{min}	lowest operation temperature
T_{max}	highest operation temperature
ΔT	T_{max}-T_{min}
N_S	nkT_{max}/q (n: subthreshold slope factor)
K	coefficient of delay
ΔV_{TH}	peak-to-peak V_{TH} variation through process
κ	temperature coefficient of V_{TH}
$V_{TH,max}$	highest V_{TH} in operation temp. and process variation range
$V_{TH,min}$	lowest V_{TH} in operation temp. and process variation range
V_{DDopt}	optimum V_{DD}
V_{THopt}	optimum $V_{TH,min}$
$I_{ON,\ min}$	drain current when V_{GS}=V_{DD} at lowest temp. and highest V_{TH} corner in process variation
$I_{OFF,\ max}$	leakage current at highest temp. and lowest V_{TH} corner in process variation
$P_{LEAK,\ max}$	leakage power at highest temp. and lowest V_{TH} corner in process variation

$$I_D = \begin{cases} I_0 e^{\alpha\left(\frac{V_{GS}-V_{TH}}{\alpha N_S}\right)^{\alpha}} & (V_{GS} \geq V_{TH} + \alpha N_S) \\ I_0 e^{\frac{V_{GS}-V_{TH}}{N_S}} & (V_{GS} \leq V_{TH} + \alpha N_S) \end{cases} \quad (3.28)$$

The notations for these formula as well as the notations for other quantities used in this study are tabulated in Table 3.2.

Fig. 3.27 shows a comparison between the proposed model and the conventional model [10]. The previous drain current model has discontinuity around V_{TH} and the present model does not have one. The difference between the proposed formula and the

measured result is within 4% when V_{GS}=0~1.5V.

Here, as a basis of optimization, the delay and the power dissipation models are explained that take into consideration the V_{TH} variation through process and temperature. The two main sources of power dissipation in CMOS VLSI's are the dynamic power dissipation due to charging and discharging of load capacitance, and the power dissipation due to subthreshold leakage. There may be short-circuit power dissipation as the third source of power dissipation but it is less than 10% in total power dissipation (see Section 3.2). Then, the short-circuit power is neglected in this study. The voltage dependency of the load capacitance is discussed in Section 3.3. The result shows that the variation of load capacitance is about 30% when V_{TH} changes from 0 to V_{DD}/2. The variation, however, decrease to about 15% when the junction capacitance and overlap capacitance are taken into account. Hence, the variation of the load capacitance is neglected.

The main device parameters that depend on the temperature are mobility, μ, V_{TH}, and subthreshold slope, N_S. The temperature dependence of these parameters are written as [13]

$$\mu = \mu' \cdot \left(\frac{T_{\max}}{T_{\min}} \right)^{-m} \qquad (3.29)$$

$$V_{TH,\min} = V_{TH,\max} - \Delta V_{TH} - \kappa \Delta T \qquad (3.30)$$

$$N_S = N_S' \cdot \frac{T_{\max}}{T_{\min}} = \frac{nkT_{\max}}{q}, \qquad (3.31)$$

where μ' and N_S' are the mobility and the subthreshold slope at the lowest temperature in use, T_{min}, respectively. $V_{TH,max}$ and $V_{TH,min}$ are the maximum and minimum threshold voltage under the temperature and process fluctuation. κ is a temperature coefficient of V_{TH}, which is typically 2.4mV/K in 0.5μm process, and m is a temperature exponent of mobility whose typical value is 1.5. Fig. 3.28 shows the temperature dependence of drain current.

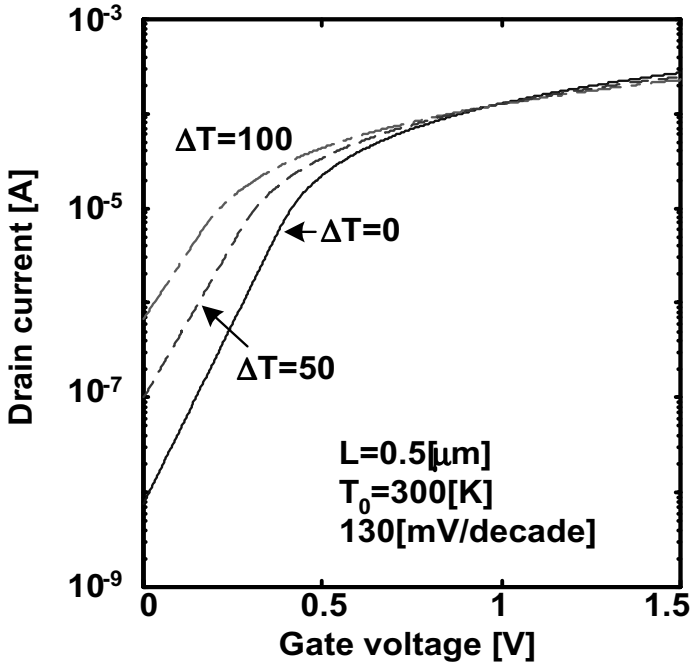


Fig. 3.28 Temperature characteristics of MOSFET

It is seen that, in sub-1V region, CMOS circuits show positive temperature dependence, because the effect caused by V_{TH} lowering is stronger than the effect caused mobility degradation [14][15]. If our interest is in sub-1V region, the worst-case delay occurs at the lowest operation temperature. The delay of interest is written as

$$t_d = K\frac{C_L V_{DD}}{\beta(V_{DD} - V_{TH,\max})^{\alpha}}, \qquad (3.32)$$

where

$$\beta = I_0\left(\frac{e}{\alpha N_S'}\right)^{\alpha}. \qquad (3.33)$$

On the other hand, the worst power consumption is observed at the highest operation temperature, because the dynamic power component, P_D, which is written as

$$P_D = afC_L V_{DD}^2 \qquad (3.34)$$

does not have temperature dependence and the main temperature dependence comes from the leakage component. The leakage component also increases when V_{TH} is lowered by V_{TH} fluctuation. Therefore, the maximum leakage current appears when the threshold voltage is $V_{TH,min}$. Consequently the maximum leakage power, $P_{LEAK,max}$ is written as

$$P_{LEAK,\max} = I_0 e^{\frac{-V_{TH,\min}}{N_S}} V_{DD} . \quad (3.35)$$

The frequency is expressed using t_d (3.32) and the logic depth of a critical path, L_d.

$$f = \frac{1}{L_d \cdot t_d} . \quad (3.36)$$

(3.34), (3.35) and (3.36) are the basic equations for the power optimization. Now we try to solve the equation system. First, by solving (3.36) in terms of $V_{TH,min}$, we get

$$\begin{aligned} V_{TH,\min} &= V_{DD} - \left(\frac{f L_d K C_L}{\beta} \right)^{1/\alpha} V_{DD}^{1/\alpha} - \Delta V_{TH} - \kappa \Delta T \\ &= V_{DD} - \chi V_{DD}^{1/\alpha} - \Delta V_{TH} - \kappa \Delta T, \end{aligned} \quad (3.37)$$

where $\chi=(fL_dKC_L/\beta)^{1/\alpha}$.

Substituting (3.37) in (3.34) and (3.35) the formula of power dissipation can be derived, which is denoted as $P(V_{DD})$. In order to obtain V_{DDopt} and V_{THopt} when the clock frequency is given, we differentiate $P(V_{DD})$ with respect to V_{DD} and set the resultant expression to zero. The resulting equation is transcendental and cannot be solved exactly. Here we can assume $V_{DD}>>N_S$, since N_S is normally less than 0.05V. Then, the equation becomes as follows.

$$V_{DD} - \chi V_{DD}^{1/\alpha} = -N_S \ln\left(\frac{2afC_L N_S}{I_0} \frac{\alpha}{\alpha - \chi} \right) + \Delta V_{TH} + \kappa \Delta T \quad (3.38)$$

Still the above equation cannot be solved for V_{DD} analytically, but optimum $V_{TH,min}$, which is denoted as V_{THopt}, can be calculated using (3.37) and (3.38) easily.

$$V_{THopt} = -N_S \ln\left(\frac{2afC_L N_S}{I_0}\frac{\alpha}{\alpha-\chi}\right) \quad (\alpha>\chi)\,, \tag{3.39}$$

where

$$\chi = \left(\frac{fL_d C_L K}{\beta}\right)^{1/\alpha}, \quad N_S = \frac{nkT_{\max}}{q}. \tag{3.40}$$

As is described above, it is difficult to solve V_{DDopt}. Some approximations are used. By using Taylor expansion of the equation around V_{DD}=1, V_{DDopt} can be solved as

$$V_{DDopt} = \frac{-N_S \ln\left(\frac{2afC_L N_S}{I_0}\frac{\alpha}{\alpha-\chi}\right) + \Delta V_{TH} + \kappa\Delta T + \frac{\alpha-1}{\alpha}\chi}{1-\frac{\chi}{\alpha}} \quad (\alpha>\chi). \tag{3.41}$$

(3.39) and (3.41) are the optimum V_{DD} and V_{TH}.

Let us make the simpler guideline for the power optimization. This is possible by using either the ratio between $P_{LEAK,max}$ and P_D or the ratio between $I_{ON,min}$ and $I_{OFF,max}$. $P_{LEAK,max}$, $I_{ON,min}$ and $I_{OFF,max}$ are defined in Table 3.2. Using (3.34) and (3.35), the ratio of $P_{LEAK,max}/P_D$ can be expressed as

$$\frac{P_{LEAK,\max}}{P_D} = \frac{2}{\frac{V_{DDopt}}{N_S} - 1 - \frac{V_{DDopt} - V_{THopt}}{\alpha N_S}}, \tag{3.42}$$

where $P_{LEAK,max}$ is leakage power dissipation at the highest temperature and at the lowest V_{TH} corner in process variation. If we confine V_{DD} around 1V (0.5V~1.5V) and V_{THopt}<<1, the ratio can be simplified as

$$\frac{P_{LEAK,\max}}{P_D} = \frac{2N_S\alpha}{\alpha-1} \quad (\alpha>1.1). \tag{3.43}$$

In terms of $I_{ON,min}$ and $I_{OFF,max}$, it is rewritten as

$$\frac{I_{ON,\min}}{I_{OFF,\max}} = K\frac{\alpha-1}{2N_S\alpha}\frac{L_d}{a} \quad (\alpha>1.1). \qquad (3.44)$$

Assuming typical values for the parameters such that N_S=0.048 (S-factor at T_{min} is 80mV/decade and T_{max}=400K) and α=1.3, $P_{LEAK,max}$ is calculated to be about 30% of the total power dissipation. This value of about 30% is not changed over a wide range of design parameters such as a, L_d and f. This is understood like below. When the target speed is changed, V_{THopt} changes slightly but V_{DDopt} changes much because V_{TH} changes the power exponentially while the dependence of power on V_{DD} is quadric. The amount of change in V_{TH} and V_{DD} cancels out the dependency of power on these parameters.

3.4.3 Comparison with Numerical Solutions

In order to confirm the validity of the V_{DDopt} and V_{THopt} formulas of (3.39) and (3.41) and the simple expression of (3.43), the proposed formulas are compared with the results of numerical solutions by (3.34), (3.35) and (3.36), and the conventional formula in [10] where it is stated that the ED product is minimized when $P_{LEAK,max}/P_D$=1.

Fig. 3.29 shows the result. In this analysis, the activity, a, is varied from 1, 0.1, to 0.01 and the logic depth, L_d, is set to 10, which is typical. ΔV_{TH} is set to 0.1V and ΔT is set to 50K. It is seen from the figure that the discrepancy in V_{THopt} between the numerical solution and the conventional calculation [10] is 0.11V, while the discrepancy is suppressed to 0.03V for the proposed formula calculation.

Fig. 3.30 shows the accuracy of the proposed formulas together with the formula in the previous publication [10]. The calculated values are compared with the results of direct numerical analysis using (3.34), (3.35) and (3.36). It is seen that the proposed formulas are in good accordance with the numerical solutions and above-mentioned approximations are found to be reasonable.

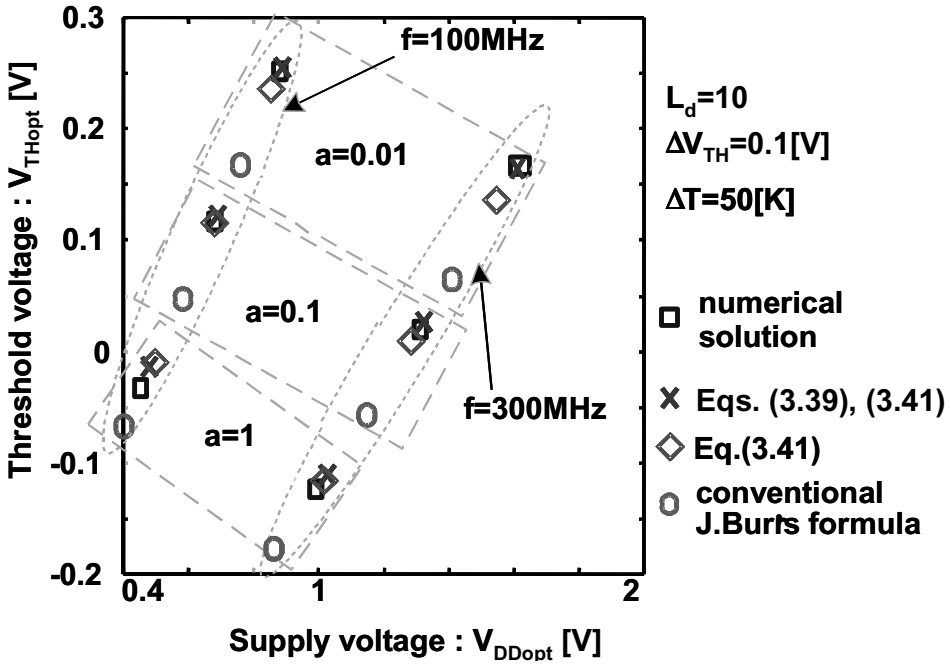


Fig. 3.29 V_{DDopt} and V_{THopt} comparison among proposed analysis formula, proposed simple expression and expression in [10]

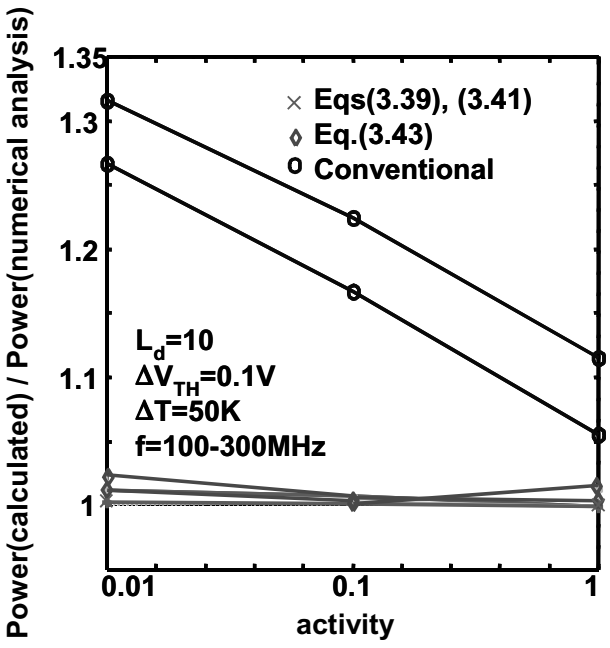


Fig. 3.30 Comparison of power among calculation using (3.39), (3.41) and (3.43) and previously published expression in [10]

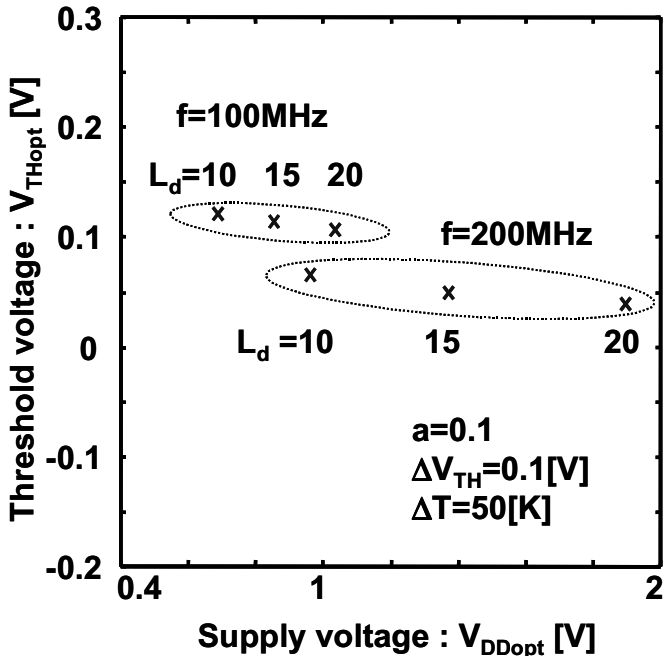


Fig. 3.31 V_{DDopt} and V_{THopt} dependence on logic depth (L_d)

3.4.4 Discussions

It is clearly seen from Fig. 3.29 that V_{THopt} decreases only 0.1V when the required frequency changes from 100MHz to 300MHz. On the other hand, V_{THopt} increases 0.3V when activity, a, changes from 1 to 0.01. Fig. 3.31 shows the V_{DDopt} and V_{THopt} dependence on the logic depth, L_d. In this figure, the variation of V_{THopt} when L_d is changed from 10 to 20 is only 0.03V. From these results, it can be said that V_{THopt} is not a strong function of either the clock frequency or the logic depth but strongly depends on the activity. Therefore, it is effective to decide V_{TH} according to the activity of macro blocks (ex. high V_{TH} for memory blocks, low V_{TH} for logic blocks and further lower V_{TH} for clock circuits). The power increases exponentially when V_{TH} decreases. Hence, to improve the speed, V_{DD} tends to increase and V_{TH} tends to stay the same. This is the reason why V_{THopt} is not a strong function of speed related constraints.

3.4.5 Future Trend of Optimum V_{TH} and Design

A future trend in V_{DD} and power dissipation has been shown in the ITRS (International Technology Roadmap for Semiconductors) [16]. V_{TH} and the logic depth, however, are not discussed in the roadmap. In this section, the trend of the optimum V_{TH}, the logic depth, and the number of transistors in logic blocks is discussed using the parameter values given in the ITRS.

When a certain device parameter is given in the ITRS, it is used in the analysis. For parameters that are not listed in the roadmap, reasonable assumptions are made as follows. α, K and N_S, are assumed to be constant in all generations, being equal to 1.3, 0.78, and 0.05, respectively. T_{min} and T_{max} are set equal to 300K and 400K, respectively. The activity, a, is set to 0.1 for logic blocks [18].

κ is a function of impurity density and can be estimated using the formula in [16]. Fig. 3.32 shows the change of κ on generations. In 0.18μm technology, V_{TH} increases about 0.11V when the temperature goes up by 100K, but when the feature size becomes 0.05μm in 2011, the V_{TH} change will be less than 0.07V.

The total number of transistors on a chip, N_{CHIP}, consists of the number of transistors in logic blocks and that in memory blocks. N_{CHIP} in 2011 is predicted to be about 70 times as large as that in 1999. The power dissipation in memory blocks can be neglected when leakage cutoff techniques are used (for example, see dynamic leakage cut-off scheme [19]). Therefore, the number of transistors in logic blocks, N_{LOGIC}, is of importance in calculating the power consumption. At present, the ratio of N_{LOGIC} to N_{CHIP} is about 20%. For a moment, let us suppose the ratio is invariant over time. L_d is also set constant at 20.

Fig. 3.33 shows the power consumption trend by the estimation through proposed formulas and that by the ITRS. In the calculation, the power will increase by a factor of

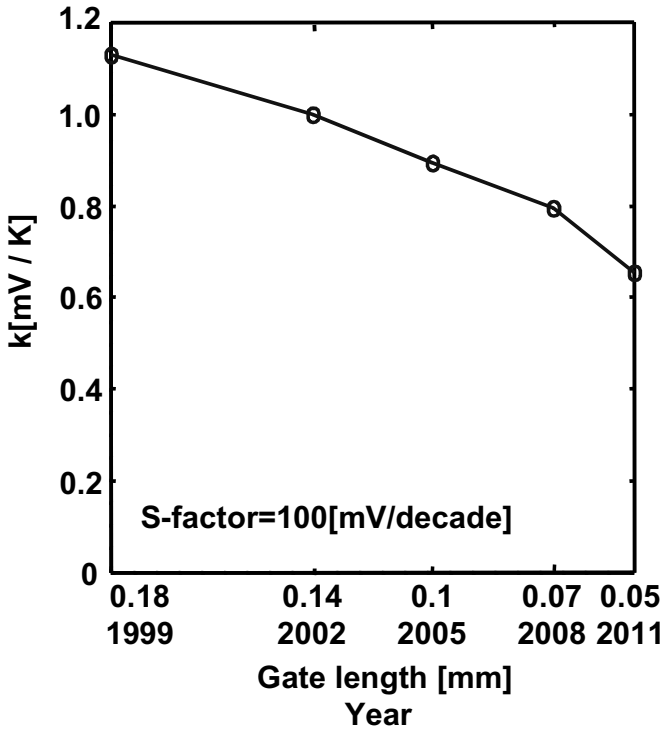


Fig. 3.32 Change in κ on generations

30. On the other hand, the ITRS tells that the total power in 2011 should be within 2 times the power in 1999. It is clear that the target in the ITRS cannot be achieved without some modifications in the scaling scenario. The main parameters, which can be modified in the design level, are the logic depth and the ratio of N_{LOGIC}/N_{CHIP}.

Three scenarios are considered here. In the first scenario, N_{LOGIC}/N_{CHIP} remains constant at 20%, while the logic depth can be changed freely. The logic depth is a function of architecture, a pipeline scheme and a design style. There are no official values for the L_d change in time. The estimated logic depth in 2008 becomes 1. Although there is a tendency that the logic depth is being decreased, this is totally unrealistic.

In the second scenario, L_d is kept constant at 20 and N_{LOGIC}/N_{CHIP} are changed freely. Then, N_{LOGIC} in 2011 will be 1.1 times of N_{LOGIC} in 1999. This scenario again is unrealistic, since it basically says that the number of transistors for logic blocks should not be increased.

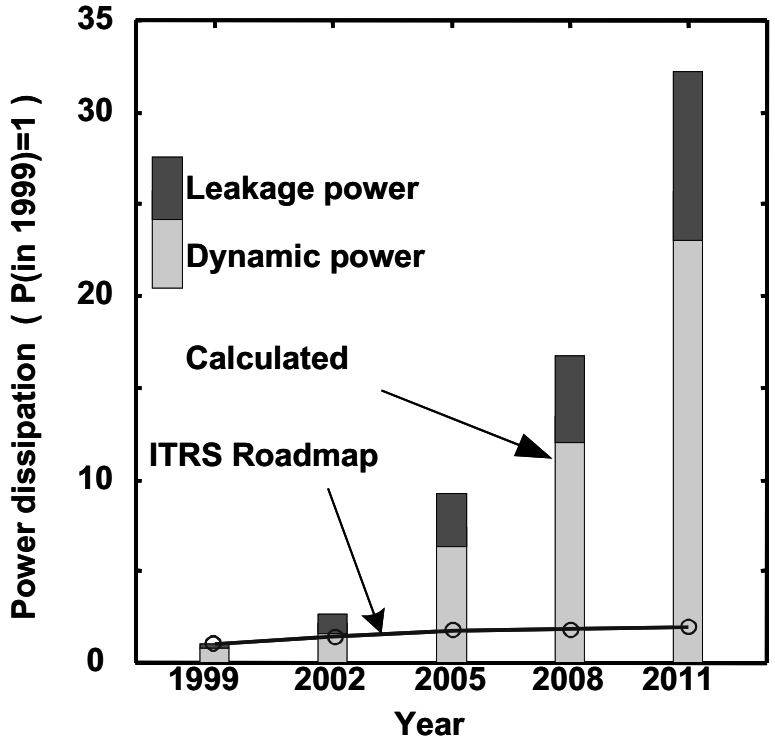


Fig. 3.33 Power trend by estimation through proposed formulas and that by ITRS (constant L_d and N_{LOGIC}/N_{CHIP})

Now, in the third scenario, more realistic values for L_d and N_{LOGIC} are searched for. In this scenario, the minimum achievable L_d is set equal to 10, a half of the current typical value and then N_{LOGIC} in 2011 can be calculated and fixed. From 1999 through 2011, N_{LOGIC} are interpolated assuming an exponential change in time. The resultant figure is shown in Fig. 3.34. This can be one possible scenario. The point is that memories can be using more transistors while logic part cannot be. Fig. 3.35 shows the trend in V_{DDopt} and V_{THopt} in this scenario. The optimum V_{TH} is varied in the hatched region due to the process and temperature variation. The lowest boundary and the highest boundary are optimum $V_{TH,min}$ and optimum $V_{TH,max}$, respectively. The target V_{TH} is the V_{TH} in lowest temperature and medium point of process variation range. From this analysis, it is shown that the target V_{TH} is almost constant at 0.2V and the optimum $V_{TH,min}$ is in the range of 0V~0.1V over generations. This conclusion is basically unchanged even if activity

increases up to 0.3 from 0.1. The resultant of target V_{TH} can be used as a guideline of device manufacturing parameter, for example, the impurity range.

V_{DDopt} coincides with the ITRS. There are many ideas presented to reduce stand-by power but up to now there are eventually no successful proposals on reducing the active power except for changing the supply voltage. In this circumstance, this third scenario is a compromised approach.

The future trend of N_{LOGIC} and N_{MEMORY} ($=N_{CHIP}$-N_{LOGIC} : the number of transistors in memory blocks) is shown in Fig. 3.36. The ratio of N_{MEMORY} to N_{CHIP} is 80% in 1999 and 97% in 2011. On the other hand, the ratio of N_{LOGIC} to N_{CHIP} is decreased to 3% in 2011 due to the constraint of power consumption in the ITRS. Therefore, memories can be using more transistors which logic part cannot be.

