High-Speed Circuit Design with Scaled-Down MOSFET's and Low Supply Voltage

Takayasu Sakurai

Semiconductor Device Eng. Lab., Toshiba Corporation
1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210, Japan
Phone : +81-44-549-2710 FAX : +81-44-549-2712

1. Introduction

The gate length of MOSFET's is getting shorter for high performance and high density. A power supply voltage must be lowered to guarantee sufficient reliability for the short-channel transistors. It is predicted that in the year 2000, the gate length and VDD will become less than 0.15μm and 1.5V respectively.

The purpose of this paper is to investigate some of the points where circuit optimization with short-channel and low V_{DD} is different from the circuit optimization with long-channel and high V_{DD}. Two effects are mainly considered in this paper. One effect is that in the short-channel MOSFET, the drain current dependence on gate voltage deviates from the Shockley's quadratic law and approaches linear law due to severe carrier velocity saturation, as shown in Fig.1. The other effect is that a threshold voltage of a MOSFET (V_{TH}) can not be scaled linearly with the size reduction. This prevents currently known high-speed circuits from being just scaled to give high-speed solution in the low voltage era.

In Section 2 and 3, the influences by the two effects are studied in a sense amplifier design and a SRAM cell design, respectively. Memory designs are investigated because a high-speed on-chip memory such as a cache memory is one of the most essential components to realize high-speed VLSI systems. This is because memories are often in the critical paths of VLSI systems and I/O delay is enormous compared with intra-chip delay.

First a design strategy of a widely-used current-mirror sense amplifier (CMSA) for an embedded SRAM based on analytical formulas is given. It is shown that the voltage gain decreases due to the carrier velocity saturation. In the low V_{DD} regime, the CMSA suffers from a speed degradation and a current latch sense amplifier (CLSA) is shown to operate faster. As for the SRAM cell design, an analytical expression is derived for a static noise margin (SNM) and it is shown that the SNM decreases by the velocity saturation.

In Section 4, an influence of the linear-law effect to an optimization strategy of basic logic gates is described. It is shown that the maximum number of a logic gate input that is allowed in high-speed designs increases to about 7 from 4 which was the maximum number in the designs with long-channel MOSFET's. Lastly, in Section 5, the effect of the V_{TH} non-scalability on a gate speed is discussed. Use of a substrate bias scheme is effective in increasing speed.

2. Sense Amplifier Design

The MOSFET model used in this paper is the following n-th power MOS mode. The salient feature of the model is the introduction of a parameter n that accounts for the velocity saturation in the short-channel devices and decreases from 2 to 1 as velocity saturation gets severer [1,2].

$$V_{DSAT}=K(V_{GS}-V_{TH})^{m},\ I_{DSAT}=\beta/2(V_{GS}-V_{TH})^{n} \quad (1)$$

$$I_D=I_{D5}=I_{DSAT}(1+\lambda V_{DS}) \quad (V_{DS}\geq V_{DSAT}: \text{saturated region}) \quad (2)$$

$$I_D=I_{D3}=I_{D5}\left(2-\frac{V_{DS}}{V_{DSAT}}\right)\frac{V_{DS}}{V_{DSAT}} \quad (V_{DS}\geq V_{DSAT}: \text{linear region}) \quad (3)$$

Current-mirror sense amplifier (CMSA) as shown in Fig.2a has long been adopted for high-speed SRAM's but the design optimization theory with short-channel devices has not been necessarily well established. First, the role of the so-called current source Q_5 is clarified. The operation of the CMSA is not affected by changing Q_5 to linear operation or even to fixed voltage source at V_S as shown in Fig.3. Therefore it can be said that the role Q_5 is pull up the V_S. Then, by setting V_S constant and equating the drain current of Q_1 and Q_3, the following equation is derived. Subscript 0 denotes the state where $\overline{V_I}$ and V_I are the same and Δ signifies the small difference from that state.

$$\frac{1}{2}\beta_N(V_{I0}+\Delta V_I-V_S-V_{TN})^{n_N}\{1+\lambda_N(V_{O0}-\Delta V_O-V_S)\}$$
$$=\frac{1}{2}\beta_P(V_{DD}-V_{O0}-V_{TP})^{n_P}\{1+\lambda_P(V_{DD}-V_{O0}+\Delta V_O)\} \quad (4)$$

As for V_{I0} and V_{O0}, the following equation holds by equating the drain current of Q_2 and Q_4.

$$\frac{1}{2}\beta_N(V_{I0}-V_S-V_{TN})^{n_N}\{1+\lambda_N(V_{O0}-V_S)\}$$
$$=\frac{1}{2}\beta_P(V_{DD}-V_{O0}-V_{TP})^{n_P}\{1+\lambda_P(V_{DD}-V_{O0})\} \quad (5)$$

By dividing (4) by (5) side by side, and using the relation $(1+x)^n \approx 1+x^n$ ($x \ll 1$), a formula for the voltage gain is obtained.