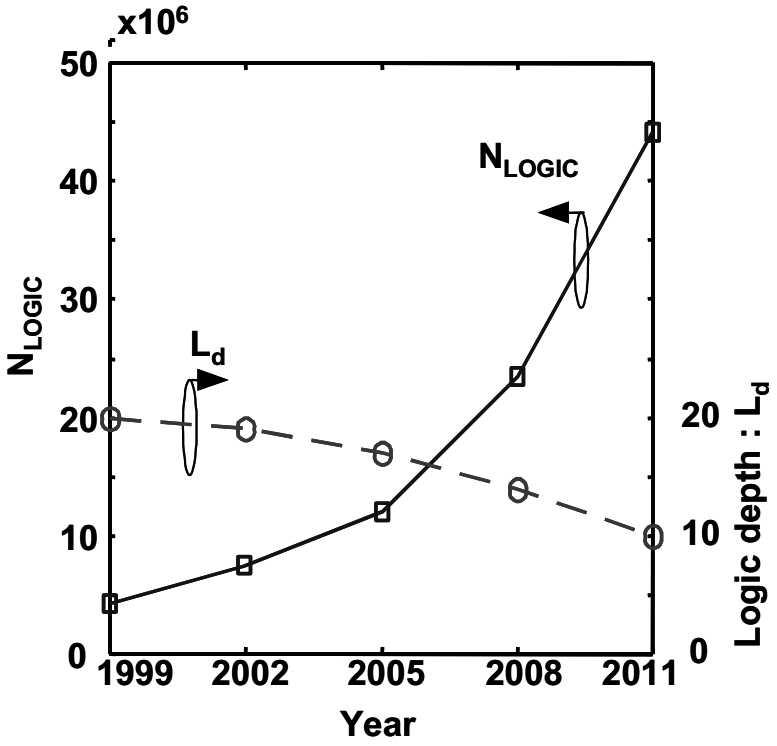


Fig. 3.34 Trend of the number of transistors in logic blocks (N_{LOGIC}) and logic depth (L_d)

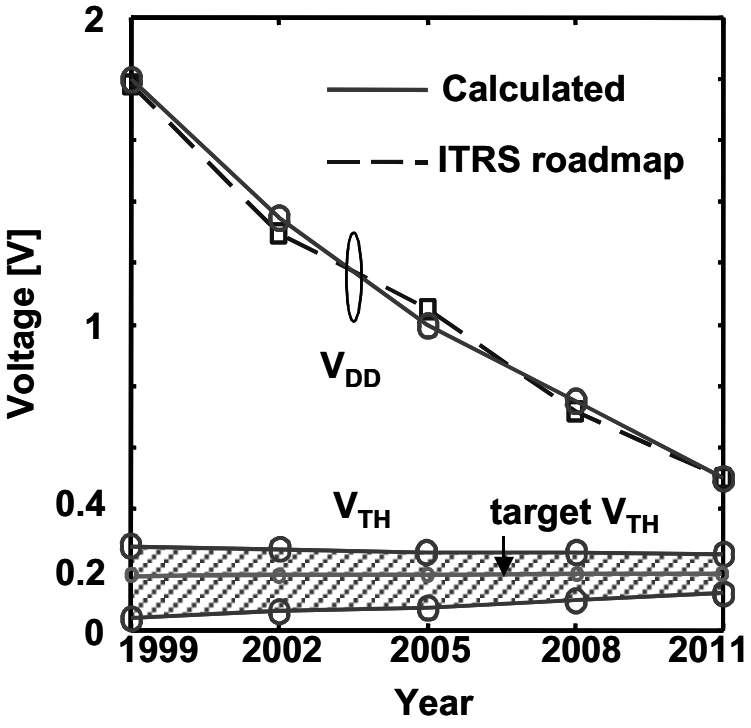


Fig. 3.35 Prediction of optimum V_{DD} and V_{TH}

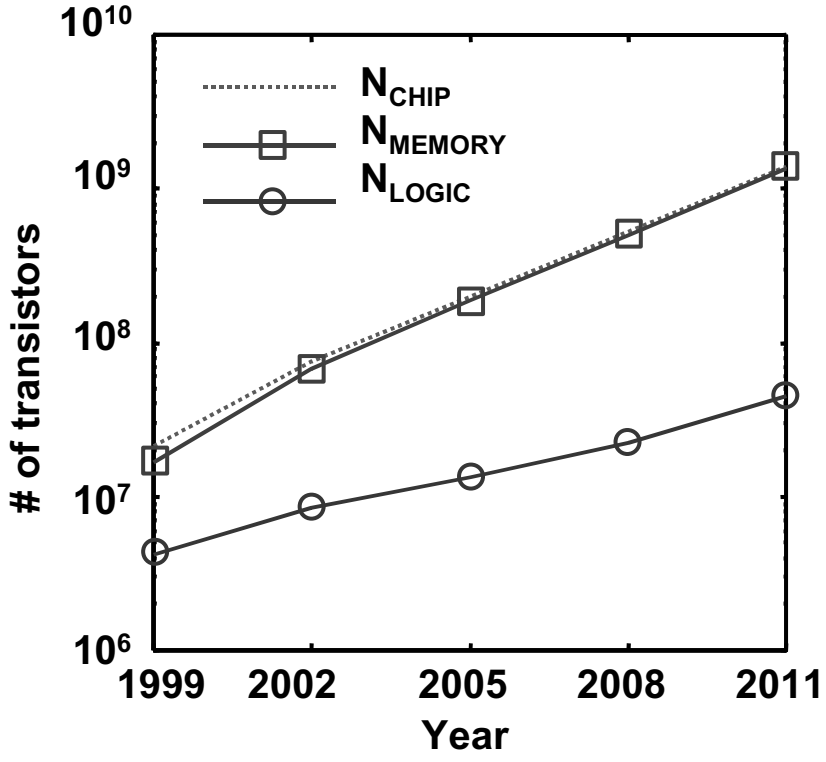


Fig. 3.36 Future trend of number of transistors for logic and memory on chip

3.5 Summary

In Section 3.2, a simple and closed-form formula for the short-circuit power dissipation is derived which correctly reproduces the dependence on various parameters such as a threshold voltage, a supply voltage, a beta ratio, transition time of input voltage, load capacitance and input capacitance.

It is shown that the short-circuit power monotonically increases as α decreases, as fanout decreases and as the ratio of the threshold voltage over V_{DD} decreases. Considering the tendency that V_{TH}/V_{DD} will be slightly increasing to keep the standby power in a tolerant level when the supply voltage is decreased as device miniaturization proceeds, the importance of the short-circuit power will not be increased (about 10%).

In Section 3.3, an appropriate effective gate capacitance, $C_{G,eff}$, has been defined and a method is proposed to extract the value by using SPICE. It is shown that the power and delay of CMOS digital circuit can be estimated accurately by introducing $C_{G,eff}$. The discrepancy between $C_{G,eff}$ and C_{OX} is increasing in low-voltage regime and adopting $C_{G,eff}$ in accurate power and delay estimation becomes more important in the future.

Closed-form formulas for optimum V_{DD} and V_{TH} are presented for low power and high-speed LSI's in Section 3.4. These formulas take the variation of threshold voltage and temperature into account. From the calculation using these formulas, it is shown that a simple guideline for power optimization is to set the ratio of the maximum leakage power to the total power around 30%. Note that the maximum leakage power is observed at the highest temperature and at the lowest V_{TH} corner in process variation.

The trend in V_{THopt} and V_{DDopt} is calculated using the device parameters given in the ITRS roadmap. The V_{DDopt} coincides with the ITRS roadmap and V_{THopt}, that is, the optimum $V_{TH,min}$ is in the range of 0V~0.1V and the target V_{TH} is almost constant at 0.2V over generations. The proposed scenario shows that more number of MOSFETs are

consumed in the memory blocks than the logic blocks in the future.

Appendix A. Deviation of Short-Circuit Power with Fast Input Transition Time

In the α-power law model, the drain current I_D is given as follows [6]

$$I_D = \begin{cases} 0 & \text{(cutoff region)} \\ I'_{D0}\left(2 - \dfrac{V_{DS}}{V'_{DS}}\right)\dfrac{V_{DS}}{V'_{DS}} & \text{(linear region)} \\ I'_{D0} & \text{(saturated region)} \end{cases} \tag{A.1}$$

where

$$I'_{D0} = I_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\alpha} \tag{A.2}$$

$$V'_{D0} = V_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\frac{\alpha}{2}}. \tag{A.3}$$

When α=2, this model becomes the Shockley model.

In this Appendix, the CMOS inverter shown in Fig. 3.1 is used for the derivation of the short-circuit power dissipation. Fig. 3.1 shows the input and output voltage waveform discharging the load capacitance. Where t_T is the transient time of the input voltage, t_0 is the time when input voltage reach at the threshold voltage of NMOS, and t_1 is the time when input voltage reach at the threshold voltage of PMOS. The short-circuit current flows between t_0 and t_1. Then, the output voltage is governed by the following differential equation.

$$C_{OUT}\frac{dV_{OUT}}{dt} = I_{DP} - I_{DN} \tag{A.4}$$

When $t_T \ll \tau_N$, however, it can be assumed that $I_{DP} \ll I_{DN}$. When the transient time of the input is slower than τ_N, it can be assumed that NMOS is in the saturated region between t_0 and t_1. Then, (A.4) can be rewritten as

$$C_{OUT}\frac{dV_{OUT}}{dt} = -I_{DON}\left(\frac{V_{GSN}-V_{THN}}{V_{DD}-V_{THN}}\right)^{\alpha_N}, \tag{A.5}$$

which should be solved with the initial condition, $V_{OUT}=V_{DD}$. Solving the above differential equation, we have

$$v_{OUT}(t) = 1-\frac{1}{\tau_N(\alpha_N+1)}\frac{(t-t_T v_{TN})^{\alpha_N+1}}{(t_T-t_T v_{TN})^{\alpha_N}}. \tag{A.6}$$

In this condition when $t_T \ll \tau_N$, PMOS is in the linear region. From (A.1), (A.2) and (A.3), the PMOS drain current, I_{DP}, is calculated as

$$I_{DP} = 2\frac{I_{D0P}}{V_{D0P}}\left(\frac{|V_{GSP}|-V_{TP}}{V_{DD}-V_{TP}}\right)^{\frac{\alpha_P}{2}}V_{DSP} - \frac{I_{D0P}}{V_{D0P}^2}V_{DSP}^2. \tag{A.7}$$

Since the output capacitance is relatively large, the output voltage moves very slowly. Then, V_{DSP} is small when the input is changing and with this assumption, the second term of (A.7) can be ignored. The second term of (A.6) becomes v_{DSP}, (A.7) can be solved in terms of I_{DP}. From these formulas, the short-circuit power dissipation, P_S, is shown as

$$P_S(t_T \ll \tau_N) = V_{DD}\int_{t_0}^{t_1} I_{DP}dt = \frac{2V_{DD}I_{D0P}t_T}{v_{D0P}\tau_N(\alpha_N+1)(1-v_{TN})^{\alpha_N}(1-v_{TP})^{\alpha_P/2}} \times \int_{t_0}^{t_1}(1-\frac{t}{t_T}-v_{TP})^{\frac{\alpha_P}{2}}(\frac{t}{t_T}-v_{TN})^{\alpha_N+1}dt. \tag{A.8}$$

Note that

$$\int_{t_0}^{t_1}(1-\frac{t}{t_T}-v_{TP})^{\frac{\alpha_P}{2}}(\frac{t}{t_T}-v_{TN})^{\alpha_N+1}dt$$
$$=\int_0^{1-v_{TP}-v_{TN}}\{(1-v_{TN}-v_{TP})-x\}^{\frac{\alpha_P}{2}}x^{\alpha_N+1}t_T dx \quad , \tag{A.9}$$

where $x=\frac{t}{t_T}-v_{TN}$. Now, the Taylor transformation is applied for the integrand.

$$\{(1-v_{TN}-v_{TP})-x\}^{\frac{\alpha_P}{2}}$$
$$=m^{\frac{\alpha_P}{2}}-\frac{\alpha_P}{2}m^{\frac{\alpha_P}{2}-1}x+\frac{\alpha_P}{2}\left(\frac{\alpha_P}{2}-1\right)m^{\frac{\alpha_P}{2}-2}x^2-\cdots \tag{A.10}$$

where m=1-v_{TN}-v_{TP}. Then the integration can be carried out as follows.

$$t_T m^{\frac{\alpha_P}{2}+\alpha_N+2}\left\{\frac{1}{\alpha_N+2}-\frac{\alpha_P}{2(\alpha_N+3)}+\frac{\alpha_P}{2\cdot 2(\alpha_N+4)}\left(\frac{\alpha_P}{2}-1\right)-\cdots\right\} \tag{A.11}$$

Let's concentrate on the quantity in the parenthesis. When the third term is multiplied by 4, a good appropriation can be obtained which fits well with SPICE simulation and this accounts for the higher terms than the fourth. Hence, the quantity in the parenthesis can be approximated as $f(\alpha)$, which is defined as

$$f(\alpha)=\left\{\frac{1}{\alpha_N+2}-\frac{\alpha_P}{2(\alpha_N+3)}+\frac{\alpha_P}{\alpha_N+4}\left(\frac{\alpha_P}{2}-1\right)\right\}. \tag{A.12}$$

With $f(\alpha)$, (3.2) in the text can be easily derived.

References

[1] H. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-19, pp.468-473, Aug. 1984.

[2] N. Hedenstierna and K. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp.270-281, Mar. 1987.

[3] S. Vemuru, N. Scheinberg and E. Smith, "Short-circuit power dissipation formula for CMOS gates," *Proceedings on IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 1333-1335, 1993.

[4] A. Hirata, H. Onodera and K. Tamaru, "Short-circuit power dissipation formula for static CMOS gates," *Karuizawa workshop on Circuits and Systems*, pp.245-250, July 1996.

[5] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. SC-25, pp.584-594, Apr. 1990.

[6] T. Sakurai and A. R. Newton, "Delay analysis of series-connected MOSFET circuits," *IEEE Journal of Solid-State Circuits*, vol. SC-26, pp.122-131, Feb. 1991.

[7] P. Yang, B.D.Epler and P.K.Chatterjee, "An investigation of the charge conservation problem for MOSFET circuit simulation," *IEEE Journal of Solid-State Circuits*, vol. SC-18, no.1, pp.128-138, Feb., 1983.

[8] D. Foty, *MOSFET Modeling with SPICE*, Prentice Hall, Inc., 1997.

[9] G. Massobrio and P. Antognetti, *Semiconductor device modeling with SPICE*, McGraw-Hill, Inc., 1993.

[10] J. Burr and A. Perterson, "Ultra low power CMOS technology," *NASA VLSI Design Symposium*, pp. 4.2.1-4.2.13, 1991.

[11] R. Gonzalez, B. M. Gordon and M. A. Horowitz, “Supply and threshold voltage scaling for low power CMOS,” *IEEE Jounal of Solid-State Circuit*, vol. 32, pp. 1210-1216, Aug., 1997.

[12] Z. Chen, C. Diaz, J. D. Plummer, M.Cao and W. Greene, “0.18um dual Vt MOSFET processing and energy-delay measurement,” *IEDM tech. Digest*, pp. 851-854, 1996.

[13] A. Bellaouar, A. Fridi, M. I. Elmasry and K. Itoh, “Supply voltage scaling for temperature insensitive CMOS circuit operation,” *IEEE Transaction on Circuit and Systems II*, vol. 45, pp. 415-417, Mar., 1998.

[14] C. Park et al, “Reversal of temperature dependence of integrated circuits operation at very low voltages,” *IEDM Tech., Digest*, pp. 71-74, 1995.

[15] K. Kanda, K. Nose, H. Kawaguchi and T. Sakurai, “Design impact of positive temperature dependence of drain current in sub 1V CMOS VLSI's,” *Proceedings of Custom Integrated Circuit Conference*, pp.563-566, May, 1999.

[16] The International Technology Roadmap for Semiconductors, ITRS Handbook, 1998.

[17] Y. Taur and T. H. Ning, “Fundamental of Modern VLSI Devices,” by Cambridge University Press, pp. 131, 1998.

[18] J. Burr and J. Shott, “A 200mV encoder-decoder circuit using stanford ultra low-power CMOS,” *ISSCC Digest of Tech. Papers*, pp. 84-85, Feb., 1994.

[19] H. Kawaguchi, Y. Itaka and T. Sakurai, "Dynamic leakage cut-off scheme for low-voltage SRAM's," *Symp. on VLSI Circuits*, pp.140-141, June, 1998.

Chapter 4. Buffer Insertion Schemes for High-Speed and Low-Power Interconnect Designs

4.1 Introduction

Optimization of the interconnect delay by the buffer insertion technique is a crucial technique for deep submicron VLSI's. RC models for MOSFET's have been used to optimize the buffered interconnect. As for the resistor, the transistor has been approximated as a linear resistor without detailed consideration on the non-linear feature of MOS I-V curves. As for the capacitance, the junction capacitance, C_J, has often been neglected [1] or even if C_J is taken into account, the delay formula including C_J is not sufficiently accurate. Moreover, the existing theories for buffered interconnect optimization are lacking in the trade-off between the delay and the power consumption although the power is one of the most important index in future giga-scale integration.

In order to overcome the shortcomings of the conventional approaches, an approximation of MOSFET as a linear resistor is investigated and the delay formula

including C_J is proposed in Section 4.2. The study also gives attention to the power consumption in the optimization process and derives closed-form formulas for optimum buffer insertion. The results have been applied to bulk and SOI technologies and implications of buffered interconnect on technologies are discussed.

From the point of view of the circuit design, the new buffer insertion scheme which can alleviate the effect of the coupling capacitance and long RC line is important for high-performance VLSI design. The original buffer insertion schemes, however, cannot be applied to the bi-directional buses because the buffer is uni-directional in nature. Some circuit configurations that can be applied to bi-directional buses have been proposed [2][3]. These circuits turn out to be prone to malfunctions when there is a noise from adjacent lines in scaled down interconnect systems where capacitive coupling is large.

In order to overcome these problems, a new buffer insertion scheme for bi-directional buses, namely dual-rail bus (DRB) scheme, which does not have noise problems is proposed and measured in Section 4.3. One more proposal is on a high-speed buffer insertion scheme for uni-directional buses by making use of staggered firing. The staggered firing bus (SFB) is proposed and measured.

4.2 Power-Conscious Interconnect Buffer Optimization with Improved Modeling of Driver MOSFET and Its Implications to Bulk and SOI CMOS Technology

4.2.1 Analytical Model for Buffer Optimization

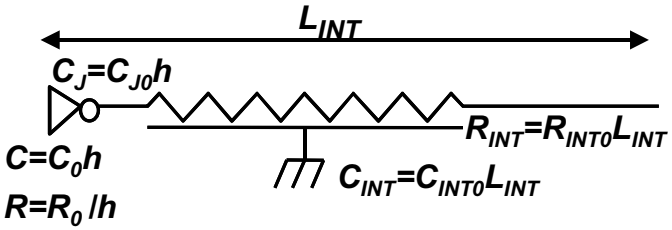
Fig. 4.1 shows a basic configuration of buffered interconnect. R, C and C_J are the resistance, gate capacitance and junction capacitance of the buffer, respectively. h is the buffer size. The inductive effect is neglected in this study since the effect on optimum buffer inserted lines will diminish and become negligible for global interconnects in the future [2]. The delay formula without buffers can be approximated as (4.1). Suffix 0 signifies quantity per unit size or length.

$$\begin{aligned} t_d &= p_1 R_{INT} C_{INT} + p_2 (RC + RC_J + RC_{INT} + R_{INT} C) \\ &= k\left[p_1 \frac{R_{INT0} L_{INT}}{k} \frac{C_{INT0} L_{INT}}{k} + p_2 \left(\frac{R_0}{h} h(C_0 + C_{J0}) + \frac{R_0}{h} \frac{C_{INT0} L_{INT}}{k} + \frac{R_{INT0} L_{INT}}{k} h C_0 \right) \right] \end{aligned} \tag{4.1}$$

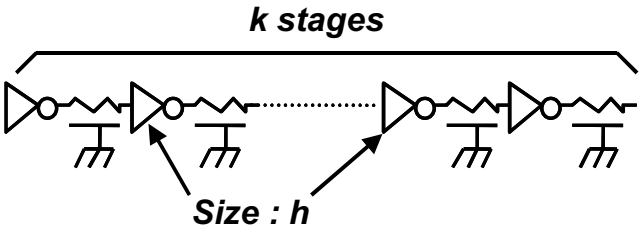
where p_1=0.377 and p_2=0.693

This expression is newly derived and the relative error is within 3% when C_J=0 and within 7.5% when C_J is equal to or less than C, which is the input capacitance of a transistor.

When the buffers are inserted like in Fig. 4.1(b), the optimum size of the buffers and the optimum number of the buffers can be derived analytically as



(a) without buffers



(b) with buffers

Fig. 4.1 Distributed RC interconnect model

$$\frac{\partial t_d}{\partial h} = 0 \rightarrow h_{OPT} = \sqrt{\frac{C_{INT0} R_0}{R_{INT0} C_0}}, \qquad (4.2)$$

$$\frac{\partial t_d}{\partial k} = 0 \rightarrow k_{OPT} = L_{INT} \sqrt{\frac{p_1}{p_2}} \sqrt{\frac{R_{INT0} C_{INT0}}{R_0 (C_0 + C_{J0})}}. \qquad (4.3)$$

h_{OPT} is the optimum size of the buffers and k_{OPT} is the optimum number of the buffers. L_{INT} is the interconnect length.

Substituting h_{OPT} (4.2) and k_{OPT} (4.3) into (4.1), the optimum delay (t_{dOPT}) can be expressed as

$$\begin{aligned} t_{dOPT} &= 2L_{INT}\left(\sqrt{p_1 p_2} + p_2\sqrt{\frac{C_0}{C_0 + C_{J0}}}\right)\sqrt{\tau_{INT0}\tau_{MOS0}} \\ &\cong 2.4 L_{INT}\sqrt{\tau_{INT0}\tau_{MOS0}} \quad (when\ C_{J0} = 0) \\ &\cong 2.0 L_{INT}\sqrt{\tau_{INT0}\tau_{MOS0}} \quad (when\ C_{J0} = C_0) \end{aligned} \qquad . \qquad (4.4)$$

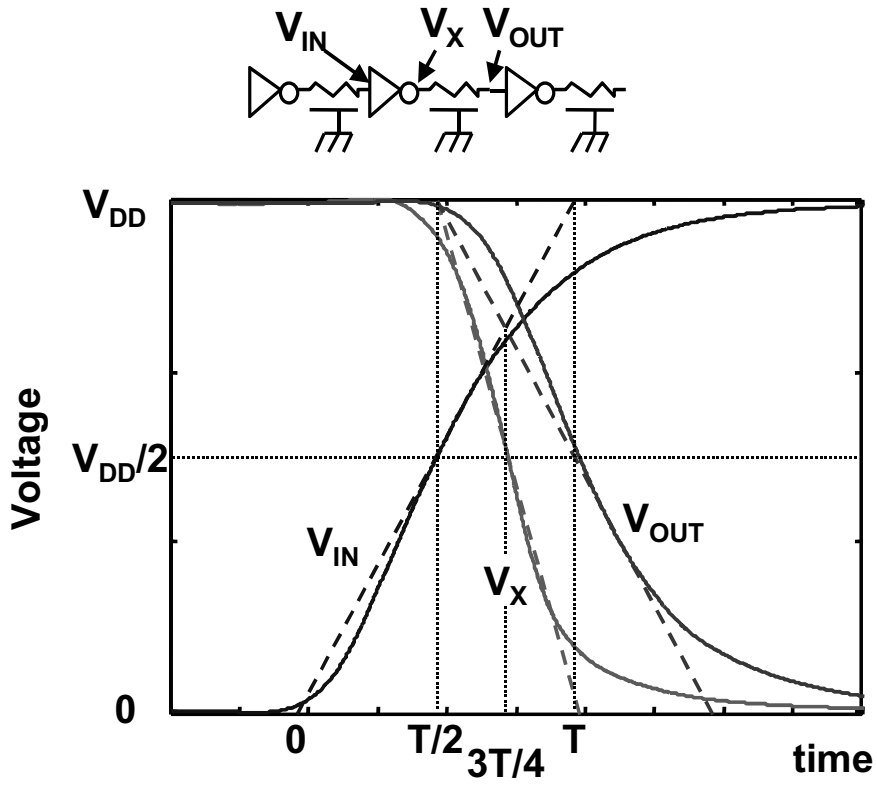


Fig. 4.2 Signal waveforms when buffers are inserted

τ_{INT0} is the time constant of interconnect ($=R_{INT0}C_{INT0}$) and τ_{MOS0} is the time constant of a buffer ($=R_0(C_0+C_{J0})$) which corresponds to the inverter delay with fanout of 1. The optimum delay is proportional to a geometric mean of the interconnect delay (τ_{INT0}) and the gate delay (τ_{MOS0}). This means that the delay of optimally buffered interconnect is approximately scaled as $\sqrt{s}$ where s is a scaling variable. It is also shown that the optimal condition is met when inserted buffer delay is approximately equal to the interconnect delay.

In order to use the derived formulas, the effective linear resistance of the unit-sized transistor (R_0) has to be determined from device characteristics. Here, we discuss an appropriate choice of the effective constant resistance when the buffers are optimized. Fig. 4.2 shows that the waveform of the input, driver output and interconnect output voltage when the buffers are inserted so as to minimize the interconnect delay.