$$\frac{\Delta V_O}{\Delta V_I}=\frac{n_N}{V_{I0}-V_S-V_{TN}}\Big/\left\{\frac{\lambda_N}{1+\lambda_N(V_{O0}-V_S)}+\frac{\lambda_P}{1+\lambda_P(V_{DD}-V_{O0})}\right\}$$

$$\therefore \text{Voltage gain}=\frac{\Delta V_O}{\Delta V_I}\approx\frac{n_N}{V_{I0}-V_S-V_{TN}}\cdot\frac{1}{\lambda_N+\lambda_P} \quad (6)$$

This expression claims that the voltage gain is increased by decreasing V_{I0} and by increasing V_S and V_{TN}. This effect is assured in Fig.4. The eq.(6) also suggests that the gain is independent of the PMOS size, which is certified in Fig.5. It is seen from the formula that the gain decreases as n goes from 2 to 1.

If V_S is too high, the output voltage swing is limited since V_O can not go low enough. In this sense, V_S is determined by the output voltage swing needed. Then, the size of Q_5 can be determined to achieve the V_S value with the current constraint which is determined by a power requirement. Make flow as large current as possible for higher speed operation within the power constraint.

The input voltage V_I should be set as low as possible for the higher gain, but since V_I is usually a bit line voltage, it can not be lowered to ensure sufficient write margin. In most cases, V_I is set equal to $V_{DD}-V_{TN}$' because of the NMOS bit line load. Then the size of Q_1 can be determined. The size of PMOS Q_3 is then determined by the requirement that the initial output voltage V_{O0} is preferably about the center of the output swing. β_P is calculated using the following formula.

$$V_{O0}\approx V_{DD}-V_{TP}-(\beta_N/\beta_P)^{1/n}\cdot(V_{I0}-V_S-V_{TN}),\ n=(n_N+n_P)/2$$

It should be noted that a little longer channel length than the minimum channel length is better be used for Q_1 and Q_2 because it diminishes the process fluctuation and moreover increases n and hence increases the voltage gain.

Although the CMSA are widely used, it suffers from a large delay to output 'High' at a low V_{DD}. This is because Q_4 can not be sufficiently biased in low V_{DD} environments. Other frequently used sense amplifier is a voltage latch sense amplifier (VLSA) which is used in almost all DRAM's. The essential part of the VLSA is a cross-coupled inverters so that the amplifier operates under low supply voltage like 1V. However, the input and output of this amplifier are

0-7803-1254-6/93$03.00 © 1993 IEEE

essentially common and it requires a complicated timing control to separate them. This control adds extra delay, which makes the amplifier not suitable for high-speed applications.

Combining I/O separate feature of the CMSA and the low voltage feature of the VLSA, a current latch sense amplifier (CLSA) as shown in Fig.2c is suitable for on-chip SRAM application at low supply voltage. The CLSA consists of a cross-coupled inverters and the input is driven from the source of the inverter MOSFET through an added transistor. This is different from the VLSA where the input is driven from the drain of the inverter MOSFET. The CLSA shows 1.5ns delay with 0.5μm MOSFET's at 1V V_{DD}, while the CMSA shows 4ns as shown in Fig.6.

Although the CLSA needs a latch timing control, the delay difference surmounts the drawback and moreover even the CMSA needs output precharge timing control for high-speed applications. The CLSA can be a candidate for high-speed SRAM amplifier in the low V_{DD} environments. The CLSA has also a low-power feature because once the data is latched there is no direct current path from V_{DD} to V_{SS}.

3. SRAM Cell Design

As for a memory cell, a 4T SRAM cell with highly resistive poly-silicon loads suffers from a stability problem and a soft-error problem when $V_{DD} \le 2V$ [3]. Consequently, 6T SRAM cell is the only candidate for the on-chip SRAM cell in the low V_{DD}.

In order to estimate the static noise margin (SNM) [3] of the memory cell, some simplifications are made to the basic MOS model. λ is set equal to zero because the channel-length modulation does not give essential effects in SNM analysis and m is set equal to n/2, which is exact in long-channel devices and is empirically true even in the miniaturized devices.

The definition of the SNM is graphically shown as SNM_{DEF} in Fig.7. The SNM_{DEF} can be approximated by SNM_{APP} ($=\sqrt{2}(V_{OH}-V_{OL})$) where V_{OH} is the turn-on input voltage of the Q_5-Q_1 inverter. This is because essentially the SNM is the voltage margin to prevent the 'Low' voltage, V_{OL}, of the cell node from turning on the inverter on the other side.