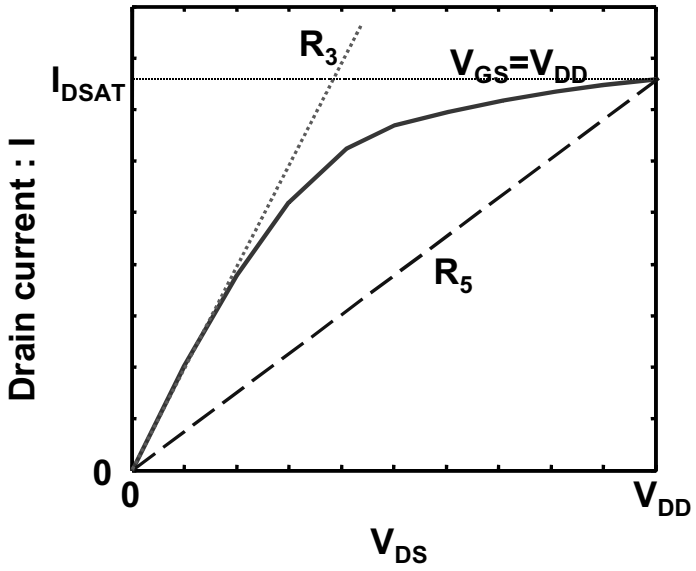


Fig. 4.3 Definition of R_5 and R_3

The waveforms can be considered as the ramp waveforms and α-power model [3] is used as the drain current model. It is assumed that V_X and V_{OUT} begin transition when $V_{IN}=V_{DD}/2$. The slope of V_X is twice as large as the slope of V_{IN} and V_{OUT} when the buffer insertion is optimized. R_5 is the transistor resistance when $V_{DS}=V_{GS}=V_{DD}$ (see Fig. 4.3). The effective linear resistance can be expressed as

$$\eta = \frac{R_0}{R_5} = \frac{3\cdot(\alpha+1)}{32\cdot\ln 2}\cdot\frac{(1-v_T)^{\alpha}}{(3/4-v_T)^{\alpha+1}-(1/2-v_T)^{\alpha+1}}. \qquad (4.5)$$

where v_T is V_{TH}/V_{DD}. The detailed derivation of (4.5) can be found in Appendix B. In order to give insight into the parametric dependence of η, the simpler formula is proposed as

$$\eta = \frac{R_0}{R_5} = \frac{R/h}{V_{DD}/I_{D0}} = 0.7\alpha + 1.5v_T. \qquad (4.6)$$

This expression acts as a bridge between the effective transistor resistance and device characteristics. In Fig. 4.4, the SPICE simulation results are compared with (4.6). Different technology models and various interconnect width and height are used for this

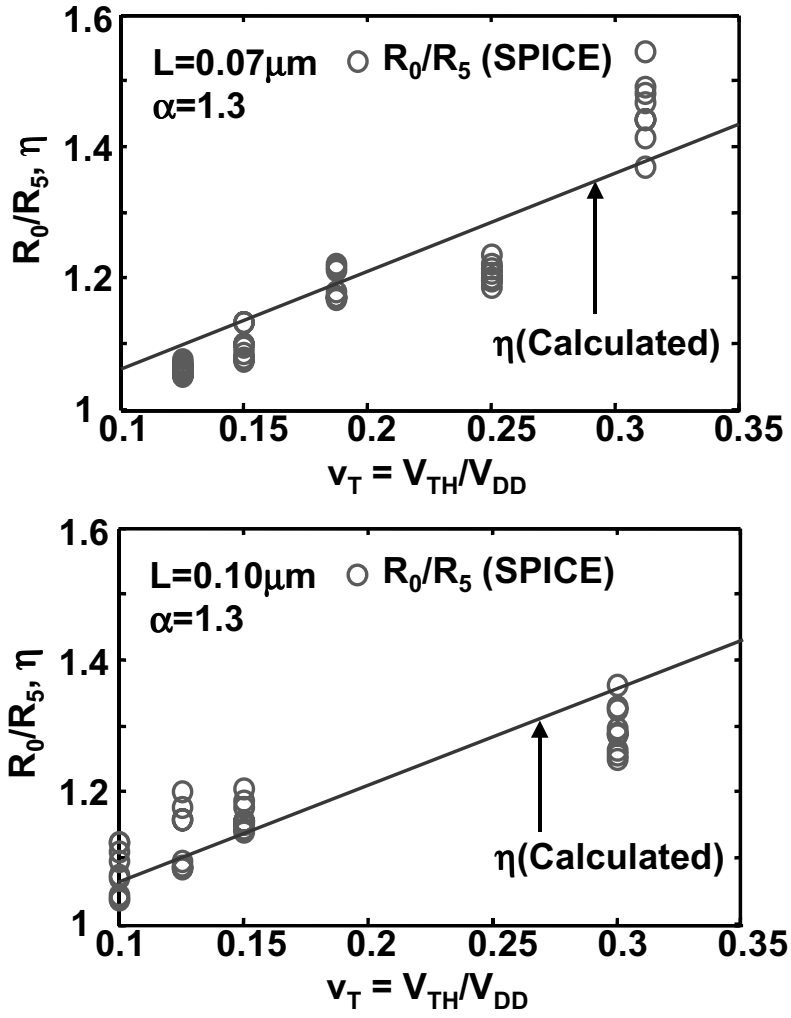


Fig. 4.4 R_0/R_5 and η dependence on v_T

simulation and the validity of (4.6) is confirmed. Fig. 4.5 shows the optimum delay comparison between the proposed method where the effective linear resistance (R_0) is used and the conventional method in [4] where the linear resistance is chosen as the R_3 (=1/(maximum drain conductance) as is shown in Fig. 4.3). The discrepancy between the delay simulated by SPICE with real buffers and a distributed RC line and the calculated delay with the effective linear resistance (R_0) is within 3%. On the other hand, the discrepancy between SPICE simulated delay and the delay calculated with the conventional R_3 is more than 30%. On the other hand, the discrepancy in power between these methods is within 6% (see Fig. 4.6). The optimum buffer size (h_{OPT}) is proportional

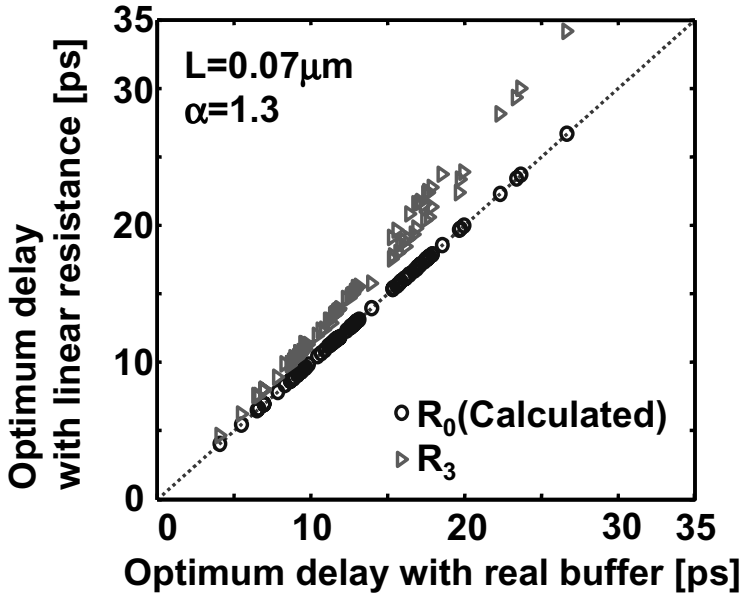


Fig. 4.5 Optimum delay comparison between R_0 and R_3

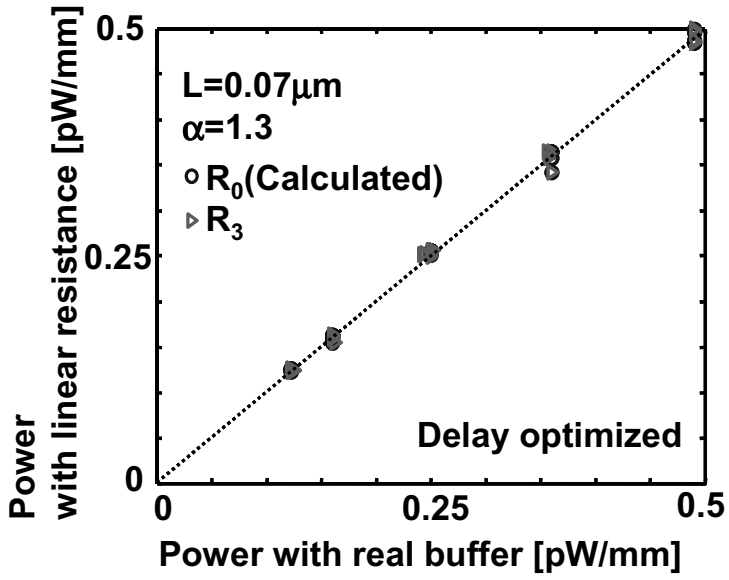


Fig. 4.6 Power comparison between R_0 and R_3

to $\sqrt{R_0}$ and the optimum number of buffers (h_{OPT}) is proportional to $1/\sqrt{R_0}$. This is why the total power with buffers, which is the function of $h_{OPT} \cdot k_{OPT}$, is unchanged even if the effective linear resistance is changed.

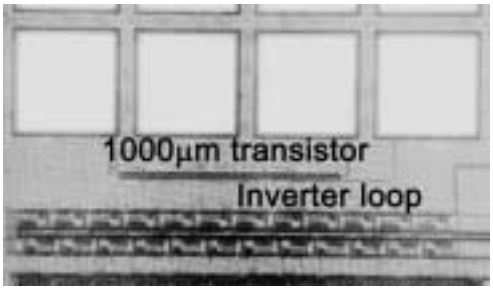


Fig. 4.7 Microphotograph of test chip fabricated by 0.25μm PD-SOI process

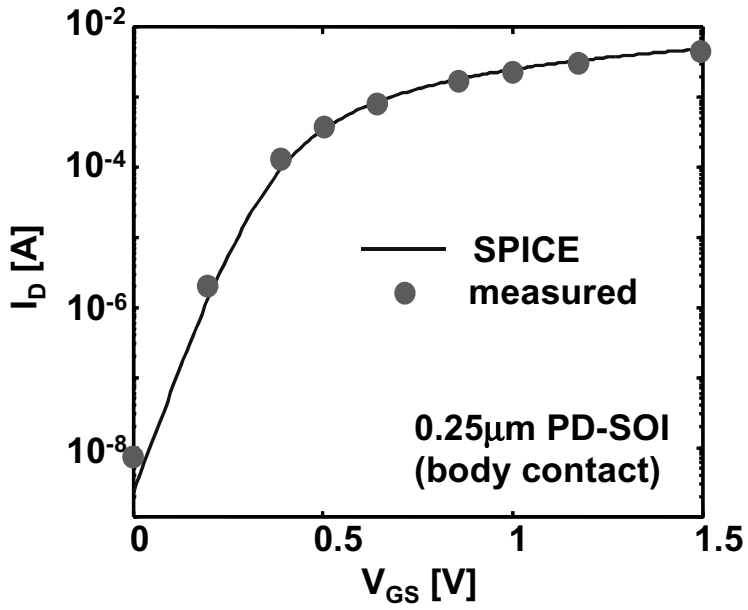


Fig. 4.8 Drain current comparison between SPICE model and measured data

Then, in order to confirm the validity of the proposed formulas for h_{OPT}, k_{OPT} and t_{dOPT}, theoretical calculations and SPICE results are compared. The model parameter set for SPICE simulation and for proposed formulas are extracted from measured data with 0.25μm PD-SOI technology whose test chip is shown in Fig. 4.7. The SPICE model agrees well with the measured results as in Fig. 4.8. Fig. 4.9 shows the h_{OPT}, k_{OPT} and t_{dOPT} comparison between rigorous optimization results with SPICE and the calculated results. Fig. 4.10 shows the power dependence on the C_{J0}/C_0. When the junction capacitance is negligible, both the optimum delay and the power with buffers are

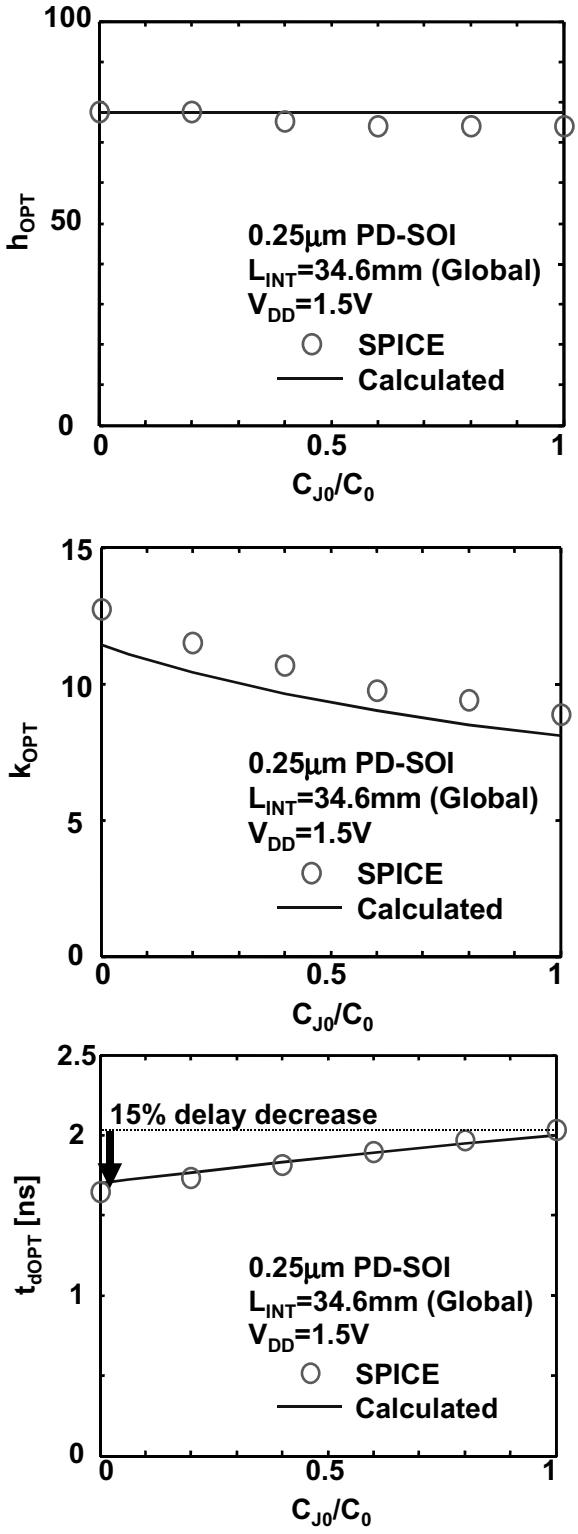


Fig. 4.9 h_{OPT}, k_{OPT} and t_{dOPT} comparison between calculated results and SPICE simulations

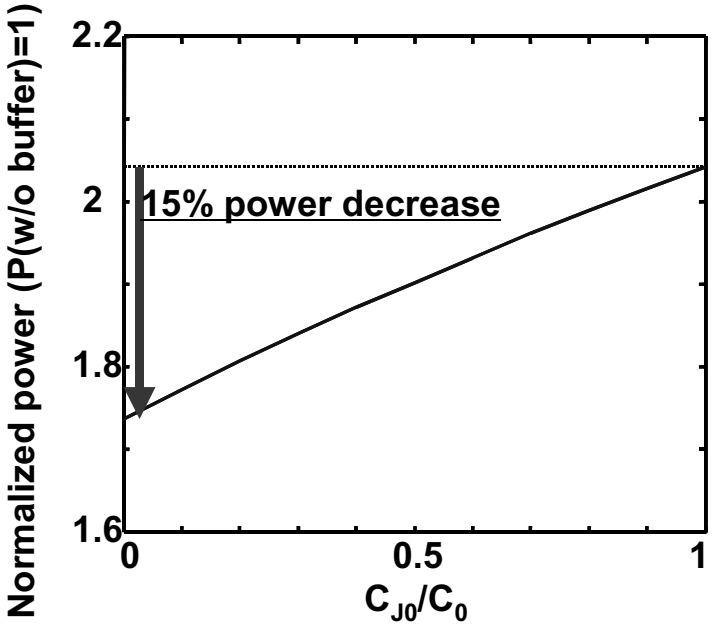


Fig. 4.10 Power dependence on C_{J0}/C_0

suppressed by 15% compared with the MOSFET with $C_{J0}=C_0$. It is shown from (4.2), (4.3) and (4.4) that the 15% reduction on power and delay is independent from the technology node.

4.2.2 Interconnect Delay and Power Comparison Between Bulk and SOI Technology

Extending the analysis, the optimum interconnect delay comparison among bulk, PD-SOI, FD-SOI and double-gate structure [5] is discussed using the simple model. The characteristics of these models are listed in Table 4.1. We set the leakage current of these structures equal to make the comparison fair. Then, V_{TH} of FD-SOI and double-gate can be lowered since the *S*-factor is smaller than other structures. C_{J0}/C_0 and V_{TH}/V_{DD} are the measured data of five different technologies. C_{J0}/C_0 of conventional bulk process are 0.7~1.3. This value does not change drastically over generations.

Table 4.1 Bulk and SOI structure

	C_{J0}/C_0	V_{TH}/V_{DD}	
Bulk	0.92	0.18	
PD-SOI (body contact)	0.13	0.18	
PD-SOI (floating)	0.13	0.18	$I_{ON} \times 1.15$ (kink)
FD-SOI	0.13	0.13	S=60mV/decade
Double-gate [5]	0.13	0.13	S=60mV/decade $I_{ON}\times 2$, $C_0\times 2$

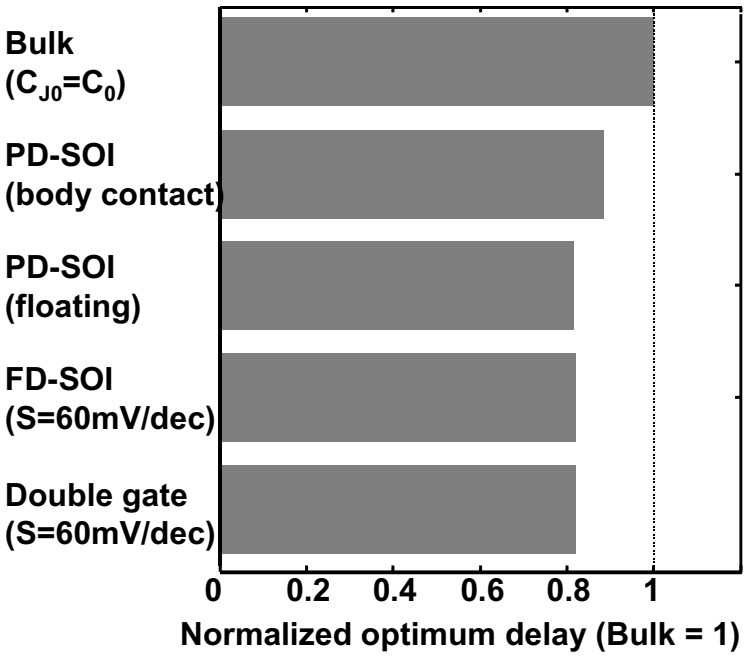


Fig. 4.11 Delay comparison between bulk and SOI

The calculated results are shown in Fig. 4.11. PD-SOI with body contact is 12% faster than bulk CMOS technology due to the small junction capacitance. It is often discussed that SOI technology does not give speed and power improvement over bulk CMOS technology in deep submicron designs, since speed and power are determined by interconnects and SOI technology does not change interconnect layers. It is not

necessarily true because deep submicron interconnect systems need relatively large buffers and due to the improvement through buffers, SOI technology still enjoys advantage over bulk CMOS. The delay can be further decreased by using PD-SOI with a floating body or FD-SOI since the drain current is enhanced by the kink effect and the lower threshold voltage. If lower C_J is achievable with bulk CMOS technology, the bulk technology approaches SOI results.

In the optimally buffered interconnect, the power dissipation increases due to the buffers. Here, the trade-off between power and delay is discussed. Let us introduce the parameter, p, which is the ratio of the total power (buffers and interconnect), P_{TOTAL}, to the power consumed by pure interconnect, P_{INT}.

$$p = \frac{P_{TOTAL}}{P_{INT}} = \frac{C_{INT} + kh(C_0 + C_{J0})}{C_{INT}} \tag{4.7}$$

If p is fixed, the optimum buffer size, h, the number of the sections, k, and the delay, t_d, can be expressed as follows.

$$\begin{aligned} \frac{h}{h_{OPT}} &= \sqrt{\frac{p(p-1)p_2 C_0}{C_P}} \\ \frac{k}{k_{OPT}} &= \sqrt{\frac{(p-1)C_P}{pp_1 C_0}} \\ \frac{t_d}{t_{dOPT}} &= \frac{1}{\sqrt{p_1}+\sqrt{p_2}}\sqrt{\frac{pC_P}{(p-1)C_0}} \end{aligned} \tag{4.8}$$

where

$$C_P = p_1(C_0 + C_{J0}) + p_2 C_0(p-1) \tag{4.9}$$

The delay dependence on the total power is calculated using the proposed formulas. The result is shown in Fig. 4.12. It can be seen from the figure that the power can be reduced by 20% if delay is allowed to increase by 5%.

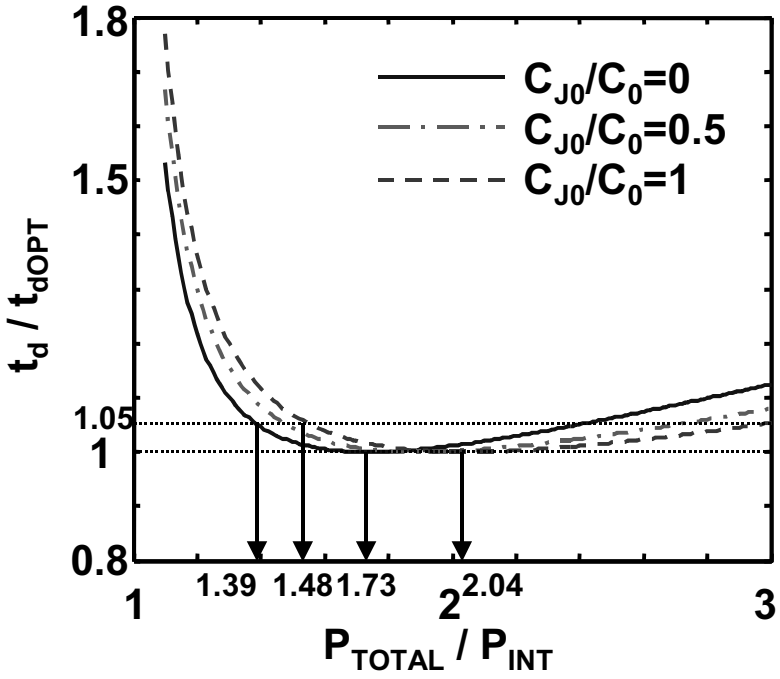


Fig. 4.12 Delay dependence on total power

4.3 Two Schemes to Reduce Interconnect Delay in Bi-Directional and Uni-Directional Buses

4.3.1 Dual-Rail Bus (DRB) for Bi-Directional Buses

Fig. 4.13 shows the schematic of the proposed dual-rail bus (DRB) scheme. This dual-rail bus consists of two buffered interconnects per bit, one of which is right-oriented and the other is left-oriented. When a certain I/O (I/O1) is output-enabled for a bus, the buffer B1 and B2 are forced to Hi-Z to get rid of the driving conflict between D1-B1 and D2-B2. It should be noted that all nodes before B1 and B2 are kept '0'. This is what the driving node does. At all other receiving nodes, the valid signal is constructed by 'OR'ing the signals of the right-oriented line and the left-oriented line. This is because one of the two lines carries a valid signal and the other line carries '0' at all location. The advantage of the proposed scheme is that there is no need for each tri-state buffer on the bus to know the direction from which the input comes. Another advantage is that the delay variation caused by the coupling capacitance among lines does not occur since the left-oriented line and the right-oriented line are placed alternately. Thus, there is no chance that the adjacent lines change state at the same time. If the bus is to be branched, the circuit shown in Fig. 4.14 should be used. The operation diagram of the branch circuit is shown in Fig. 4.15. In order to operate the dual-rail bus scheme properly, it is necessary to keep the principle that one of the two lines carries a valid signal and the other line carries ‘0’ at all locations. By using the branch circuit, the principle can be kept at all times.

Some may think about a ring-structured bus by shorting right-oriented line and left-oriented line at both ends. The ring-structured bus, however, is slow because in the worst-case, the signal should travel twice as long as the length of the bus.

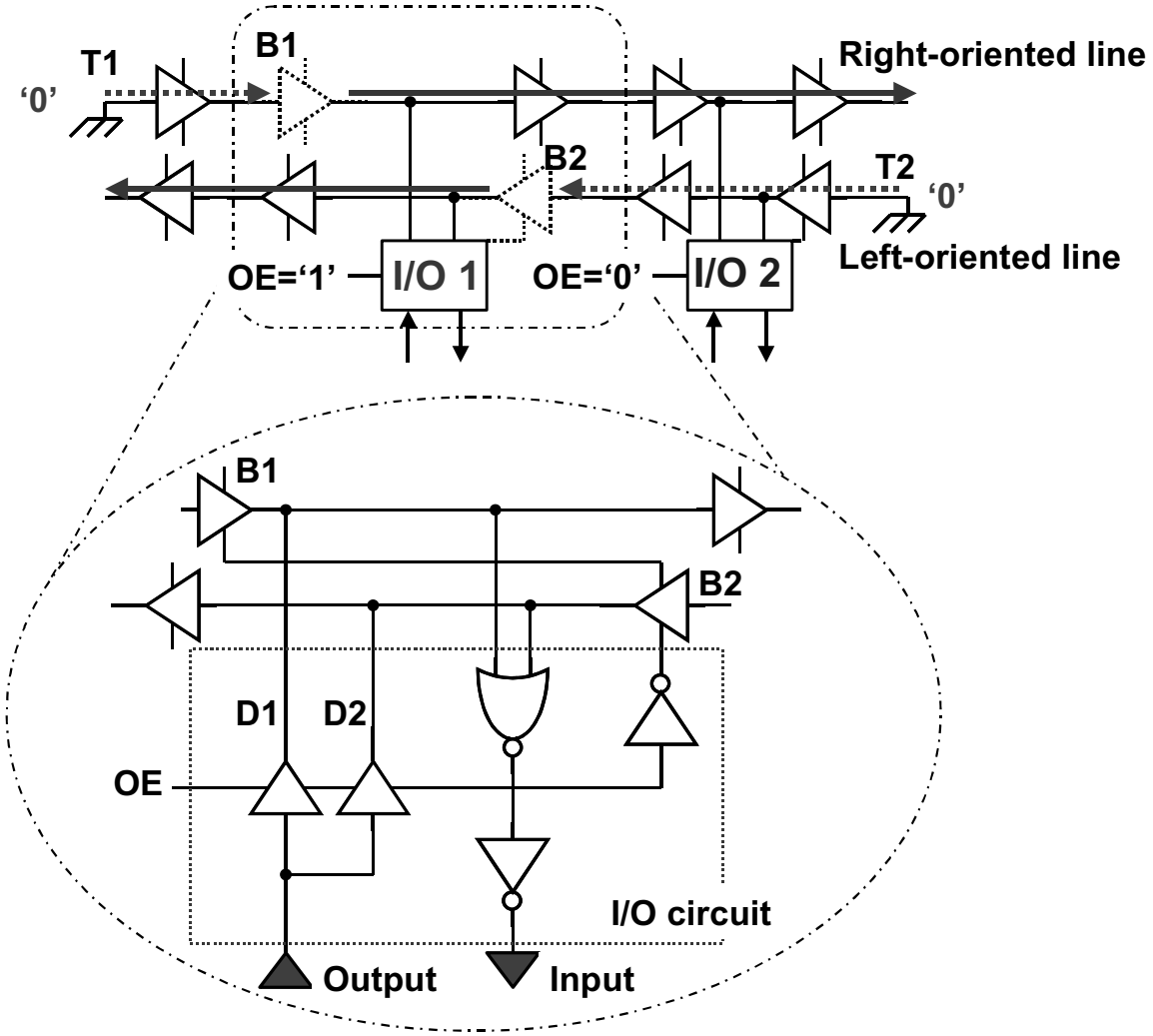


Fig. 4.13 Schematic diagram of dual-rail bus scheme

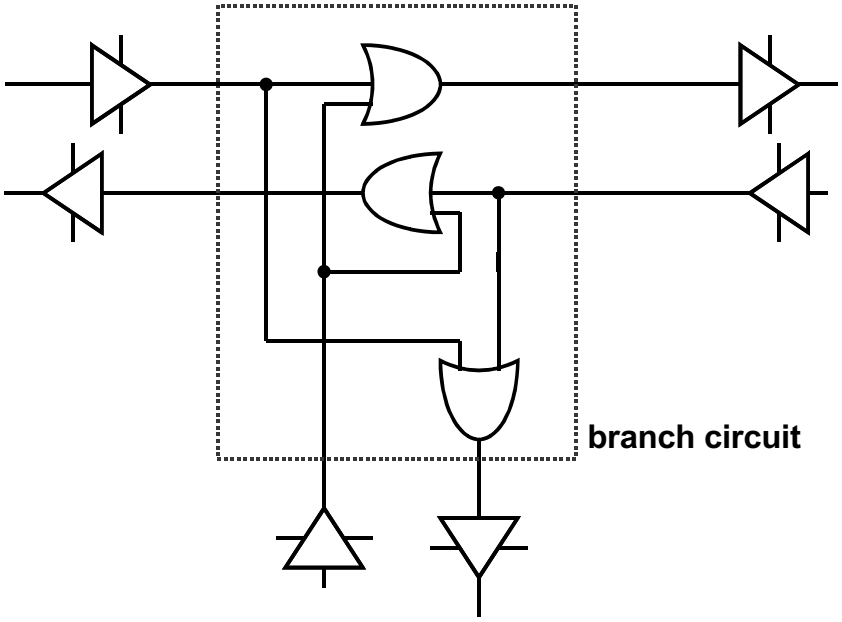


Fig. 4.14 Branch circuit for dual-rail bus

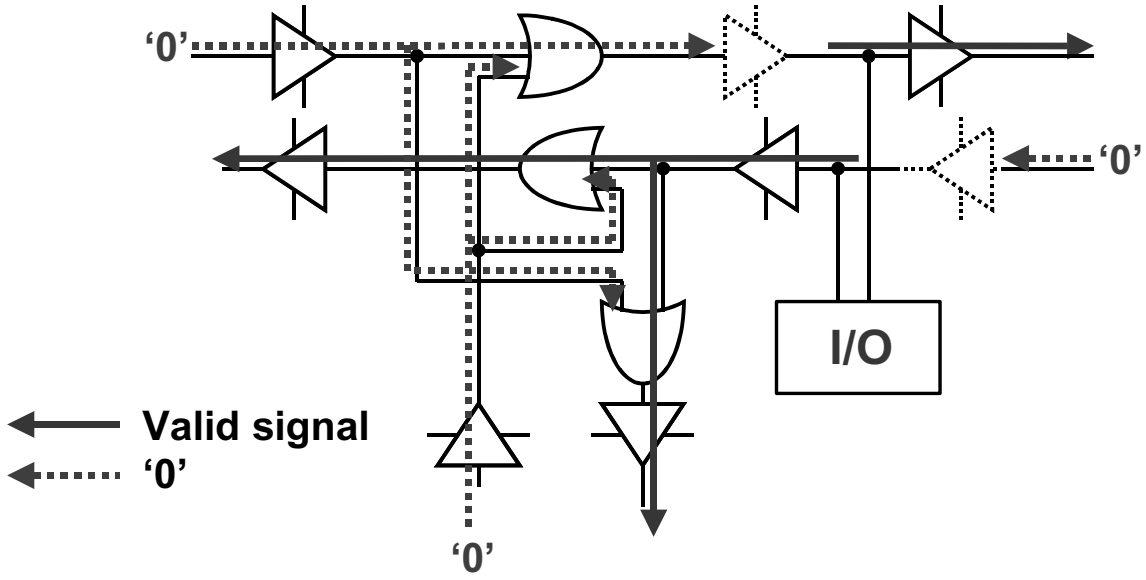


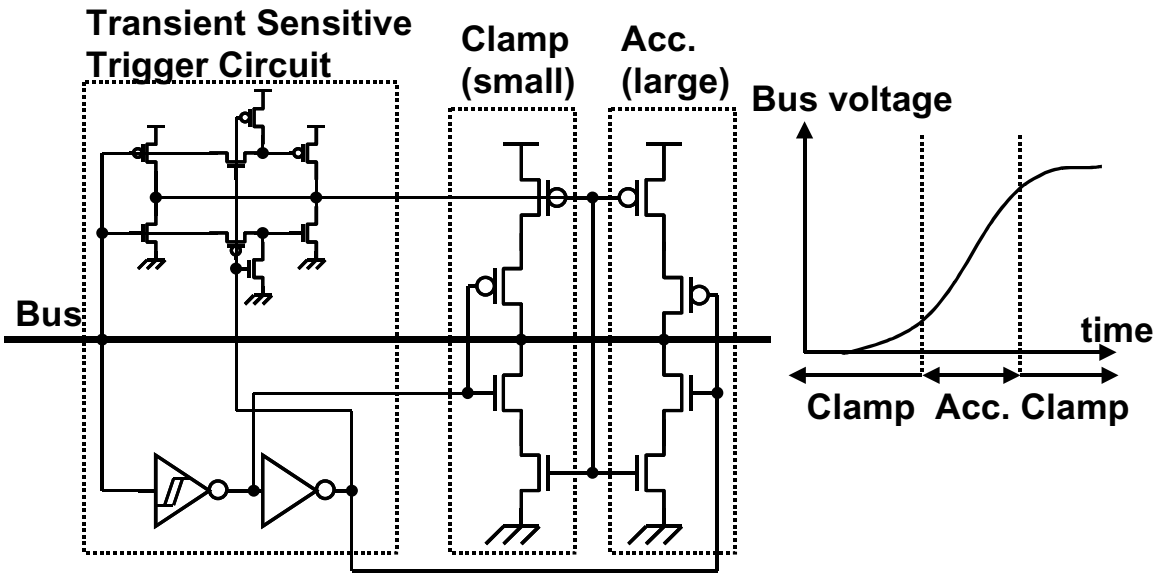
Fig. 4.15 Operation of branch circuit

In order to show the effectiveness of the proposed scheme, the noise resiliency of the proposed scheme and the previously published schemes is compared.

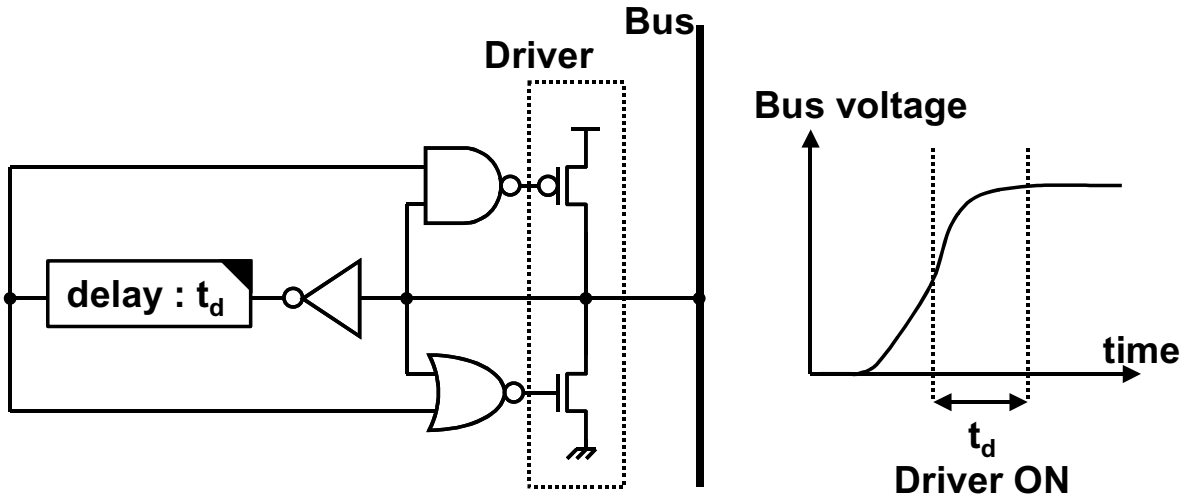
Some high-speed bus schemes have previously published for bi-directional buses. One is transient sensitive accelerator, which is called TSA [7]. The schematic of TSA is shown in Fig. 4.16(a). TSA consists of a transient sensitive trigger circuit, an accelerator and a clamp circuit. When signal voltage starts to change, the transient sensitive trigger circuit senses the signal transition. Then, the accelerator circuit is turned on, whose output driver is large. Thus, the transition of the bus is accelerated. In a steady state, the output of the accelerator circuit becomes high-impedance, and the clump circuit is activated. The clump circuit is used to hold the bus voltage. In order to avoid the conflict between the output of the clump circuit and the bus input, large-sized transistors cannot be used for the clump circuit. Consequently, in a steady state the bus lines are connected to '0' or '1' only weakly.

Another scheme is CRF [8], which stands for complimentary regenerative feed back repeater. The schematic of the CRF scheme is shown in Fig. 4.16(b). Like the TSA scheme, the CRF circuit senses the transition of the bus voltage first. Then the driver circuit is turned on for the duration of a predetermined delay t_d. In a steady state, the output of the CRF circuit is not actively driven since the driver is in a high impedance state. Therefore, the CRF scheme is the same as the TSA scheme in that the line is connected to '0' or '1' only weakly in a steady state.

Fig. 4.17 shows noise resiliency of the proposed DRB scheme and the previously published schemes. Since the ratio of the coupling capacitance to the grounding capacitance (C_C/C_G) is now about 1.5 but will be increasing more than 3 in the future. This means that the noise induced by the coupling becomes larger. In the DRB, the noise resiliency is high because the line sections are shunted to '0' or '1' at each section. Other schemes, however, has smaller noise resiliency because the line is not shunted to '0' or '1' at each section. When coupling noise is applied, a victim line which wants to be static at '0'



(a) Transient sensitive accelerator (TSA) [7]



(b) Self-timed complementary regenerative feedback repeater (CRF) [8]

Fig. 4.16 Conventional bi-directional buffers

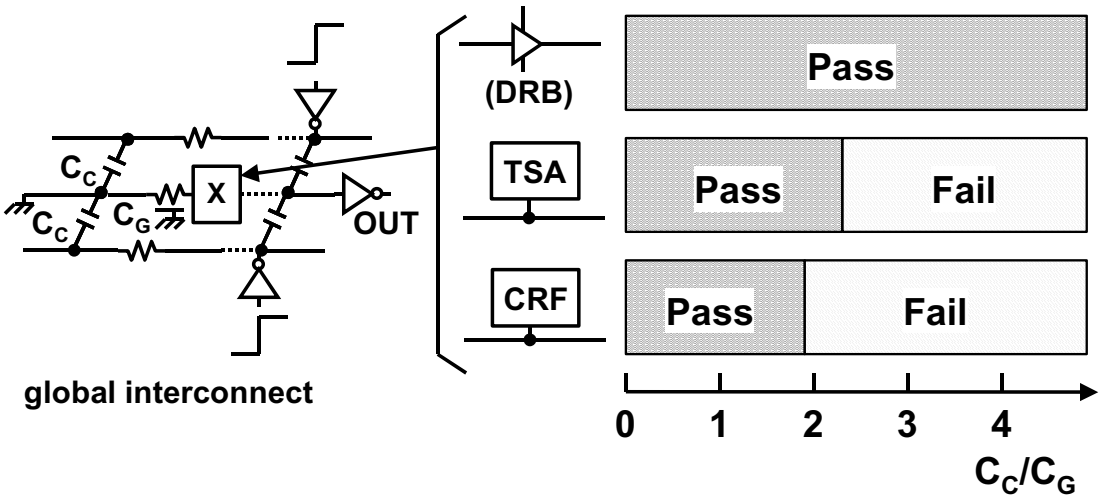


Fig. 4.17 Noise resiliency comparison among dual-rail bus scheme (DRB), transient sensitive accelerator (TSA) and self-timed complementary regenerative feedback repeater (CRF)

flips its state to '1' and flips back to '0' after a while. The transition takes several ns, which is too slow to take it as a glitch and error occurs in other schemes.

4.3.2 Staggered Firing Bus (SFB) Scheme for Uni-Directional Buses

Fig. 4.18 shows the delay fluctuation by the behavior of adjacent lines in capacitively coupled buses. Capacitive coupling induces the delay fluctuation when the adjacent lines are switching simultaneously. Even if the buffers are inserted to minimize the interconnect delay, the worst-case delay increase more than 30% compared with the normal case delay. In order to decrease the worst-case bus delay, staggered firing bus (SFB) scheme is effective.

The schematic of the SFB scheme is shown in Fig. 4.19. The interconnects are driven at a different timing by applying additional delay (firing delay) at alternate lines. The firing delay can be tuned by a couple of ways, two of which are depicted in the figure.

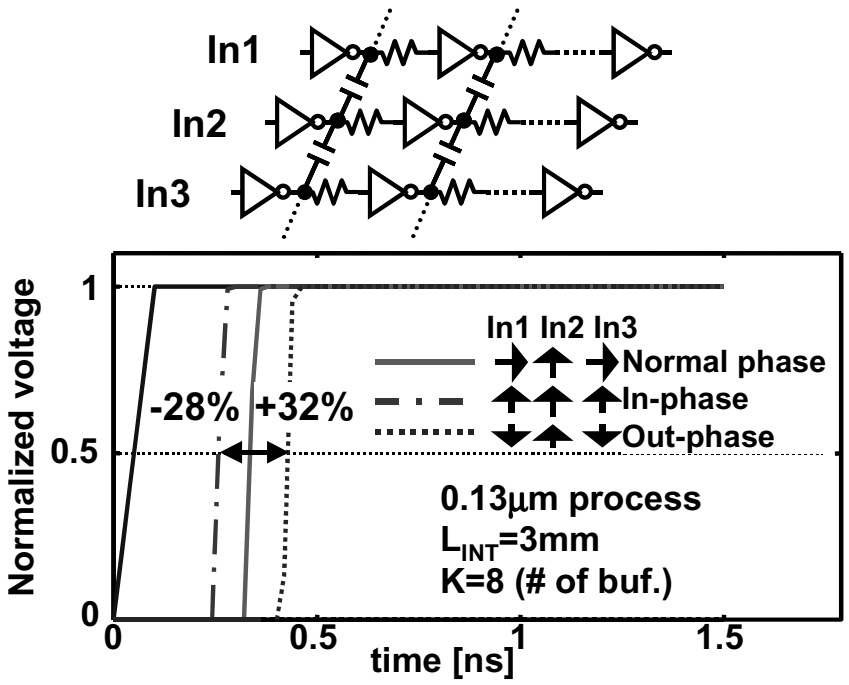


Fig. 4.18 Delay fluctuation by capacitive coupling

Fig. 4.19(a) is a tunable delay buffer which can generate the delay using two reference signals, V_{REFN} and V_{REFP}. Fig. 4.19(b) is another delay buffer which can adjust the firing delay by varying inverter stages.

The detailed operation of the staggered firing bus scheme is shown in Fig. 4.20. Here, we assume that the out-phase signals are applied simultaneously at the driving point. When the transition edge of the victim signals arrive at the first region, the transition edge of the aggressor signals have already passed through the same region due to the firing delay. Then, the victim lines do not slow down since the adjacent aggressor lines do not change when the victim lines are in transition. Similarly, the aggressor lines do not slow down since the transition edge of the victim signals have not yet arrived. Using the staggered firing bus scheme, the worst-case bus delay can be improved since the out-phase signals do not interfere each other. If the firing delay is too large, however, the firing delay itself increases the worst-case delay of the system. There is the optimum delay in the staggered firing, which is to be realized by the above-mentioned tunable delay buffer.

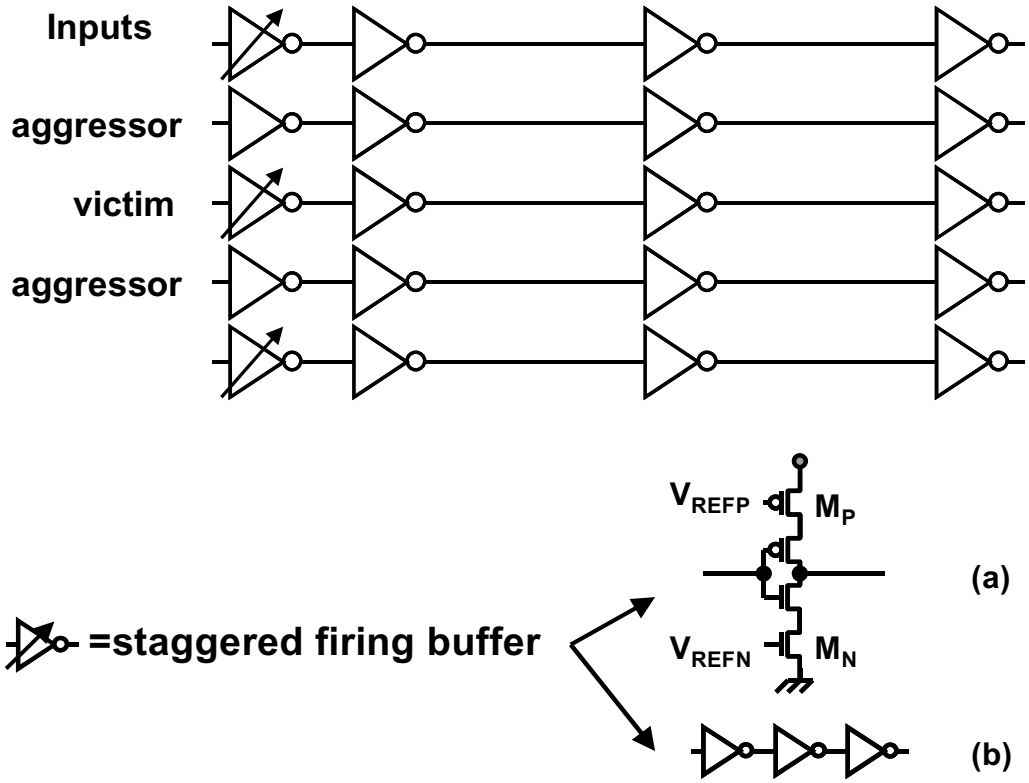


Fig. 4.19 Schematic diagram of staggered firing bus scheme

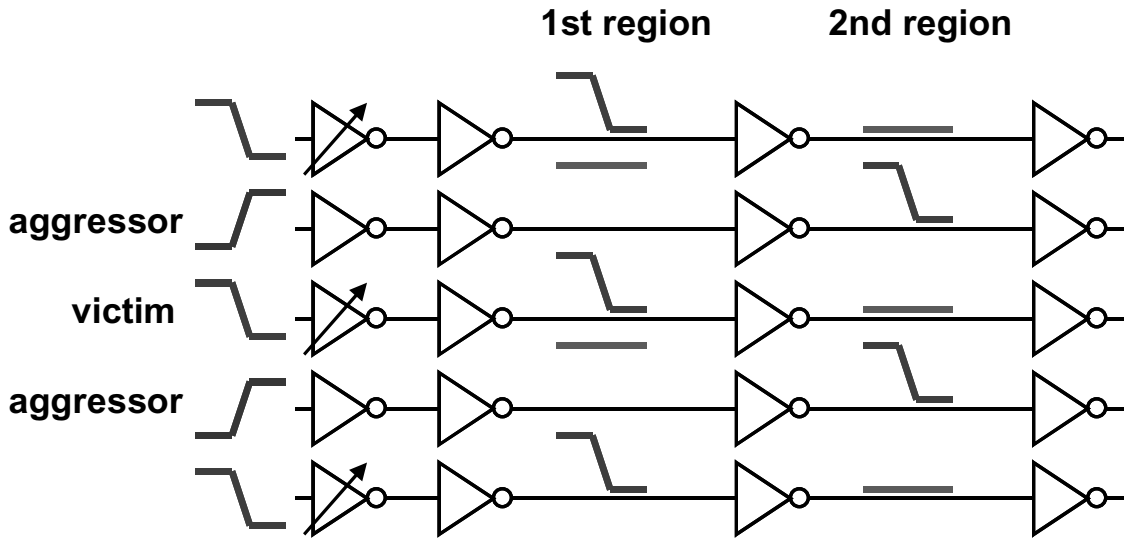


Fig. 4.20 Operation of staggered firing bus scheme

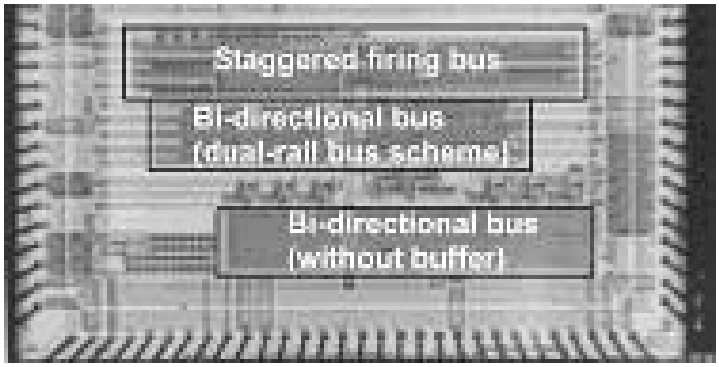
The staggered firing is not necessary for the DRB, since the left-oriented line and the right-oriented line are placed alternately and there is no chance for the adjacent signal to be out-phase.

4.3.3 Measurement Results

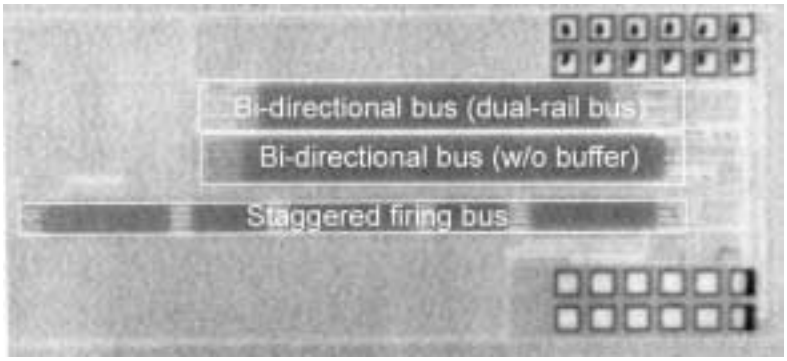
Experimental circuits of the proposed schemes are fabricated using 0.6μm CMOS technology. A microphotograph of the test chip is shown in Fig. 4.21(a). The bus lines are 60mm in length and the 8 buffers are inserted. The other experimental circuits of the proposed schemes are fabricated using 0.13μm CMOS technology. Fig. 4.21(b) shows the microphotograph of the test chip. The bus lines are 10mm in length. 5 buffers are inserted for the dual-rail bus scheme and 11 buffers are inserted for the staggered firing bus scheme.

Fig. 4.22 shows the measurement results for the bi-directional bus delay. With 0.6μm CMOS process, the interconnect delay of the proposed scheme is 31% faster than that of the conventional bi-directional buses where no buffers are inserted. With 0.13μm process, 44% delay reduction can be achieved by using the proposed dual-rail bus scheme. It is to be noted that two lines are used per bit in the proposed scheme while a single line is used per bit in the conventional bi-directional buses. In order to make the comparison fair, the line width and spacing are doubled for the conventional buses.

Fig. 4.23 shows measured results for the staggered firing bus. The firing delay is zero if the staggered firing buffer is not used. The delay shows the minimum when the firing delay is about 1ns in 0.6μm process and about 0.1ns in 0.13μm process. If the firing delay is smaller than the optimum value, the delay is increased by the capacitive coupling. If the firing delay is larger than the optimum value, the firing delay itself delays the signal propagation.



(a)



(b)

Fig. 4.21 Microphotograph

(a) 0.6μm CMOS process (b) 0.13μm CMOS process

0.6μm process (bus length (L_{INT})=60mm)

conv. (w/o buffer)	31% faster	Dual-rail bus
17.7 ns	⇨	12.2 ns

0.13μm process (bus length (L_{INT})=10mm)

conv. (w/o buffer)	44% faster	Dual-rail bus
2.25 ns	⇨	1.27 ns

Fig. 4.22 Measurement result of bi-directional buses

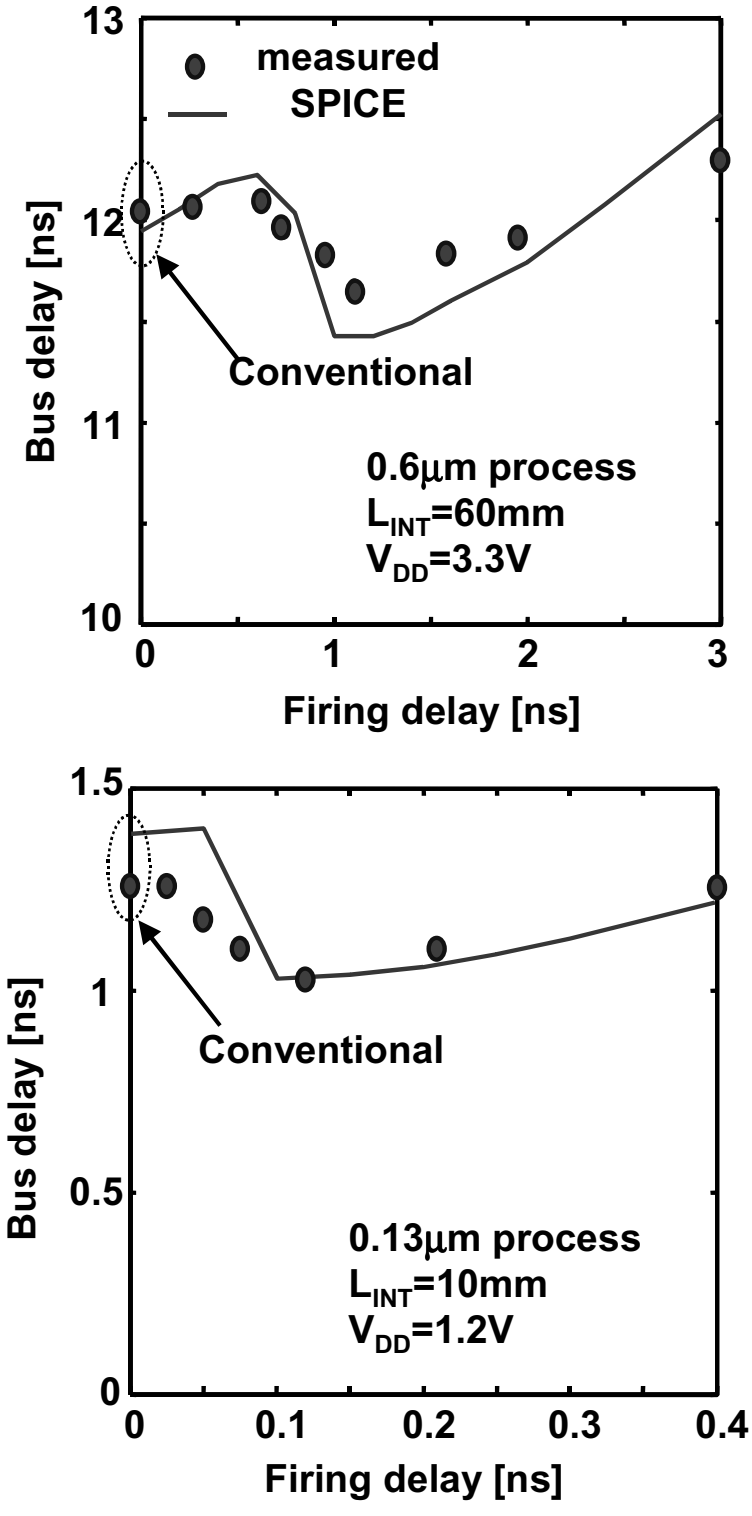


Fig. 4.23 Measured results of staggered firing buses

4.3.4 Future Trend

Since the measurement was carried out not with deep submicron interconnects and the real advantage of the proposed schemes increases as design rules scales down, the SPICE simulation is conducted to estimate the future benefit of the proposed schemes

Fig. 4.24 is the SPICE simulation results for the dual-rail bus scheme. The interconnect parameters are taken from the ITRS[6]. The length of the lines is assumed to be twice as long as a chip size. As is mentioned in 4.3.3, to make the comparison fair, the line width and spacing are assumed to be doubled for the conventional cases where no buffers are inserted and a single line is used per bit. As seen from Fig. 4.24, global interconnects can benefit from the use of the proposed scheme. When 0.07μm design rule is used, the delay is improved by an order of magnitude.

Fig. 4.25 shows the future perspective of the effectiveness of the staggered firing bus scheme. As seen from this figure, the proposed scheme can suppress the delay by about 20% at 0.18μm generation and beyond.

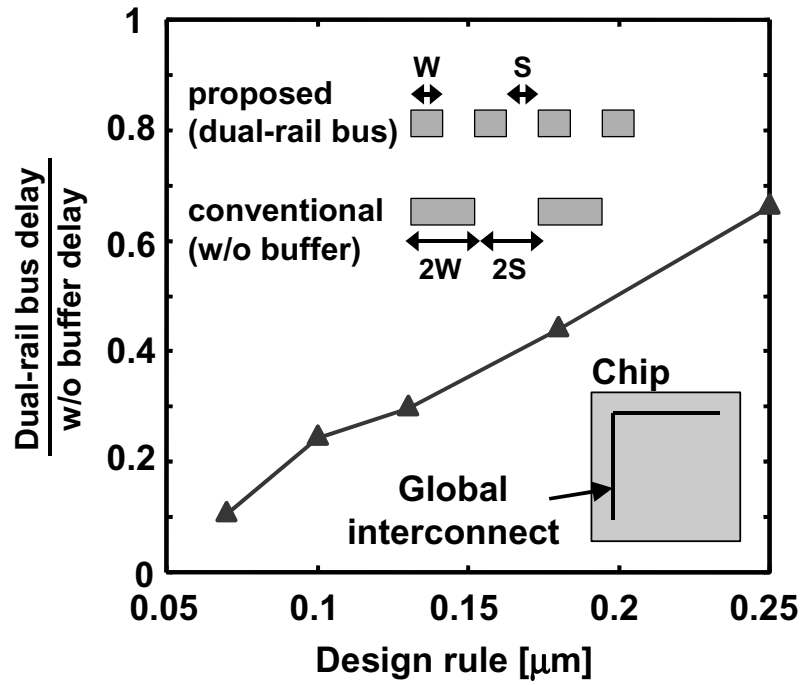


Fig. 4.24 SPICE estimation of benefit of dual-rail bus scheme for the future

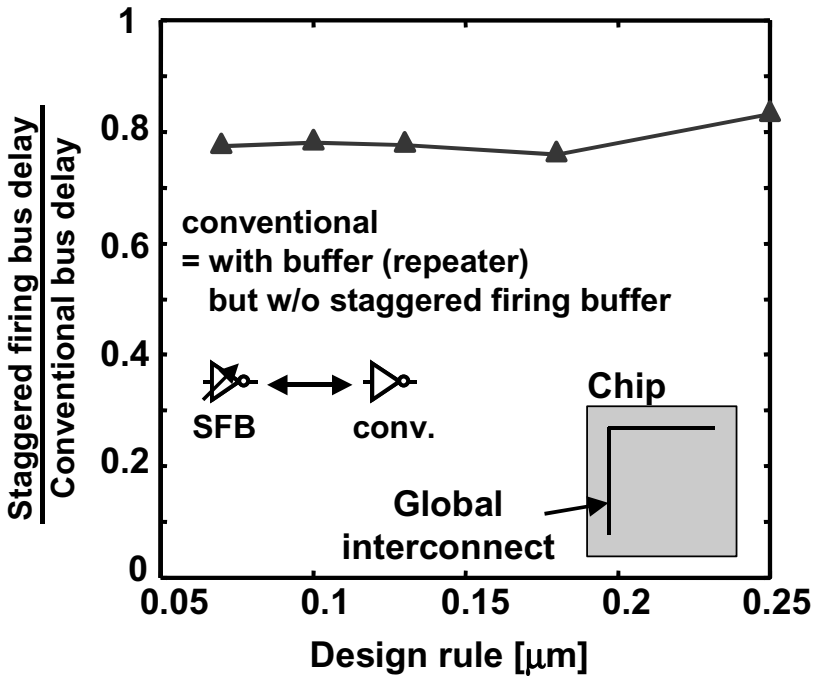


Fig. 4.25 SPICE estimation of benefit of staggered firing bus scheme for the future

4.4 Summary

In Section 4.2, closed-form formulas for optimum buffer insertion where the junction capacitance is taken into account are proposed. In order to use the derived formulas, we clarified an appropriate choice of the effective linear resistance of the driving transistor when the buffers are inserted so as to minimize the interconnect delay.

Using these formulas, the optimum interconnect delay comparison among bulk, PD-SOI, FD-SOI and double-gate structure is discussed. If the junction capacitance can be negligible, the optimum interconnect delay is 15% smaller than the delay when $C_{J0}=C_0$. MOSFET with small junction capacitance, like SOI, can suppress the interconnect delay and power by 15% compared with MOSFET with $C_{J0}=C_0$, like conventional bulk MOSFET.

In Section 4.3, a new buffer insertion scheme for bi-directional buses, namely dual-rail bus (DRB) scheme, which does not have noise problems, and a high-speed buffer insertion scheme for uni-directional buses, namely staggered firing bus (SFB) scheme, are proposed and measured. When 0.07μm design rule is used, DRB scheme can improve the performance of bi-directional buses by an order of magnitude and SFB scheme can suppress the delay of uni-directional buses by about 20% at 0.18μm generation and beyond.

It can be seen from Fig. 4.12 that the power can be reduced by 30% if the total delay is allowed to increase by 20%. Therefore, if SFB scheme is used instead of conventional uni-directional buses, 30% power reduction can be achieved while the performance of SFB is the same as that of conventional buses.

Appendix B. Deviation of Effective Linear Resistance

In order to derive R_0, one section of buffered interconnect is approximated by one-step π RC circuit connected to R_0 [4], depicted in Fig. 4.26. R_I and C_I are the interconnect resistance and interconnect capacitance of one section, respectively. C_X is the sum of C_J and $C_I/2$ and C_{OUT} is the sum of C_G (input gate capacitance) and $C_I/2$. The expression for R_0 is calculated first assuming the following points and then evaluated using rigorous simulations.

(a) Fanout is set to 1, since sections are repeated.

(b) V_X and V_{OUT} are start to fall at $T/2$ simultaneously as in Fig. 4.2.

(c) The time constant of V_{OUT} (τ_{OUT}) is twice as large as that of V_X (τ_X), as in Fig. 4.2.

(d) C_X=C_{OUT}

τ_{OUT} and τ_X are described as

$$\tau_X = R_0(C_X + C_{OUT}) \qquad (B.1)$$

$$\begin{aligned}\tau_{OUT} &= R_0(C_X + C_{OUT}) + R_I C_{OUT} \\ &= 2\tau_X\end{aligned} \qquad (B.2)$$

V_X is expressed as the function of τ_X.

$$V_X = V_{DD}\left(1 - e^{-\frac{t}{\tau_X}}\right) \qquad (B.3)$$

V_X is V_{DD} at $T/2$ and falls to $V_{DD}/2$ at $3T/4$ as is shown in Fig. 4.2. Then, T can be derived by (B.3).

$$e^{-\frac{T/4}{\tau_X}} = \frac{1}{2} \qquad \text{(B.4)}$$
$$\rightarrow T = (4\ln 2)\cdot R_0(C_X + C_{OUT}) = (4\ln 2)\cdot R_0 C$$

where C=(C_X+C_{OUT}).

The total charge which is discharged at T/2~3T/4 is written as

$$\Delta Q = \frac{1}{2}C_X V_{DD} + \frac{1}{4}C_I V_{DD} = \frac{1}{4}CV_{DD} + \frac{1}{8}CV_{DD} = \frac{3}{8}CV_{DD}\,. \qquad \text{(B.5)}$$

From the point of view of the drain current formula which can be expressed as

$$I = \beta(V_{GS} - V_{TH})^{\alpha}, \qquad \text{(B.6)}$$

the total charge supplied from the input buffer between T/2 and 3T/2 (ΔQ) can be expressed as

$$\begin{aligned}\Delta Q &= \int_{T/2}^{3T/4} \beta(V_{GS} - V_{TH})^{\alpha}\, dT \\ &= \int_{V_{DD}/2}^{3V_{DD}/4} \beta(V_{GS} - V_{TH})^{\alpha}\cdot\frac{T}{V_{DD}}\, dV_{GS} \quad , \\ &= \frac{\beta T V_{DD}^{\alpha}}{\alpha+1}\left[\left(\frac{3}{4} - v_T\right)^{\alpha+1} - \left(\frac{1}{2} - v_T\right)^{\alpha+1}\right]\end{aligned} \qquad \text{(B.7)}$$

where v_T=V_{TH}/V_{DD}.

R_5, which is the transistor resistance when V_{DS}=V_{GS}=V_{DD}, can be expressed as

$$R_5 = \frac{V_{DD}}{\beta(V_{DD} - V_{TH})^{\alpha}} = \frac{1}{\beta V_{DD}^{\alpha-1}(1 - v_T)^{\alpha}} \qquad \text{(B.8)}$$

Substituting (B.4), (B.5) and (B.8) into (B.7), following equation can be derived.

$$\begin{aligned}\Delta Q &= \frac{TV_{DD}}{R_5}\frac{1}{(\alpha+1)(1-v_T)^{\alpha}}\left[\left(\frac{3}{4}-v_T\right)^{\alpha+1}-\left(\frac{1}{2}-v_T\right)^{\alpha+1}\right] \\ &= \left(\frac{R_0}{R_5}\cdot(4\ln 2)CV_{DD}\right)\cdot\frac{1}{(\alpha+1)(1-v_T)^{\alpha}}\left[\left(\frac{3}{4}-v_T\right)^{\alpha+1}-\left(\frac{1}{2}-v_T\right)^{\alpha+1}\right] \\ &= \frac{3}{8}CV_{DD}\end{aligned} \qquad (B.9)$$

From (B.9), the effective linear resistance can be solved as the function of R_5.

$$\eta = \frac{R_0}{R_5} = \frac{3\cdot(\alpha+1)}{32\cdot\ln 2}\cdot\frac{(1-v_T)^{\alpha}}{(3/4-v_T)^{\alpha+1}-(1/2-v_T)^{\alpha+1}} \qquad (B.10)$$

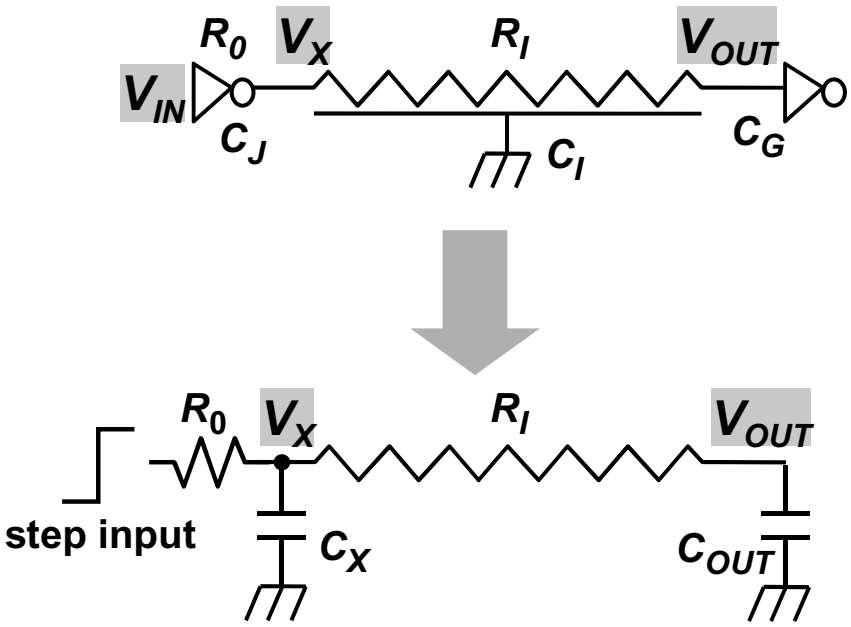


Fig. 4.26 Simple model for deviation of effective linear resistance

References

[1] Y. I. Ismail and E. G. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits," *IEEE Trans. VLSI systems*, vol. 8, No. 2, pp.195-206, Apr. 2000.

[2] K. Banerjee and A. Mehrotra, "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," *Symposium on VLSI Circuits, Dig. of Tech. Papers*, pp.195-198, 2001.

[3] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol.25, pp.584-593, Apr., 1990.

[4] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE Journal of Solid-State Circuits*, vol. SC-18, no.4, pp.418-426, Aug., 1983.

[5] T. Tanaka, H. Horie, S. Ando and S. Hijiya, "Analysis of P^+ poly Si double-gate thin-film SOI MOSFETs," *IEEE International Electron Device Meeting (IEDM)*, pp.683-686, 1991.

[6] International Technology Roadmap for Semiconductors, SIA Handbook, 1999.

[7] T. Iima, M. Mizuno, T. Horiuchi and M. Yamashina, "Capacitance coupling immune, transient sensitive accelerator for resistive interconnect signals of sub-quarter micron ULSI," *Symposium on VLSI Circuits, Dig. of Tech. Papers*, pp.31-32, 1995.

[8] I. Dobbelaere, M. Horowitz and A. E. Gamel, "Regenerative feedback repeaters for programmable interconnects," *Proc. of International Solid-State Circuits Conference*, pp.116-117, 1995.

[9] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE Journal of Solid-State Circuits*, vol.SC-18, no.4, pp.418-426, Aug., 1983.

Chapter 5. Hardware-Software Cooperative Systems for Low-Power Processors

5.1 Introduction

High-performance VLSI design with low supply voltage (V_{DD}) becomes one of the most important issues in CMOS VLSI's, since main-stream V_{DD} will be scaled down to below 0.5V in the coming years. The power and the delay dependence on the threshold voltage at 0.5V V_{DD} are shown in Fig. 5.1. As seen from the figure, the threshold voltage (V_{TH}) has to be decreased to achieve high performance. Reducing V_{TH}, however, could cause a significant increase in the static leakage power component. Especially, when the threshold voltage is lower than 0.1V, the leakage power becomes a dominant component in the total power consumption even in the active mode. In order to suppress the power consumption in low-voltage processors, it is necessary to reduce the leakage power component in the active mode.

Table 5.1 and Fig. 5.2 shows the power reduction techniques. Dual-V_{DD} [1] and

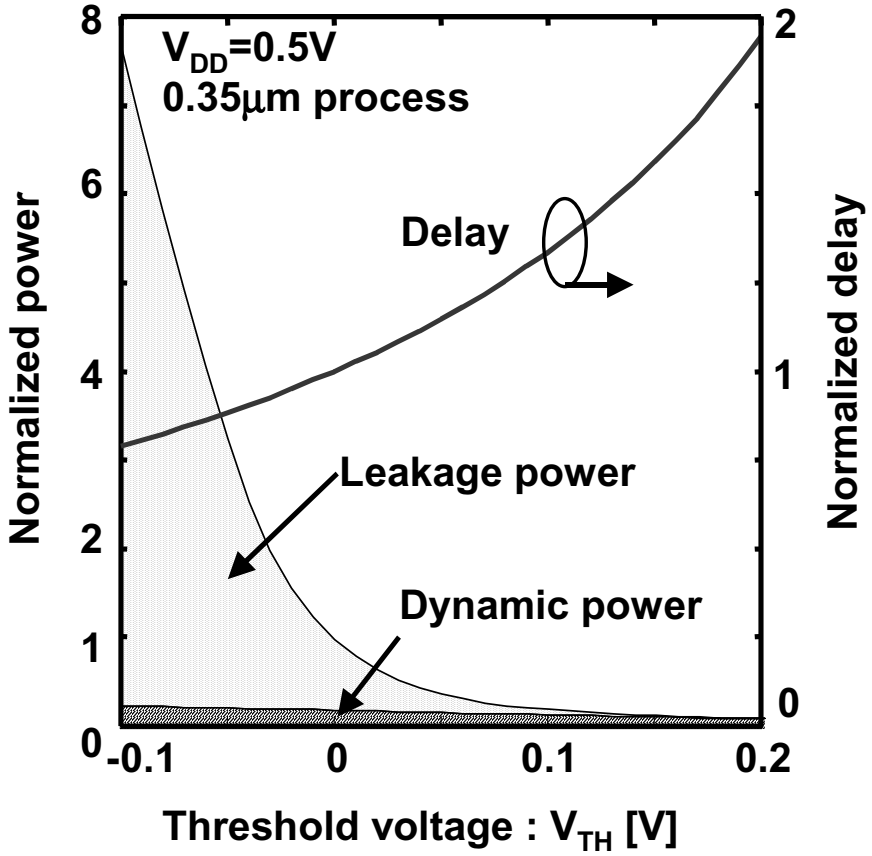


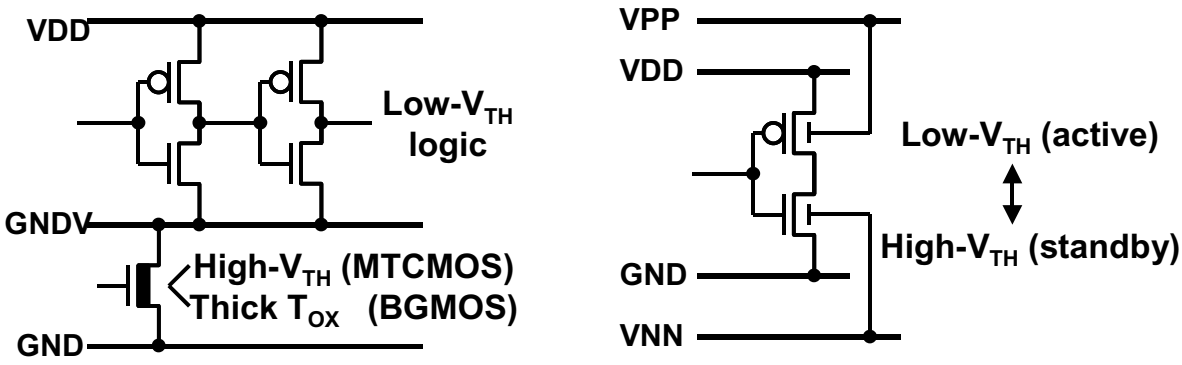
Fig. 5.1 Power and delay dependence on threshold voltage (V_{TH})

V_{DD}-hopping [2] mainly reduce the dynamic power and not the leakage power. Boosted gate MOS [3], MTCMOS [4] and VTCMOS [5] reduce the stand-by leakage power and not the power in the active mode. Thus, these schemes cannot suppress the leakage power in the active mode, which becomes the dominant component in total power consumption in low-voltage processors. Another approach is dual-V_{TH} [6]. In dual-V_{TH} technique, logic gates are partitioned into critical and non-critical paths, and low-V_{TH} transistors are only used for the logic gates in the critical paths. The drawback of dual-V_{TH} is that the leakage current cannot be sufficiently suppressed since the large leakage current always flows through the low-V_{TH} transistors. In order to suppress the active leakage power in a single-V_{TH} design, a stacked MOS [7] has been proposed. Fig. 5.3 is a diagram of the stacked MOS. The intermediate node voltage approaches V_{INT}

Table 5.1 Power reduction techniques

Target	Dynamic	Active leakage	Stand-by leakage
	Dual-V_{DD}	Dual-V_{TH}	MTCMOS
	V_{DD}-hopping	V_{TH}-hopping	BG-MOS
	DIBL-hopping		VTCMOS

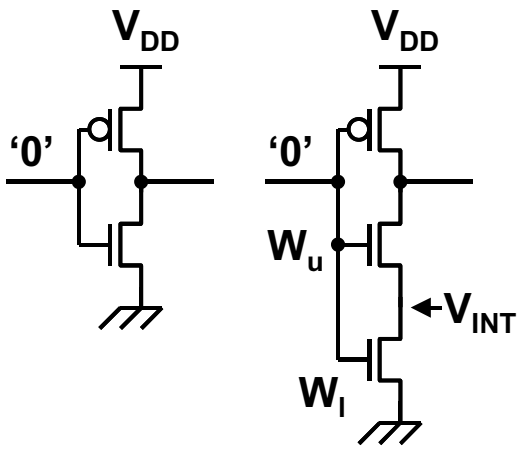
Software-hardware cooperation



(a) (b)

Fig. 5.2 (a) MTCMOS [4] and BGMOS [3] (b) VTCMOS [5]

such that the leakage currents in the upper and lower transistors are equal. In this case, the leakage current of the upper transistor decreases compared with the single off device due to the negative gate-source voltage (=$-V_{INT}$). On the other hand, the leakage current of the lower transistor also decreases since lower drain-source voltage alleviates the DIBL effect [8]. The disadvantage of this scheme is that the gate delay with the stacked MOS increases more than three times, although the leakage power can be suppressed about one



(a) (b)

Fig. 5.3 Stack effect [7]

(a) Normal MOS (b) Stacked MOS

order of magnitude. The trade-off between delay and leakage power is similar to the dual-V_{TH} scheme. Then, the leakage power cannot be suppressed sufficiently since the stacked MOS scheme can be used only for the non-critical paths.

Section 5.2 presents a dynamic threshold voltage hopping (V_{TH}-hopping) scheme that can solve above-mentioned problems. This scheme utilizes dynamic adjustment of frequency and V_{TH} through back-gate bias control depending on the workload of a processor. When the workload is decreased, less power would be consumed by increasing V_{TH}. This approach is similar to the dynamic V_{DD} scaling (DVS) [8]. In the DVS scheme, V_{DD} and the frequency are controlled dynamically based on the workload variation. The DVS, however, is effective when the dynamic power is dominant. On the other hand, V_{TH}-hopping is effective in the low V_{DD} designs where V_{TH} is low and the active leakage component is dominant in total power consumption.

As another scheme to control the active leakage power, DIBL-hopping scheme is proposed in Section 5.3. This scheme utilizes dynamic adjustment of frequency and V_{TH}

through V_{DD} control depending on a workload of a processor. V_{TH} can be changed by controlling V_{DD} due to the DIBL (Drain Induced Barrier Lowering) effect. When the workload is decreased, less power would be consumed by lowering V_{DD}. The V_{DD} control scheme is the same as the V_{DD}-hopping scheme [2]. The target of V_{DD}-hopping, however, is the reduction only in the dynamic power. On the other hand, DIBL-hopping has aimed at not only the reduction in the dynamic power but also the reduction in the active leakage power. Then, the DIBL-hopping scheme is effective in low-voltage processors where both the dynamic power and the leakage power has to be decreased.

5.2 V_{TH}-Hopping Scheme to Reduce Subthreshold Leakage

5.2.1 V_{TH}-Hopping Scheme

Fig. 5.4 shows the total power consumption depending on the workload. V_{THlow} signifies V_{TH} applied when the workload is maximum. The dynamic power (P_D) and subthreshold leakage power (P_{LEAK}) are written as

$$P_D = afCV_{DD}^2, \qquad (5.1)$$

$$P_{LEAK} = I_0 \cdot 10^{-\frac{V_{TH}}{S}} \cdot V_{DD}, \qquad (5.2)$$

where a is the switching activity, f is the operation frequency, C is the load capacitance, I_0 is the leakage current when V_{TH}=0 and S is the subthreshold slope factor. Figure 3 is calculated from these formulas.

The broken line represents a fixed V_{TH} case with only a frequency control. If the workload is less than the peak workload, frequency can be decreased to the level where the speed requirement is just satisfied. The dynamic power consumption decreases in proportion to the workload, since the dynamic power is proportional to the frequency (see (5.1)). The leakage power, however, is not reduced since it does not depend on the frequency, as is seen from (5.2). The straight line in the figure shows the power dependency of the variable V_{TH} system on the workload. When the workload is lower than the maximum workload (i.e. workload<1), the higher threshold voltage can be used while guaranteeing the logic blocks to work with the lower frequency. As is shown in Fig. 5.4, it is clear that the total power is decreased effectively with dynamic V_{TH} control

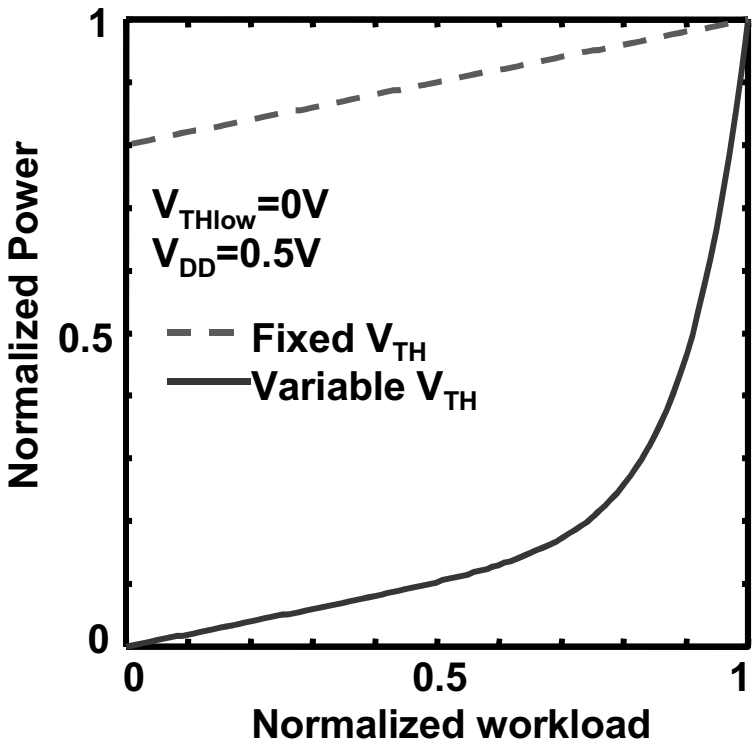


Fig. 5.4 Power dependence on workload

depending on the workload. This sets the basis for the V_{TH}-hopping.

The schematic diagram of the V_{TH}-hopping scheme is shown in Fig. 5.5. Using the control signal (CONT) which is sent from the processor, the power control block generates select signals of V_{TH}'s, VTHlow_Enable and VTHhigh_Enable, which in turn control substrate bias for the processor. When the V_{TH} controller asserts VTHlow_Enable, V_{TH} in the target processor becomes V_{THlow}. On the other hand, when the V_{TH} controller asserts VTHhigh_Enable, V_{TH} in the target processor becomes V_{THhigh}. CONT is controlled by software through a software feedback loop scheme [2], which has been proposed for dynamic V_{DD} scaling (DVS) but is also effective for V_{TH}-hopping. The software feedback scheme can guarantee hard real-time for multimedia applications with the DVS and the same algorithm guarantees the real-time operation with V_{TH}-hopping, since software-wise, the DVS and V_{TH}-hopping are the same.

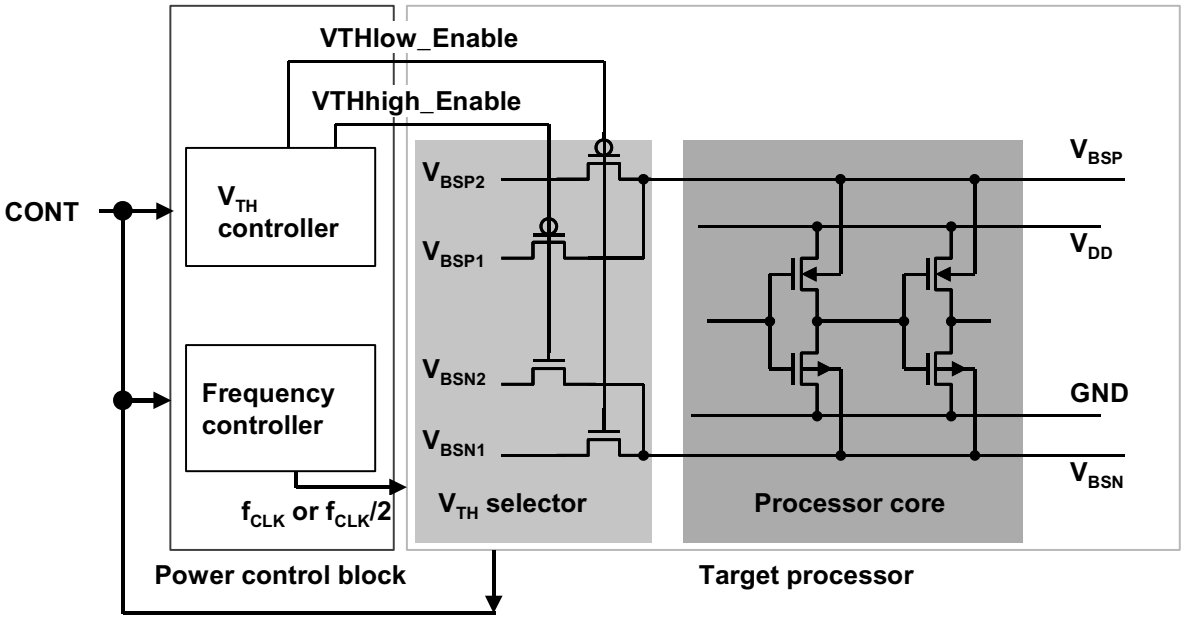


Fig. 5.5 Schematic diagram of V_{TH}-hopping

CONT also controls the operation frequency of the target processor. When the V_{TH} controller asserts VTHlow_Enable, the frequency controller generates f_{CLK}, and when the V_{TH} controller asserts VTHhigh_Enable, the frequency controller generates $f_{CLK}/2$. If necessary, the power control block can be extended so that more than two sets of frequency and threshold voltage can be generated. In order to avoid the synchronization problem at the interface of the processor with the external systems, the frequency has only discrete values of $f_{CLK}, f_{CLK}/2, f_{CLK}/3, \cdots$.

V_{THlow} is determined so that the maximum performance of the processor achieves the required clock frequency of f_{CLK}. On the other hand, V_{THhigh} is determined so that the processor operates at $f_{CLK}/2$.

Fig. 5.6 shows the power and the performance dependence on the back-gate bias (V_{BS}). The back-gate bias for V_{TH}-hopping is not only limited to negative value but also can be positive. The negative back-gate biasing is effective in the low-V_{TH} design in which the active leakage power is dominant. Using the negative back-gate biasing, the active

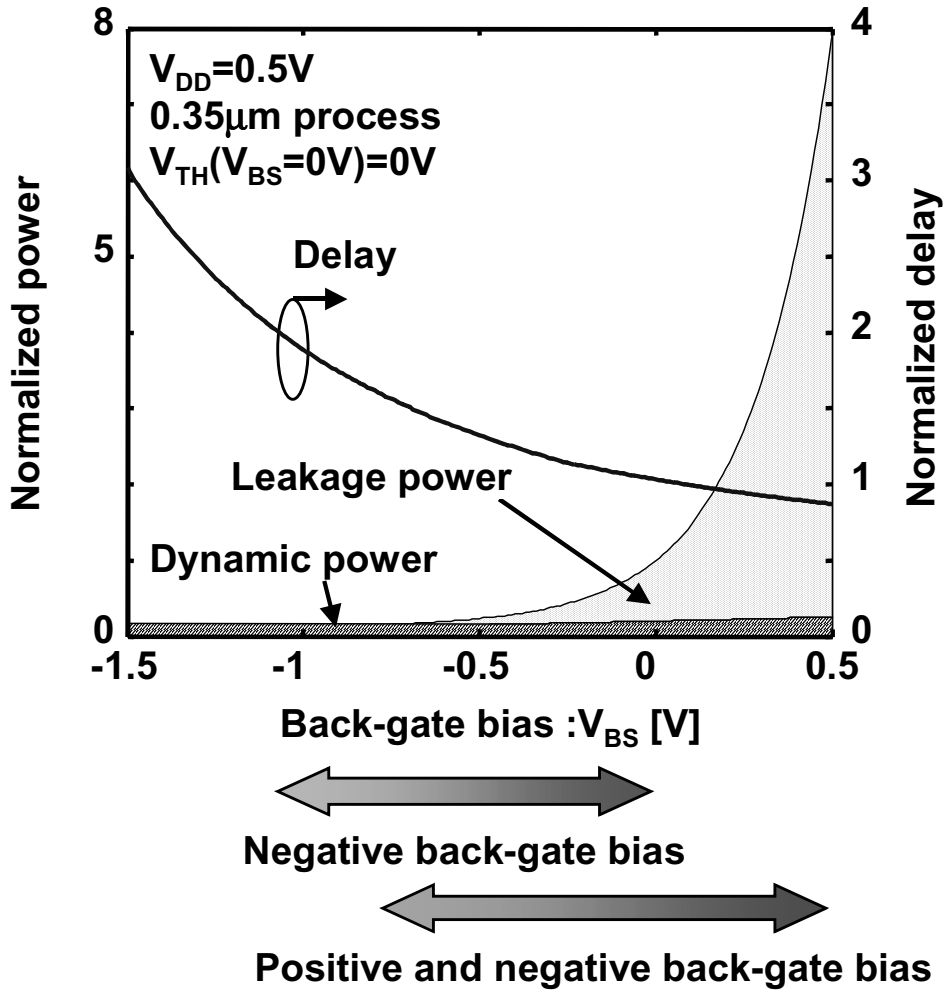


Fig. 5.6 Power and delay dependence on back-gate bias (V_{BS})

leakage power can be suppressed effectively. It is suspected, however, that the strong negative biasing may be difficult in the future since the strong negative biasing enhances a short-channel effect and the band-to-band tunneling, which is called BTBT and induces leakage [10][11]. In order to improve the effect of V_{TH}-hopping, a positive and negative combined back-gate bias scheme would be ideal. The lowest V_{TH} is achieved by positive back-gate bias and the highest V_{TH} is obtained by negative back-gate bias. Compared with the negative back-gate bias scheme, the effect of V_{TH}-hopping with the positive and negative combined back-bias scheme is improved since the wider range of the threshold voltage can be realized when the negative back-gate bias is limited.

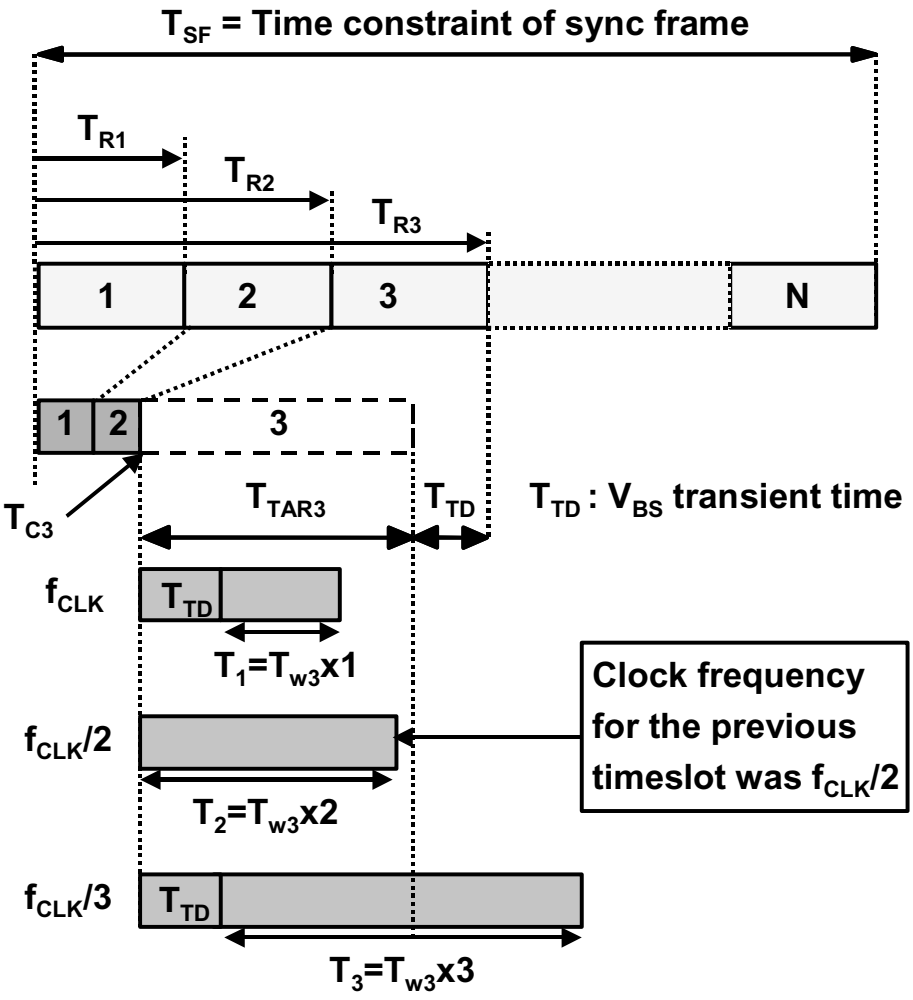


Fig. 5.7 Determination of V_{TH} and frequency

The algorithm to adaptively change V_{TH} depending on the workload is of importance. Since the workload strongly depends on data, the control should be dynamic in real-time, and should not be static at compile time. On the other hand, it is impossible to predict the workload of the task to be done in the future without error.

In order to solve this problem, the algorithm with software feedback loop, which is shown in Fig. 5.7, is used for V_{TH}-hopping. Most real-time applications have a given time interval in which a certain amount of tasks should be executed. For example, real-time MPEG4 application performs video coding at 15 frames per second. This time interval is called a sync frame (T_{SF}). Here, the following algorithm is used to guarantee

the real time execution.

(a) Every sync frame is divided into N slices called timeslots. The frequency and the threshold voltage of the target processor are determined for each timeslot.

(b) For each timeslot, target execution time, T_{TAR}, is calculated. The execution time accumulated from 1st to $(i\text{-}1)^{th}$ timeslots, T_{Ci}, can be taken from the internal timer in the power control chip. The target execution time for timeslot i, T_{TARi}, is calculated as $T_{TARi}=T_{Ri}-T_{Ci}-T_{TD}$, where T_{TD} is the transition delay to change a clock frequency and threshold voltage and T_{Ri} is execution time limit of the timeslot i.

(c) For each clock frequency (f_{CLK}/j, j=1,2,3,…), estimated worst case execution time (WCET) is calculated as $T_j=T_{Wi}\times j$. T_{Wi} is the worst case execution time of timeslot i. There is no transition delay (T_{TD}) if the clock frequency is the same as the clock frequency used in the previous timeslot. On the other hand, if the frequency is not equal to the previous clock frequency, $T_j=T_j+T_{TD}$.

(d) The clock frequency is determined as a minimum frequency whose estimated worst case execution time does not exceed the target time (T_{TARi}).

Thus, the frequency and V_{TH} are dynamically controlled on a timeslot-by-timeslot basis inside each task by software.

This algorithm is based on the concept of the run-time voltage hopping scheme [2]. The algorithm can be applied to such real-time applications whose WCET (worst case execution time) is known for example, MPEG2 and VSELP speech encoding. In [2], three typical real-time applications such as MPEG4 video encoding, MPEG2 video decoding and VSELP speech encoding were simulated and the effectiveness of the scheme has been verified. As for robustness, this algorithm guarantees the hard real-time execution of an application if a processor can execute the application in real-time with the constant higher frequency.

5.2.2 Simulation Results of MPEG4 Encoding using V_{TH}-Hopping

In order to show the effectiveness of the scheme, performance evaluation is conducted using MPEG-4 video coding.

Fig. 5.8 shows a simulation result of power transition in time for MPEG4 encoding case using V_{TH}-hopping. In this simulation, the transition delay (T_{TD}) is set to 0.5ms. If more than two clock levels, hence more than two V_{TH} levels, are provided, more power reduction is possible but the improvement is minor (only 6%) as is shown in the figure. Moreover, if more levels are provided, there are test issues since speed test should be run at more than two frequencies and more area overhead is needed for the control block and selectors, and controlling V_{TH} through V_{BS} becomes difficult since the negative back-gate bias for $f_{CLK}/3$ and slower is higher than 1V where the short-channel effects and the band-to-band tunneling are enhanced. This is why the number of V_{TH} levels is limited to two. Only two levels, that is, f_{CLK} and $f_{CLK}/2$, are sufficient, meaning that the proposed scheme is simple, in both software and hardware.

It is seen from Fig. 5.9 that f_{CLK} is used only 6% of the time while the processor is run at $f_{CLK}/2$ for 94% of the time. f_{CLK} is still needed because the processor will run at f_{CLK} for 100% of the time when the worst-case data comes, which is very unlikely and for most of the time, the workload is about a half on average. This tendency holds for other applications such as MPEG2 decoding and VSELP voice codec.

Fig. 5.10 shows the simulation result of a power comparison among fixed single V_{TH}, dual-V_{TH} and V_{TH}-hopping cases for MPEG4 encoding. The dual-V_{TH} [6] can reduce the power only to 65% of the fixed single V_{TH} case since the leakage power of the low-V_{TH} gates cannot be suppressed. V_{TH}-hopping can reduce the power to 18% of fixed low-V_{TH} circuit and 27% of the dual-V_{TH} scheme in 0.5V V_{DD} regime. Thus, V_{TH}-hopping is

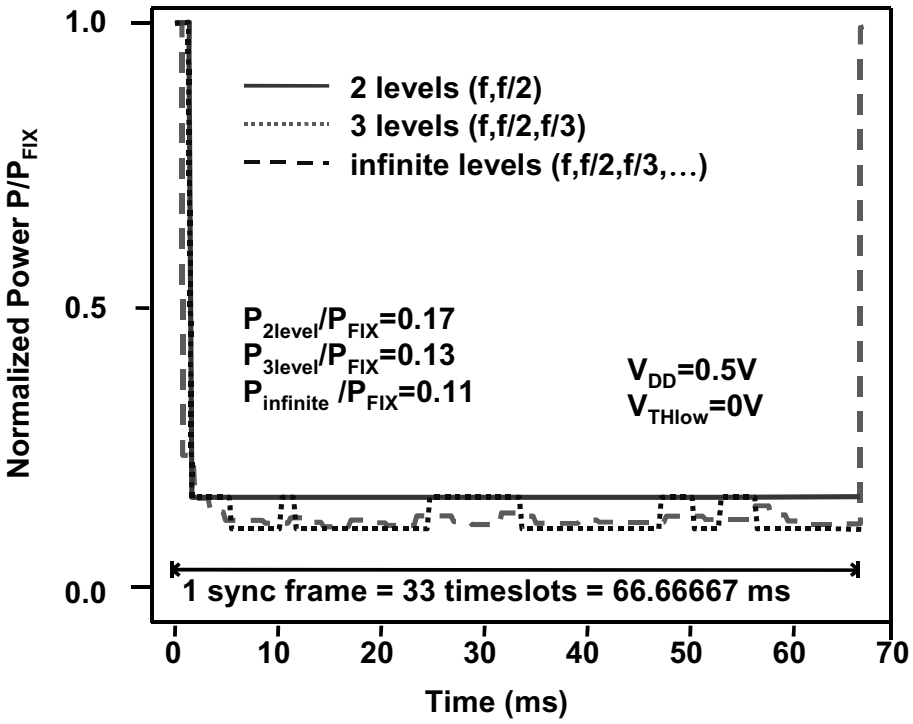


Fig. 5.8 Power transition of V_{TH}-hopping

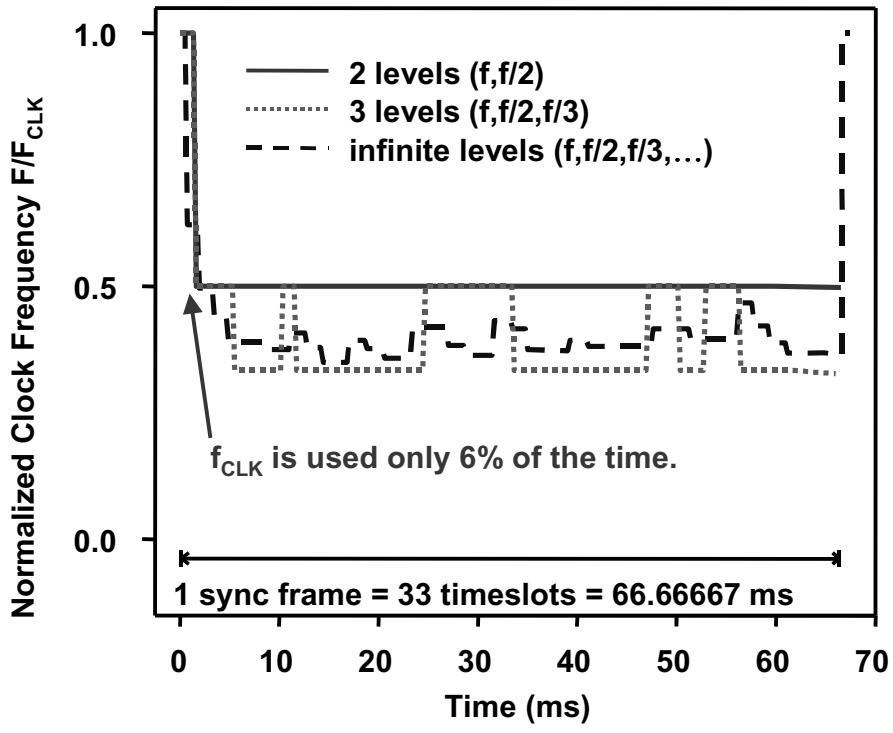


Fig. 5.9 Frequency transition of V_{TH}-hopping

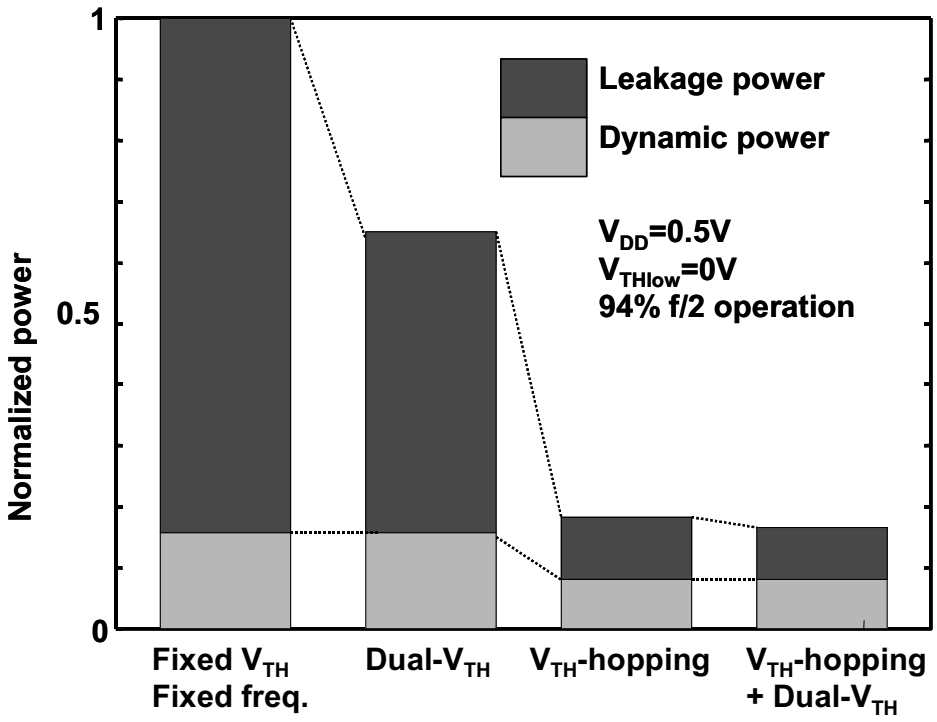


Fig. 5.10 Power comparison among single fixed V_{TH}, dual-V_{TH} and V_{TH}-hopping

effective in the low supply voltage design where the threshold voltage is low and the active leakage component is dominant in total power consumption.

In order to suppress the leakage power further, combining the V_{TH}-hopping scheme and the dual-V_{TH} scheme could be useful. Fig. 5.11 shows the schematic of this scheme. In this scheme, V_{TH}-hopping is used only in the critical paths. On the other hand, V_{TH} of the non-critical gates is set to considerably higher value (V_{THnon_crit}), which is not changed for all the time.

As shown in Fig. 5.10, however, the above mentioned combination scheme hardly improves the power (only 9%) compared with the V_{TH}-hopping scheme. The reason is that the difference between the leakage power in the critical paths and the leakage power in the non-critical paths is small since the leakage power in the critical paths has already been suppressed by using V_{TH}-hopping. Therefore, it can be said that the scheme using only V_{TH}-hopping is the most effective.

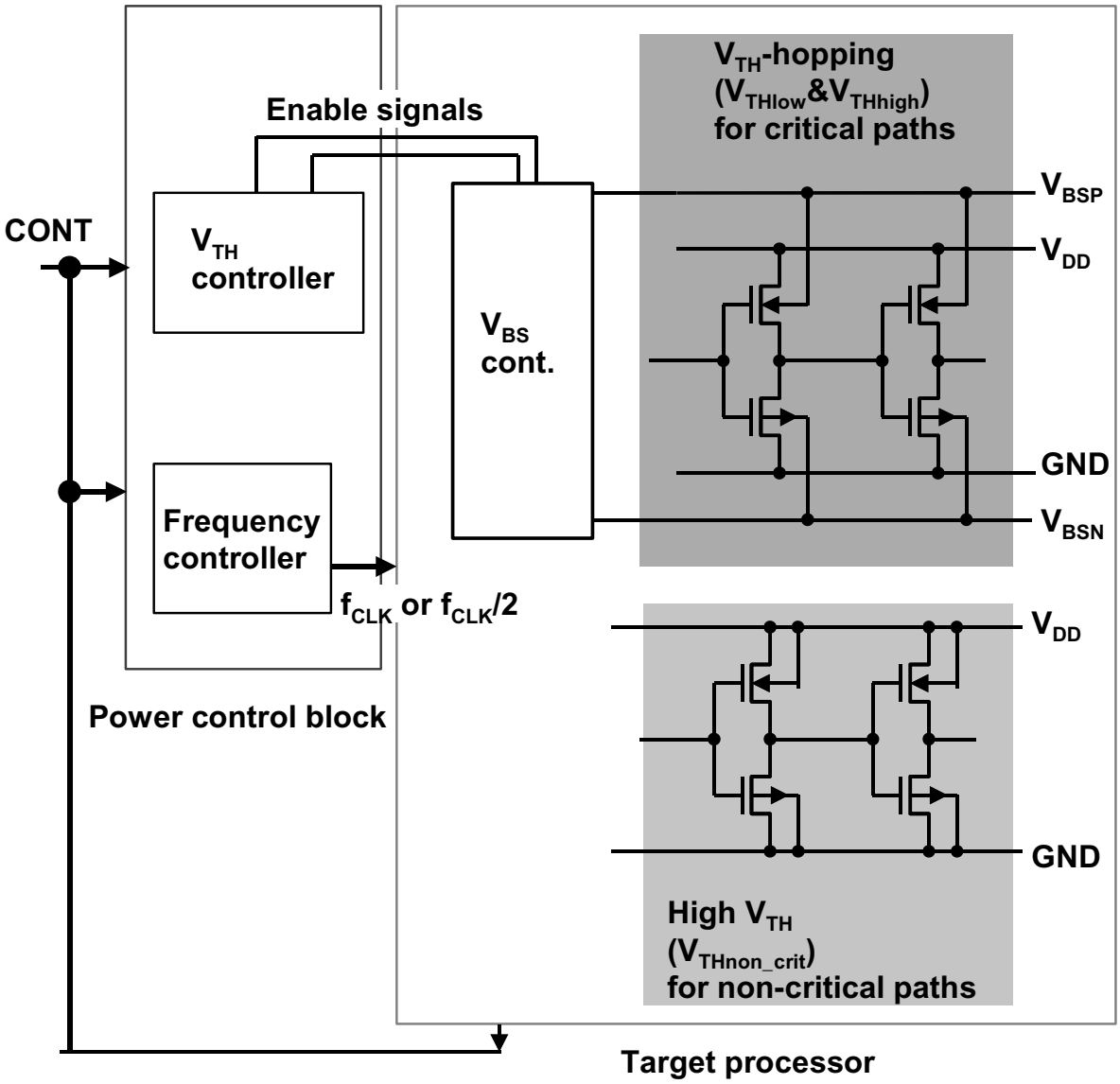


Fig. 5.11 Schematic diagram of design which combines V_{TH}-hopping and dual-V_{TH}

5.2.3 Measurement of RISC Processor with V_{TH}-hopping

As is mentioned in Section 3, the negative back-gate bias scheme and positive and negative back-gate bias scheme are effective in the low-V_{TH} design where the active leakage power is dominant. The negative back-gate biasing, however, has little effect on the total power in the conventional high threshold voltage designs where the active leakage power is much smaller than the dynamic power. On the other hand, the positive back-gate bias scheme is compatible with the conventional high threshold voltage design. The performance of the processor can be improved since lower threshold voltage can be achieved by positive back-gate biasing [12]. The drawback of this scheme is the forward junction leakage current, which occurs between drain and back-gate, increases exponentially.

In order to suppress the forward junction leakage current, combining V_{TH}-hopping and the positive back-gate bias is effective. For example, the low threshold voltage (V_{THlow}) is realized by positive back-gate bias to improve the performance and the high threshold voltage (V_{THhigh}) is achieved by zero back-gate bias to suppress the leakage power.

A small scale RISC processor with V_{TH}-hopping capability and the positive back-gate bias scheme is fabricated in a 0.6μm CMOS technology. The area overhead of the V_{TH}-hopping scheme is 14 %. This includes the additional V_{BSP} and V_{BSN} lines in the standard cell area and the area of V_{BS} selector. A microphotograph of the RISC processor appears in Fig. 5.12. The size of RISC core is 2.1 mm × 2.0 mm and the size of the V_{BS} selector is 0.2mm × 0.6mm.

In normal standard cells, the n-well bias voltage and the p-well bias voltage are fixed at V_{DD} and ground respectively, since the well contacts and substrate contacts are connected to the V_{DD} and ground lines. In order to design a processor with V_{TH}-hopping, it is

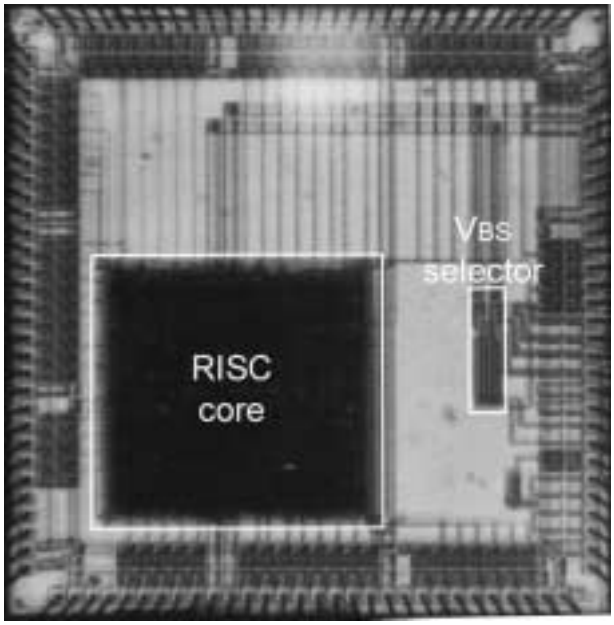


Fig. 5.12 Microphotograph of RISC processor

necessary that the n-well bias and the p-well bias can be changed freely. In this study, a simple yet effective design methodology for the V_{TH}-hopping is adopted using the commercially available CAD tools [13][14] and normal standard cells. Fig. 5.13 shows the detailed process of the place and route (P&R) [13] for V_{TH}-hopping, which is summarized as follows.

Place and route is executed using the conventional standard cells. In order to add metal lines for V_{BSP} and V_{BSN}, the standard cells are placed at appropriate intervals, which can be done by using the conventional place and route tool with an appropriate parameter.

(a) Well contacts located on the V_{DD} line and well contacts (or substrate contacts) located on the ground line are removed by using SKILL script [14].

(b) The n-well pattern, p-well pattern, V_{BSP} lines, V_{BSN} lines and well/substrate contacts are added to the gap between the standard cells.

(c) The advantage of this technique is the standard cells need not be modified at all. If the standard cells can be modified, the area overhead could be reduced to 9%.

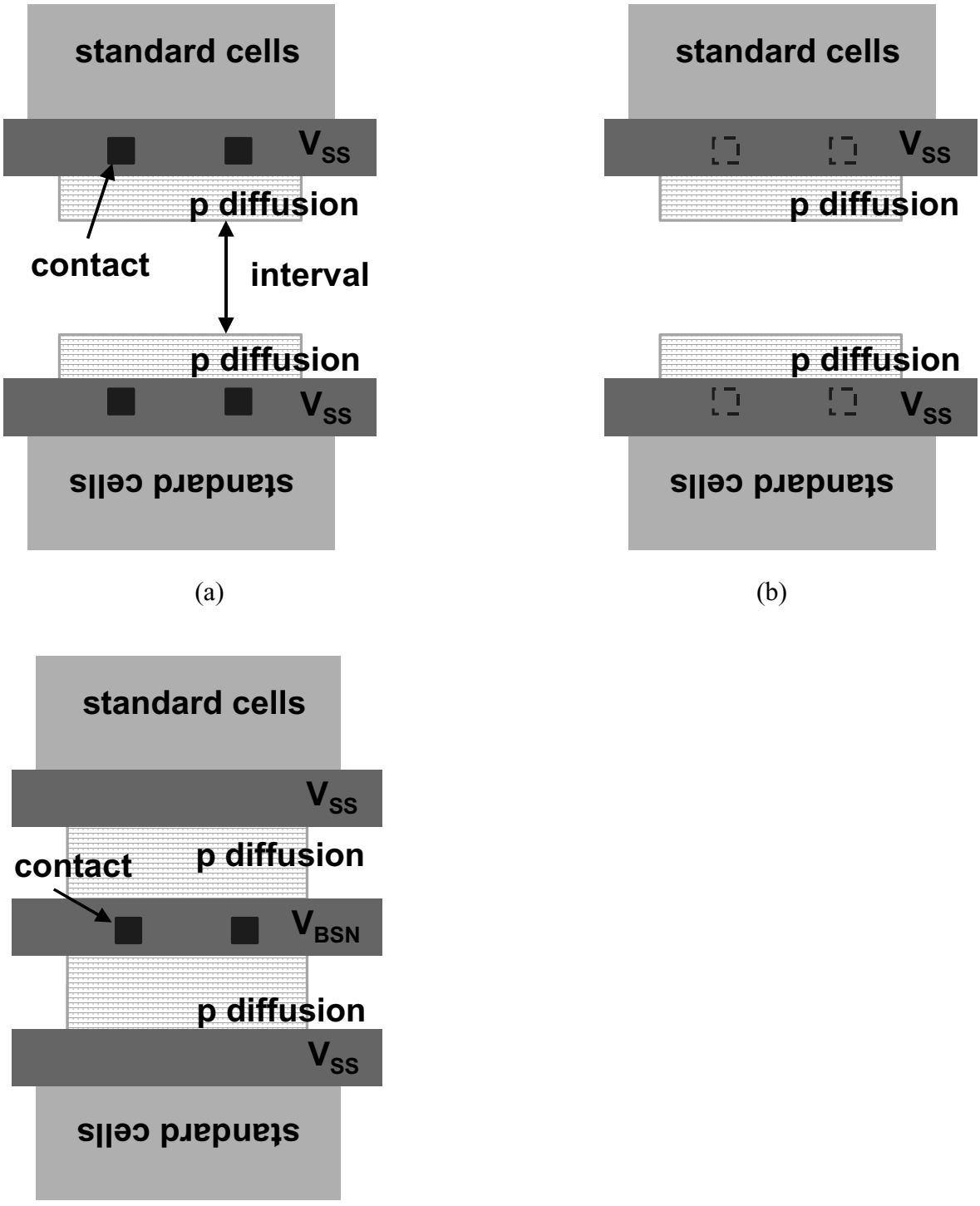


Fig. 5.13 Place and route using conventional standard cells

Fig. 5.14 shows the measurement results and SPICE simulation results of the RISC processor using simple hand-coded programs. V_{FW} is the positive back-gate bias voltage and ΔV_{FW} is the peak-to-peak V_{FW} variation which is set to 0.1V (±6% of V_{DD}). The variation of V_{FW} includes process variation, temperature variation and noise of V_{BSN} and V_{BSP} lines. We assumed that the lowest positive back-gate bias is 0.6V and the highest positive back-gate bias is 0.7V. When the positive back-gate bias is asserted, the worst delay occurs at the lowest V_{FW}. The delay improves 29% at 0.9V V_{DD} with 0.6V V_{FW}. On the other hand, the worst-case leakage power occurs at the highest back-gate bias voltage, which is 0.7V in this case. The leakage power increases exponentially when V_{FW} is higher than 0.6V due to the forward junction leakage. If zero back-gate bias is applied, 91% power reduction can be achieved compared with the fixed 0.7V positive back-gate bias scheme.

In order to verify the effectiveness of V_{TH}-hopping with positive back-gate bias scheme, MPEG4 encoding is simulated based on the measured data. The simulation result shows that 86% power saving can be achieved by using V_{TH}-hopping compared with the fixed positive back-gate bias scheme.

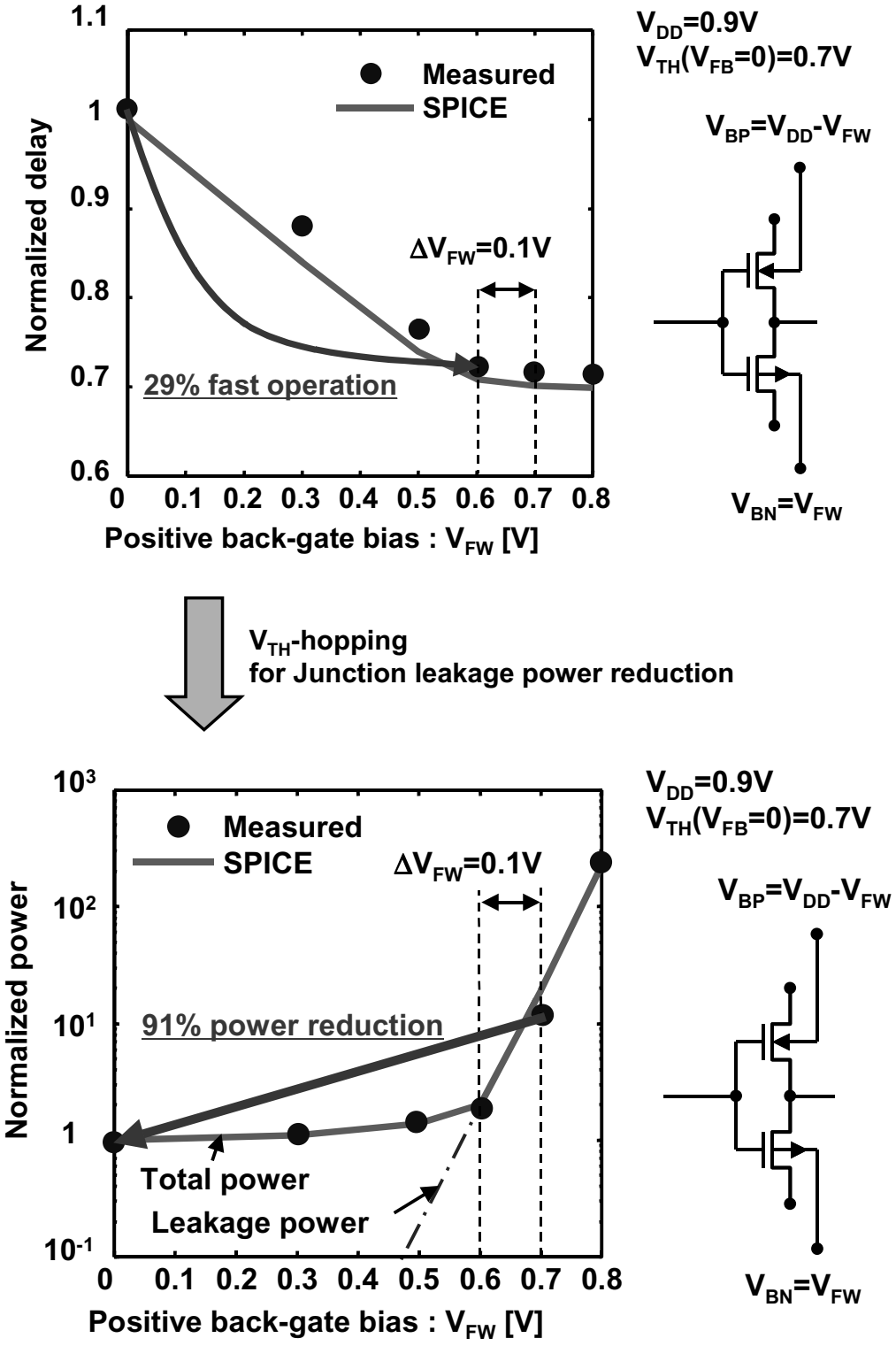


Fig. 5.14 Measured results of delay and power of V_{TH}-hopping with positive back-gate bias

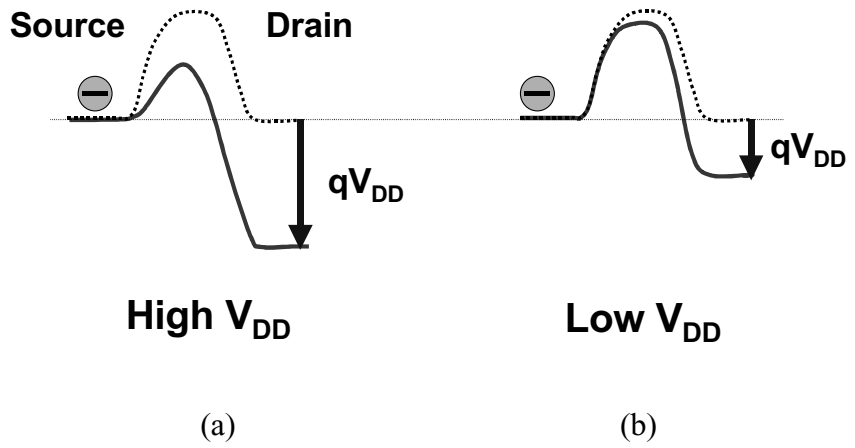
5.3 DIBL-Hopping Scheme

5.3.1 Schematic of DIBL-Hopping

Fig. 5.15 shows the schematic of the potential barrier diagram versus lateral distance from the source to the drain. The gate voltage is assumed to be 0. In short-channel MOSFET, the potential barrier between drain and source decrease since the source and drain field lengthen into the middle of the channel. This causes the increase of the subthreshold leakage current. When a high drain voltage is applied to the short-channel MOSFET, the subthreshold leakage current increases further since the barrier between drain and source is more lowered (Fig. 5.15(b)). This effect is called drain induced barrier lowering (DIBL) [8].

Conversely, the subthreshold leakage current can be suppressed if the drain voltage, that is, the supply voltage is lowered.

Fig. 5.16 shows the total power consumption depending on a workload. The straight line represents a fixed V_{DD} case with only a frequency control. If the workload is less than the peak workload, frequency can be decreased to the level where the speed requirement is just satisfied. The dynamic power consumption decreases in proportion to the workload, since the dynamic power is proportional to the frequency. The leakage power, however, is not reduced since it does not depend on the frequency. The broken line in the figure shows the power dependency of the variable V_{DD} system on the workload. When the workload is lower than the maximum workload (i.e. workload<1), lower V_{DD} can be used while guaranteeing the logic blocks to work with the lower frequency. It is clear that the total power is decreased effectively with dynamic V_{DD} control depending on the workload. This sets the basis for the DIBL-hopping.



(a) (b)

Fig. 5.15 Surface potential versus lateral distance

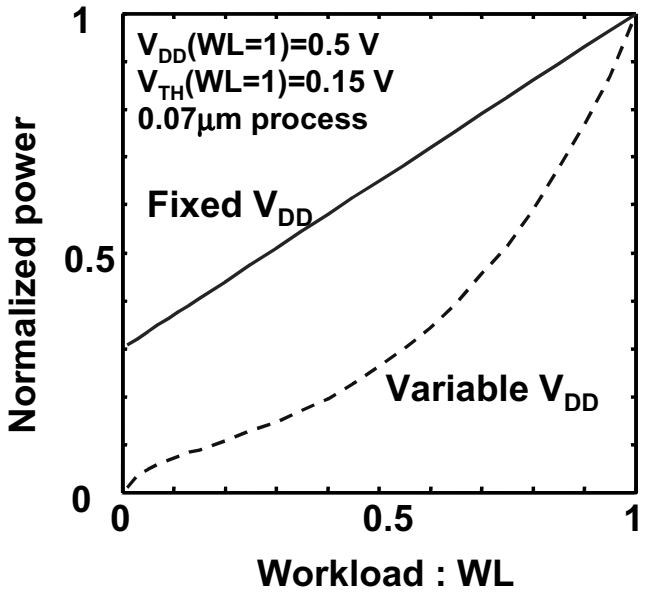


Fig. 5.16 Power dependence on workload

Fig. 5.17 shows a schematic of the DIBL-hopping. The higher supply voltage, V_{DDH}, is asserted while the processor is run at a maximum frequency, f_{CLK}. On the other hand, when the half of the maximum frequency, $f_{CLK}/2$, is assigned, the supply voltage is lowered to V_{DDL}. The power reduction can be achieved by following effects.

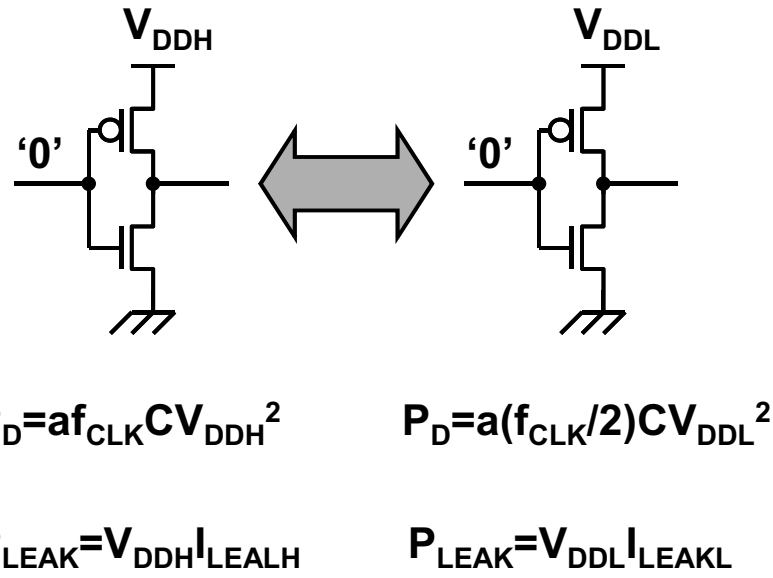


Fig. 5.17 Schematic of DIBL-hopping

(1) The dynamic power is suppressed due to the lowered frequency ($f_{CLK} \rightarrow f_{CLK}/2$) and the lowered supply voltage ($V_{DDH} \rightarrow V_{DDL}$).

(2) The subthreshold leakage current decreases since the DIBL effect is alleviated and the leakage power is reduced by lowering V_{DD}.

(3) The gate tunnel leakage [15], which is occurred by the thin gate electrode, is reduced since the gate voltage is lowered.

The algorithm to adaptively change V_{DD} depending on the workload is of importance. In DIBL-hopping, V_{DD} is controlled by software through a run-time voltage hopping scheme [2]. This algorithm is used in V_{TH}-hopping scheme (see Section 5.2.1). The run-time voltage hopping scheme can guarantee hard real-time for multimedia applications with the V_{DD}-hopping and the same algorithm guarantees the real-time operation with DIBL-hopping, since software-wise, the V_{TH}-hopping and DIBL-hopping are the same.

5.3.2 Simulation Results of DIBL-Hopping

In order to confirm the effectiveness of the DIBL-hopping scheme, the SPICE simulation is evaluated. The MOSFET model for the SPICE simulation is used the predictive technology model (PTM) [16] which is provided by the device group at the U. C. Berkeley. The higher supply voltage, V_{DDH}, is set to the value listed in ITRS (International Technology Roadmap for Semiconductors) [17]. Table 5.2 shows the main parameters which are extracted from the characteristics of PTM and ITRS. The accurate relation between V_{TH} and V_{DD} is complicated, but in the SPICE models, the relation is simply expressed as [18]

$$\Delta V_{TH} = \lambda \cdot \Delta V_{DD} = \lambda (V_{DDH} - V_{DDL}), \tag{5.3}$$

where λ is called a DIBL factor.

Fig. 5.18 shows the SPICE simulation results and the calculated results of the inverter loop whose fanout is 1. V_{DDL} is determined so that the delay with V_{DDL} is twice as large as the delay with V_{DDH}. The leakage power ratio, Pratio, is the ratio of $P(V_{DDL})$ (the subthreshold leakage power when V_{DDL} is assigned) to $P(V_{DDH})$ (the subthreshold leakage power when V_{DDH} is assigned). In this simulation, the gate tunnel leakage current is not taken into account since the present BSIM3 model does not support the gate tunnel leakage.

On the other hand, the calculated result is derived from the following simple delay and power expressions.

$$P_{LEAK} \propto 10^{-\frac{V_{TH}+\Delta V_{TH}}{S}} \cdot V_{DD} \tag{5.4}$$

Table 5.2 Parameter sets of PTM model

Channel length [μm]	0.18	0.13	0.10	0.07
V_{DDH} [V]	1.8	1.45	1.22	1
V_{TH} [V]	0.3	0.27	0.25	0.2
α (@V_{DS}=V_{DDH})	1.25	1.4	1.3	1.3
S-factor (mV/decade)	87	87	89	99
DIBL factor (λ)	0.038	0.071	0.098	0.111

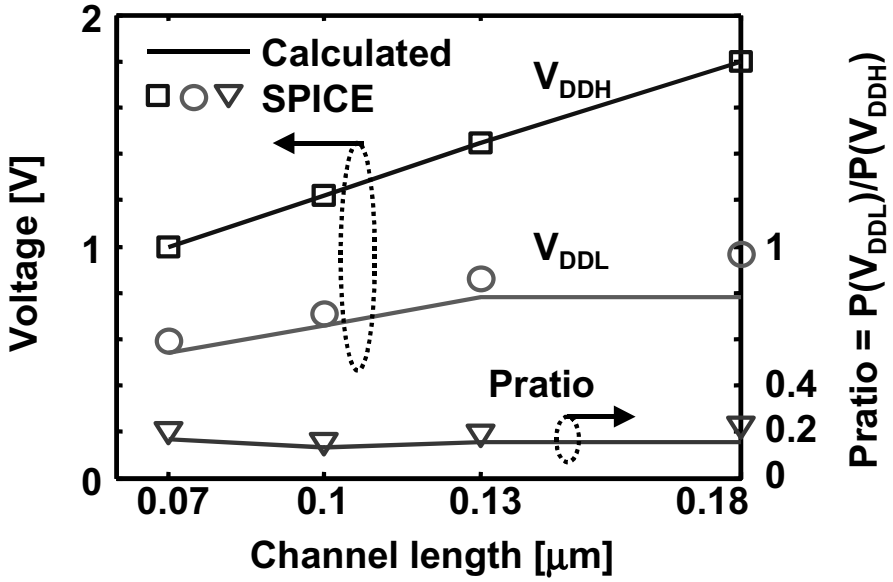


Fig. 5.18 Leakage power ratio (Pratio) of DIBL-hopping

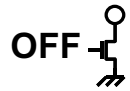
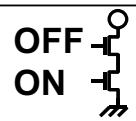
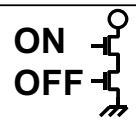
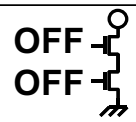
$$T_d \propto \frac{V_{DD}}{(V_{DD} - (V_{TH} + \Delta V_{TH})^{\alpha}} \quad (5.5)$$

From (5.3) and (5.5), the relation between V_{DDH} and V_{DDL} can be expressed as

$$\frac{V_{DDL}}{(V_{DDL} - (V_{TH} + \lambda \cdot (V_{DDH} - V_{DDL})))^{\alpha}} = 2 \cdot \frac{V_{DDH}}{(V_{DDH} - V_{TH})^{\alpha}} \quad (5.6)$$

Substituting the calculation result of V_{DDL} in (5.4), the leakage power ratio can be derived.

Table 5.3 Effectiveness of DIBL-hopping with series-connected MOSFET

structure	$P(V_{DDH})$	$P(V_{DDL})$	$\frac{P(V_{DDL})}{P(V_{DDH})}$
	1.15e-7	1.86e-8	16.2%
	1.15e-7	2.00e-8	17.4%
	6.43e-8	1.38e-8	21.5%
	5.93e-9	3.05e-9	51.4%

The simulation and calculated results show that if the supply voltage is lowered to V_{DDL}, 80% leakage power reduction is possible compared with the fixed V_{DDH} scheme. The result that the power ratio is about 0.2 does not change over generation.

It is important to discuss the effectiveness of DIBL-hopping when the DIBL-hopping is used for more complex circuits. In the complex circuits, the effectiveness of DIBL-hopping depends on the gate voltages of the series-connected MOSFETs. Table 5.3 shows the subthreshold leakage power dependence on the gate voltages in the series-connected MOSFET. The 0.1μm process PTM model is used for the SPICE model. The transistor width is set to 10μm. When two or more gates are turning off, $P(V_{DDH})$ has already suppressed sufficiently due to the stack effect [7]. This is why the effectiveness of DIBL-hopping is diminished in the series-connected MOSFET which two or more gates are turning off (that is, stacked MOS). The stack effect, however, does not affect the total power since the subthreshold leakage power with stacked MOS is about one digit smaller than that with other cases. Thus, the effectiveness of the DIBL-hopping does not change

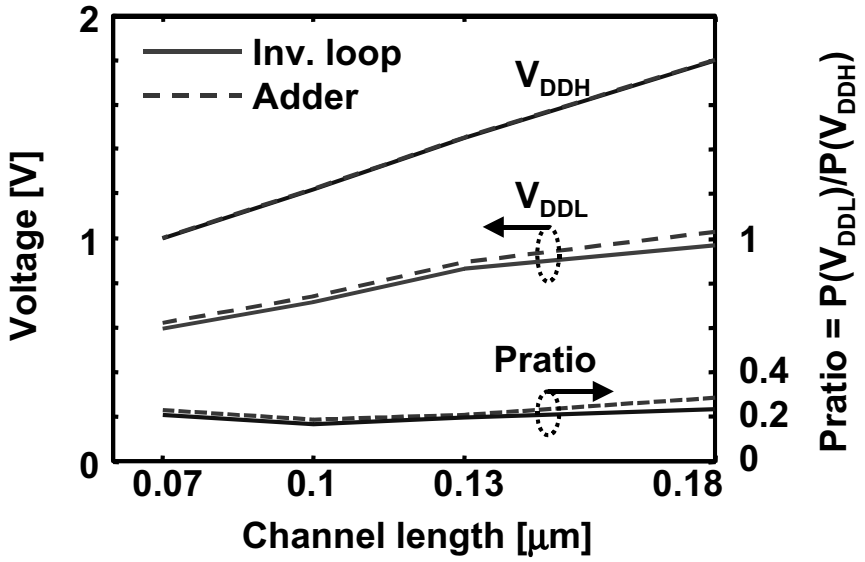


Fig. 5.19 Leakage power ratio comparison between inverter loop and Brent-Kung adder

drastically even in the complex circuits.

The effectiveness of DIBL-hopping with complex circuits is verified by using a 16-bit Brent-Kung adder [19]. V_{DDH} is the same as Table 5.2 and V_{DDL} is determined so that the delay of the critical path may double. Fig. 5.19 shows the leakage power ratio (Pratio) comparison between the 16-bit Brent-Kung adder and inverter loop by using the PTM models. The discrepancy of Pratio between the inverter loop and the Brent-Kung adder is within 5%. Hence, it can be said that the effectiveness of DIBL-hopping does not strongly depend on the complexity of the circuits.

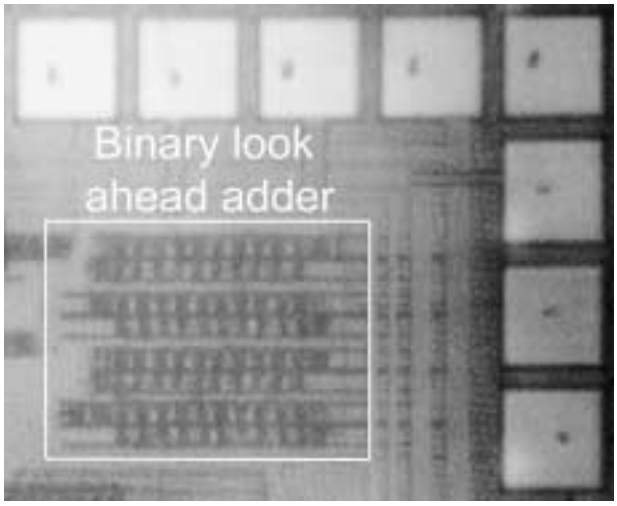


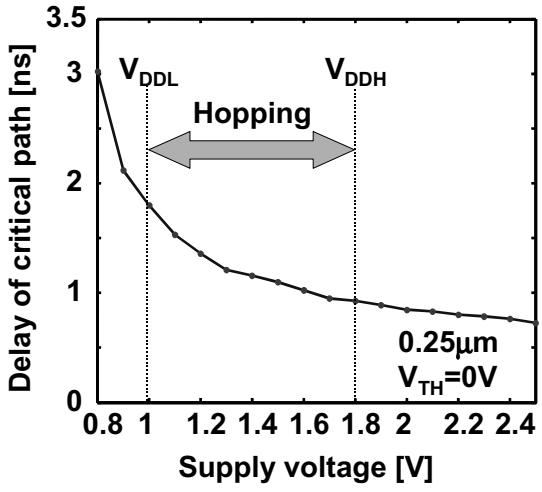
Fig. 5.20 Microphotograph of 16-bit Brent-Kung adder with DIBL-hopping

5.3.3 Measurement of Adder with DIBL-hopping

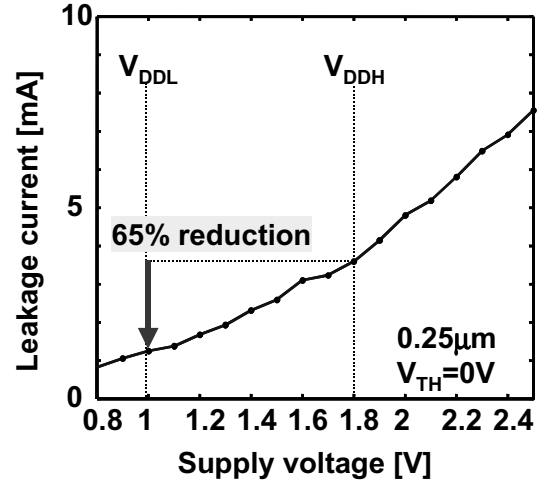
A 16-bit Brent-Kung adder with DIBL-hopping capability is fabricated in a 0.25μm PD-SOI (body contacted) technology. A microphotograph of the adder appears in Fig. 5.20. Low-V_{TH} process (V_{TH}=0V) is used to measure the leakage power.

Fig. 5.21(a) shows the measured result of delay dependence on supply voltage. If V_{DDH} is set to 1.8V (typical value in 0.25μm technology), V_{DDL} can be lowered to 1.0V. In V_{DD}-hopping [2], the leakage current reduction by using V_{DDL} is not taken into consideration. When V_{DDL} is asserted, 65% leakage current reduction can be achieved compared with the fixed V_{DDH} scheme due to the DIBL effect (see Fig. 5.21(b)). The power dependence on supply voltage is shown in Fig. 5.21(c). If V_{DDL} is applied, 80% power saving can be achieved compared with the fixed V_{DDH} scheme.

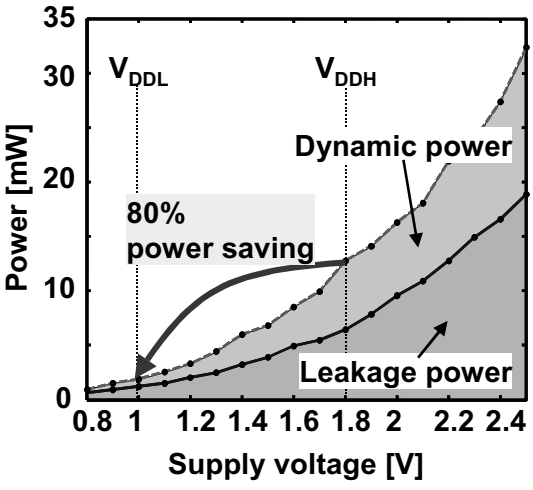
In order to verify the effectiveness of the DIBL-hopping with real-time applications, an MPEG4 encoding is simulated based on the measured data. The simulation result shows that 75% power saving can be achieved by using DIBL-hopping compared with the fixed V_{DDH} scheme.



(a) delay



(b) Leakage current



(c) Power

Fig. 5.21 Measured results of delay, leakage current and power of DIBL-hopping

5.3.4 Comparison Between DIBL-hopping and V_{TH}-hopping

DIBL-hopping scheme utilizes dynamic adjustment of frequency and V_{TH} through V_{DD} control depending on the workload of a processor. On the other hand, V_{TH}-hopping scheme (Section 5.2) is proposed to suppress the active leakage power. In V_{TH}-hopping, V_{TH} is controlled by the back-gate bias. If the V_{TH}-hopping is used, however, the dynamic power cannot be suppressed sufficiently since V_{DD} is constant. On the other hand, the effectiveness of DIBL-hopping scheme is limited by the DIBL-factor and ΔV_{DD}. Then, there is a limit in the effectiveness of reduction of the active leakage power since V_{TH} is not freely changed. In this section, effectiveness of the V_{TH}-hopping scheme and effectiveness of the DIBL-hopping scheme are compared.

Fig. 5.22 shows the power comparison between V_{TH}-hopping and DIBL-hopping. $P_{LEAK}(f_{CLK})/P_{TOTAL}(f_{CLK})$ is the ratio of the active leakage power to the total power when the maximum frequency, f_{CLK}, is assigned. $P_{TOTAL}(f_{CLK}/2)/P_{TOTAL}(f_{CLK})$ is the total power when the frequency is set to $f_{CLK}/2$ and normalized by $P_{TOTAL}(f_{CLK})$.

The result shows that $P_{TOTAL}(f_{CLK}/2)$ with DIBL-hopping is lower than that with V_{TH}-hopping if $P_{LEAK}(f_{CLK})/P_{TOTAL}(f_{CLK})$ is relative small. This means that if the dynamic power is relatively large, DIBL-hopping is more effective than V_{TH}-hopping since the dynamic power can be suppressed by lowering V_{DD}. On the other hand, if the leakage power is dominant, that is, if $P_{LEAK}(f_{CLK})/P_{TOTAL}(f_{CLK})$ is large, V_{TH}-hopping is more effective since V_{TH} with V_{TH}-hopping can be higher than V_{TH} with DIBL-hopping when $f_{CLK}/2$ is asserted. The reason is that the performance degradation by the decrease of V_{DD} is not occurred in V_{TH}-hopping.

In order to suppress the active leakage power and dynamic power further, the combined scheme between V_{TH}-hopping and DIBL-hopping is proposed. The combined scheme utilizes the adjustment of V_{DD} and V_{TH} depending on the frequency. Lower V_{DD} and

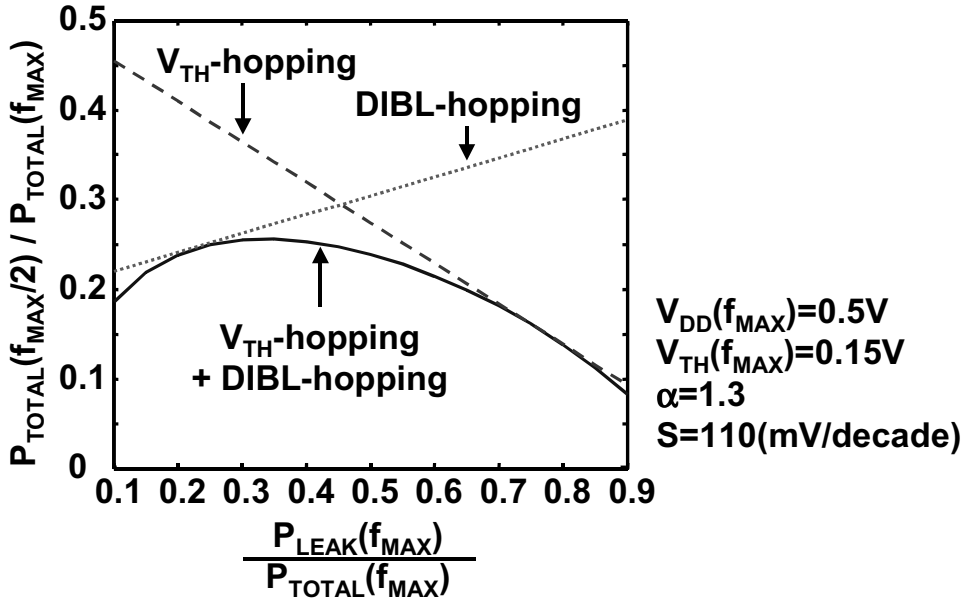


Fig. 5.22 Power comparison among V_{TH}-hopping, DIBL-hopping and combined scheme

higher V_{TH} are assigned simultaneously while the processor is run at $f_{CLK}/2$. The combined scheme is effective when the leakage power and dynamic power are comparable, that is, $P_{LEAK}(f_{CLK})/P_{TOTAL}(f_{CLK})$ is about 0.5. This tendency does not change even if the device and design parameters are changed.

5.4 Summary

In Section 5.2, a threshold voltage hopping (V_{TH}-hopping) scheme is proposed where the threshold voltage, V_{TH}, is dynamically controlled through software depending on a workload of a processor. The V_{TH}-hopping scheme can achieve 82% power saving compared with the fixed low-V_{TH} circuits in 0.5V supply voltage regime for multimedia applications. V_{TH}-hopping is effective in the low V_{DD} designs where V_{TH} is low and the active leakage component is dominant in total power consumption.

A small-scale RISC processor with V_{TH}-hopping and the positive back-gate biased scheme is fabricated. The measured data shows that if zero back-gate bias is applied, 91% power reduction was possible compared with the fixed 0.7V positive back-gate bias scheme. Based on the measured data, performance evaluation is conducted using MPEG-4 video coding. The simulation result shows that 86% power saving can be achieved by using V_{TH}-hopping compared with the fixed positive back-bias scheme.

In Section 5.3, DIBL-hopping scheme is proposed to suppress the active leakage power. The simulation results show that if the supply voltage is lowered so that the processor can run at $f_{CLK}/2$, 80% leakage power reduction is possible compared with the fixed supply voltage scheme. This result does not depend on the complexity of the circuits.

A 16-bit Brent-Kung adder with DIBL-hopping is fabricated. The measured data shows that if supply voltage is lowered to run at the half of the maximum frequency, 80% power reduction is possible compared with the fixed supply voltage scheme. In order to verify the effectiveness of DIBL-hopping, MPEG-4 encoding is simulated based on the measured data. The simulation result shows that 75% power saving can be achieved by using DIBL-hopping compared with the fixed V_{DD} scheme.

References

[1] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low-power design," *Proc. Int. Symposium on Low Power Electronics and Design*, pp.3-8, 1995.

[2] S. Lee and T. Sakurai, "Run-time voltage hopping for low-power real-time systems," *IEEE/ACM Proc. Design Automation Conference*, pp.806-809, 2000.

[3] T. Inukai, M. Takamiya, K. Nose, H. Kawaguchi, T. Hiramoto and T. Sakurai, "Boosted gate MOS (BGMOS): device/circuit cooperation scheme to achieve leakage-free giga-scale integration," *Proc. Custom Integrated Circuits Conference*, pp.409-412, 2000.

[4] S. Mutoh, et al, "1-V power supply high-speed digital circuits technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, pp.847-854, Aug., 1995.

[5] T. Kuroda, et al, "A 0.9-V 150-MHz, 10-mW, 4mm^2, 2-D discrete cosine transform core processor with variable threshold-voltage (VT) scheme," *IEEE J. Solid-State Circuits*, vol. 31, pp.1770-1778, Nov. 1996.

[6] Q. Wang and S. Vrudhula, "Static power optimization of deep submicron CMOS circuits for dual vt technology," in *International Conference on Computer-Aided Design*, pp.490-494, 1998.

[7] Y. Ye, S. Borkar and V. De, "A technique for standby leakage reduction in high-performance circuits," *Dig. Tech. Papers Symp. of VLSI Circuits*, pp.40-41, 1998.

[8] R. R. Troutman, "VLSI limitations from drain induced barrier lowering," *IEEE Trans. Electron Devices*, vol. ED-26, no.4, pp.461, Apr., 1979.

[9] A. Chandrakasan, V. Gutnik and T. Xanthopoulos, "Data driven signal processing: an approach for energy efficient computing," *Proc. Int. Symposium on Low Power Electronics and Design*, pp.347-352, 1996.

[10] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Roy and V. De, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in CMOS IC's," *Proc. Int. Symposium on Low Power Electronics and Design*, pp. 252-254, 1999.

[11] T. Miyake, et al, "Design methodology of high performance microprocessor using ultra-low threshold voltage CMOS," *Proc. Custom Integrated Circuits Conference*, pp.275-278, 2001.

[12] C. Wann et al, "CMOS with active well bias for low-power and RF/analog applications," *Dig. Tech. Papers Symp. of VLSI Tech*., pp.158-159, 2000.

[13] Apollo user guide, Avant! co., 1998.

[14] Diva interactive verification reference manual, Cadence Design Systems Inc., 1997.

[15] S. –H. Lo et al., "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin oxide nMOSFET's," *IEEE Electron Device Letter*, vol.18, pp.209-211, 1997.

[16] Y. Cao, T. Sato, M. Orshansky, D. Sylvester and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuits simulation," *Proc. Custom Integrated Circuits Conference*, pp.201-204, 2000.

[17] International Technology Roadmap for Semiconductors, 1999.

[18] D. Foty, *MOSFET Modeling with SPICE*, Prentice Hall, Inc., 1997.

[19] R. P. Brent and H. T. Kung, "A regular layout for parallel adders," *IEEE Trans. on Computers*, vol.C-31, no.3, pp.260-264, Mar., 1982.

Chapter 6. Conclusion

In this thesis, low power and high-performance design methodologies for low voltage VLSI's are proposed. In this chapter, the results of the thesis are summarized.

In Chapter 3, analytical formulas of the CMOS power dissipation are proposed. Using the formulas, future trend of the optimum CMOS design is discussed.

(1) In Section 3.2, a simple and closed-form formula for the short-circuit power dissipation is derived. As a results, the importance of the short-circuit power will not be increased (about 10%).

(2) In Section 3.3, appropriate effective gate capacitance, $C_{G,eff}$, has been defined and a method is proposed to extract the value by using SPICE. The discrepancy between $C_{G,eff}$ and oxide capacitance, C_{OX}, is increasing in low-voltage regime and adopting $C_{G,eff}$ in accurate power and delay estimation becomes more important in the future.

(3) Closed-form formulas for optimum V_{DD} and V_{TH} which take the variation of V_{TH}

and temperature into consideration are presented for low power and high-speed LSI's in Section 3.4. From the calculation using these formulas, it is shown that a simple guideline for power optimization is to set the radio of the maximum leakage power to the total power around 30%. Extending the analysis, the future trend in the optimum threshold voltage (V_{THopt}) and the optimum supply voltage (V_{DDopt}) is calculated using the device parameters given in the ITRS. The V_{DDopt} coincides with the ITRS and V_{THopt}. The lowest V_{TH}, $V_{TH,min}$, is in the range of 0V~0.1V and the target V_{TH} is almost constant at 0.2V over generations. The proposed scenario shows that more number of MOSFET is consumed in the memory blocks than the logic blocks in the future.

In Chapter 4, new buffer insertion techniques for high-performance bus interconnects are discussed.

(4) In Section 4.2, closed-form formulas for optimum buffer insertion where the junction capacitance effect is taken into account are proposed and an appropriate choice of the effective constant resistance is investigated. Using these formulas, the optimum interconnect delay and power comparison among bulk, PD-SOI, FD-SOI, and the double-gate structure is discussed. MOSFET with small junction capacitance, like SOI, can suppress both the optimum delay and power by 15% compared with the conventional bulk MOSFET whose junction capacitance is assumed to be equal to the gate capacitance.

(5) In Section 4.3, a new buffer insertion scheme for bi-directional buses, namely dual-rail bus (DRB) scheme, which does not have noise problems, and a high-speed buffer insertion scheme for uni-directional buses, namely staggered firing bus (SFB) scheme, are proposed and measured. In 2008 when 0.07μm design rule is used, DRB scheme can improve the performance of bi-directional buses by an order of magnitude and SFB scheme can suppress the delay of uni-directional buses by about 20% at 0.18μm generation and beyond. If we use

SFB scheme instead of conventional uni-directional buses, the 27% power reduction can be achieved while the performance of SFB is the same as that of conventional buses.

In order to suppress the active leakage power which becomes an crucial issue in low-voltage processors, new hardware-software cooperative schemes are implemented in Chapter 5.

(1) A threshold voltage hopping (V_{TH}-hopping) scheme is proposed where the threshold voltage, V_{TH}, is dynamically controlled through software depending on a workload of a processor. The V_{TH}-hopping scheme can achieve 82% power saving compared with the fixed low-V_{TH} circuits in 0.5V supply voltage regime for multimedia applications. A small-scale RISC processor with V_{TH}-hopping and the positive back-gate biased scheme is fabricated. Based on the measured data, performance evaluation is conducted using MPEG-4 video coding. The simulation results shows that 86% power saving can be achieved by using V_{TH}-hopping compared with the fixed positive back-bias scheme.

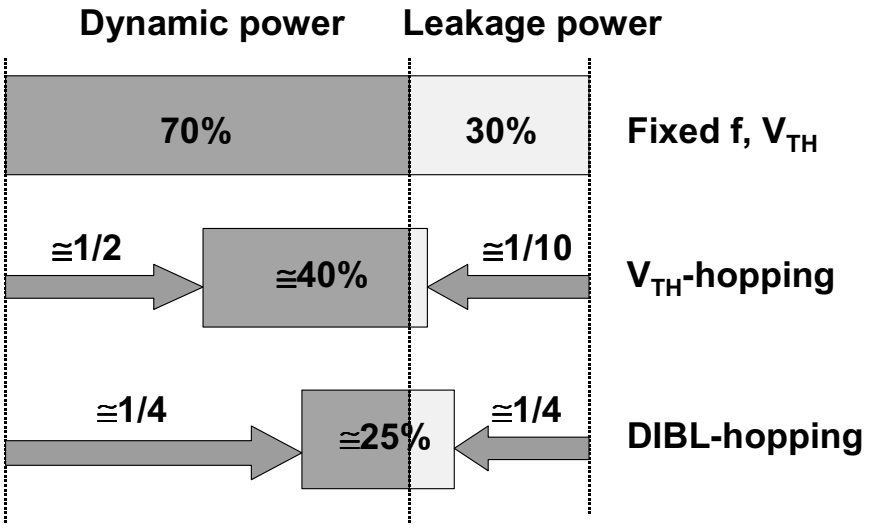
(2) As another method which can suppress the active power consumption dynamically, DIBL-hopping scheme is proposed. The subthreshold leakage current can be suppressed by lowering V_{DD} since V_{TH} increases by the DIBL (drain induced barrier lowering) effect. An MPEG-4 encoding is simulated based on the measured data. The result shows that the total power can be suppressed by 75% compared with the fixed V_{DD} scheme.

In the combination of the proposed techniques, the following power reduction effects are expected.

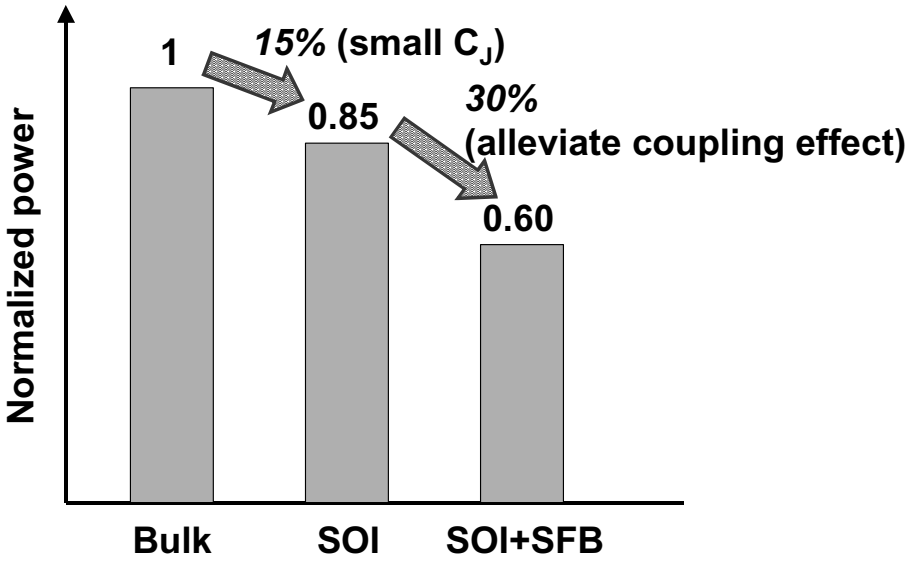
As for logic blocks, the leakage power can be suppressed by using V_{TH}-hopping and DIBL-hopping. The effectiveness of these schemes is shown in Fig. 6.1(a). We assumed that the ratio of the leakage power to the total power is 0.3 (the optimum ratio

discussed in Section 3.4) when the fixed frequency and voltages scheme is used. The effectiveness of V_{TH}-hopping and DIBL-hopping are estimated from the results in Chapter 5. From the figure, the power in the logic block can be suppressed by 60% when the V_{TH}-hopping is used and 75% when the DIBL-hopping is adapted. (As is mentioned in Section 5.3.4, which scheme is more effective depends on the ratio of the leakage power to the total power. If the leakage power component is dominant, V_{TH}-hopping is more effective.)

As for the global buses, 15% power reduction can be achieved by using SOI technology instead of conventional bulk technology. Furthermore, the staggered firing bus (SFB) scheme can decrease the power by 30% without dropping the performance. Then, the power of the bus wires is suppressed by 40% by using SOI and SFB scheme (see Fig. 6.1(b)).



(a) Logic block



(b) Bus interconnects

Fig. 6.1 Summarize of proposed power reduction techniques

List of Publications and Presentations

Publications

[1] K. Nose and T. Sakurai, "Analysis and future trend of short-circuit power," *IEEE Transactions on Computer-Aided Design of Integrated Circuits And Systems*, vol. 19, pp.1023-1030, Sep., 2000.

[2] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee and T. Sakurai, "V_{TH}-hopping scheme to reduce subthreshold leakage for low-power processors," *IEEE Journal of Solid-State Circuits*, vol. 37, pp.413-419, Mar., 2002.

[3] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for low-power and high-speed applications and which implication to low-voltage 1G-scale integration," *IEEE Transaction on Very Large Scale Integration Systems* (submitted).

Presentations

[4] K. Nose and T. Sakurai, "Closed-form expressions for short-circuit power of short-channel CMOS gates and its scaling characteristics," *Proceedings of International Technical Conference on Circuit/Systems, Computers and Communication*, pp.1741-1744, July, 1998.

[5] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for low-power and high-speed applications," *Proceedings of the Asia and South Pacific Design Automation Conference*, pp.469-474, Jan., 2000.

[6] K. Nose, S. -I. Chae and T. Sakurai, "Voltage dependent gate capacitance and its impact in estimating power and delay of CMOS digital circuits with low supply voltage," *Proceeding of International Symposium on Low Power Electronics and Design*, pp.228-230, July, 2000.

[7] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee and T. Sakurai, "V_{TH}-hopping scheme for 82% power saving in low-voltage processors," *Proceedings of Custom Integrated Circuits Conference*, pp.93-96, May, 2001.

[8] 野瀬浩一, 平林雅之, 川口博, 李誠洙, 桜井貴康,「閾値ホッピング(V_{TH}-hopping)手法を用いた低電圧・低消費電力プロセッサ」電子情報通信学会技術研究報告, vol. 101, no. 85, ICD2001-33, pp.67-73, 2001.

[9] K. Nose and T. Sakurai, "Two schemes to reduce interconnect delay in bi-directional and uni-directional buses," *Symposium on VLSI Circuits Digest of Technical Papers*, pp.193-194, June, 2001.

Other publications

[10] 野瀬浩一, 桜井貴康,「マイクロ IDDQ テストのための電流測定デバイス」,電子情報通信学会論文誌, Vol.J83-C, No.6, pp.516-522, June, 2000.

[11] K. Nose and T. Sakurai, "Current sensing device for micro-IDDQ test," *Electronics and Communications in Japan*, part 2, vol.84, no.9, 2001.

Other presentations

[12] K. Nose and T. Sakurai, "Integrated current sensing device for micro IDDQ test," *Proceedings of the seventh Asian Test Symposium*, pp.323-326, Dec., 1998.

[13] K. Nose and T. Sakurai, "Micro IDDQ test using lorentz force MOSFET's," *Symposium on VLSI Circuits Digest of Technical Papers*, pp.169-170, June, 1999.

[14] H. Kawaguchi, K. Nose and T. Sakurai, "A CMOS scheme for 0.5V supply voltage with pico-ampere standby current," *International Solid-State Circuits Conference Dig. of Tech. papers*, pp.192-193, Feb. 1998.

[15] H. Kawaguchi, K. Nose and T. Sakurai, “A CMOS scheme for 0.5V supply voltage with pico-ampere standby current,” *International Workshop on Advanced LSIs*, vol. ICD-98-82, pp.45-49, July 1998.

[16] グェン・ドュック・ミン, 野瀬 浩一, 桜井 貴康,「低電源電圧 depletion 型 CMOS の最低動作閾値電圧」, 1999 年電子情報通信学会総合大会.

[17] K. Kanda, K. Nose, H. Kawaguchi and T. Sakurai, “Design impact of positive temperature dependence of drain current in sub 1V CMOS VLSI's,” *Proceedings of Custom Integrated Circuits Conference*, pp.563-566, May, 1999.

[18] T. Inukai, M. Takamiya, K. Nose, H. Kawaguchi, T. Hiramoto and T. Sakurai, “Boosted gate MOS (BGMOS): device/circuit cooperation scheme to achieve leakage-free giga-scale integration,” *Proceedings of Custom Integrated Circuits Conference*, pp.409-412, May, 2000.

[19] M. Hirabayashi, K. Nose and T. Sakurai, “Design methodology and optimization strategies for dual Vth schemes using commercially available tools,” *Proceeding of International Symposium on Low Power Electronics and Design*, pp.283-286, Aug., 2001.

Patent

[20] 桜井》 貴康, 川口》 博, 野瀬》 浩一, 「電力制御装置及び方法並びに電力制御プログラムを記録した記憶媒体」, 特願 2000-221676, 平成 12 年 7 月 24 日出願, 日本