The line L1 is determined by Q_4-Q_2 inverter and the line L2 by Q_5-Q_1 inverter. To obtain V_{OL}, the saturated drain current of the access transistor $I_{D,a}$ is set equal to the linear drain current of the driver transistor $I_{D,d3}$ as follows. The quadratic term in V_{DS} can be neglected in $I_{D,d3}$ because $V_{OL} \ll V_{DD}-V_{TN}$.

$$I_{D,a}=\frac{1}{2}\beta_a(V_{DD}-V_{OL}-V_{TN})^{n_N} \approx \frac{1}{2}\beta_a(V_{DD}-V_{TN})^{n_N}\left(1-\frac{V_{OL}}{V_{DD}-V_{TN}}\right)$$

$$I_{D,d3} \approx \frac{\beta_d}{K_N}(V_{DD}-V_{TN})^{n/2}V_{OL}$$

By introducing $V_Z=V_{DD}-V_{TN}$ and equating the two current terms,

$$I_{D,a}=\frac{1}{2}\beta_a(V_Z-V_{OL})^{n_N} \approx \frac{1}{2}\beta_a V_Z^{n_N}\left(1-\frac{V_{OL}}{V_Z}\right) = I_{D,d3} \approx \frac{\beta_d}{K_N}V_Z^{n/2}V_{OL}.$$

Then, V_{OL} is solved as follows.

$$V_{OL} \approx \frac{V_Z}{n_N+2rV_Z^{1-n/2}/K_N} \approx \frac{V_Z}{n_N+2rV_Z/V_{D0}}$$

On the other hand, the V_{OH} can be obtained by equating the saturated drain current of the driver transistor $I_{D,d5}$ and the linear region current of the PMOS $I_{D,p}$ as follows. The V_∂ in Fig.7 is a fitting parameter and $V_\partial=0.1V_{DD}$ turns out to give a good fit.

$$I_{D,d5}=\frac{1}{2}\beta_5(V_{OH}-V_{TN})^{n_N} = I_{D,p} \approx \frac{\beta_a}{K_P}V_\partial(V_{DD}-V_{TP}-V_{OH})^{n_P/2},$$

Assuming $n_N \approx n_P \approx (n_N+n_P)/2 \equiv n$, V_{OH} can be solved as follows.

$$V_{OH} \approx V_{TN} - 0.5V_\Delta + \sqrt{V_\Delta(V_{DD}-V_{TN}-V_{TP})},\quad V_\Delta=(2qV_\partial/K_P r)^{n/2}$$

SNM_{APP} can obtained by calculating $\sqrt{2}(V_{OH}-V_{OL})$ as

$$SNM_{APP} \approx \sqrt{2}\left\{\frac{V_Z}{n_N+2rV_Z/V_{D0}} - V_{TN} - 0.5V_\Delta + \sqrt{V_\Delta(V_{DD}-V_{TN}-V_{TP})}\right\}$$

The comparisons between the SNM_{APP} and the simulated SNM_{DEF} are made in Figs.8 and 9 for various configurations and good agreement is observed. Figure 10 is the calculated SNM_{APP} for various n. It is seen that with decreasing n, the SNM decreases. This is mainly because the noise margin of an inverter made with short-channel devices is smaller than that with long-channel devices as shown in Fig.11.

4. Basic Logic Gate

It has been qualitatively discussed that the speed degradation of N serially connected MOSFETs of size W is less than 1/N compared with an inverter speed where only one MOSFET of size W drives the output.

In this paper, a quantitative simulation is carried out to know what N is a cross-over point, over which a two-stage configuration should be used instead of a single-stage N-input logic gate to optimize speed. A gate array implementation is considered where the gate width of P-channel MOSFET is the same as that of N-channel MOSFET and the load fanout is assumed to be 7 which is typical. An input slope is chosen as an output slope of a 2-input NAND gate with fanout of 7. In Fig.12, delay of N-input NAND gate is compared with that of NOR-NAND two-stage configuration of the same function with an input phase inverted.

The delay cross-over point with 2μm MOSFET's (n ~ 1) was observed between 4-input and 5-input, while it was between 7 and 8 with 0.5μm MOSFET's where n is about 1.2. This corresponds to a design practice of old days that more than 5-input gate should be avoided. This design rule of thumb should be changed to "avoid more than 8-input logic gate" in lower sub-micron designs.

5. Mitigating Non-Scalability of Threshold Voltage

Threshold voltage is not a scalable parameter. This fact may casts the most stringent constraints on the low voltage high-speed circuits. Sub-threshold current I_{SUB} is expressed as

$$I_{SUB} \propto 10^{\left\{\frac{V_{GS}-V_{TH}}{s}\right\}},\qquad s=\frac{kT}{q}\left\{1+\frac{C_{DEP}}{C_{OX}}\right\}\ln 10$$

s and called an s factor. s is about 110mV / decade and can not be scaled. The effect of V_{TH} on propagation delay time is estimated for various V_{DD} using a simple delay expression as follows [4].

$$tpd=\frac{C_L V_{DD}}{(V_{DD}-V_{TH})^n}\left[\left(\frac{1}{2}-\frac{1-V_{TH}/V_{DD}}{1+n}\right)\left(\frac{0.9}{0.8}+\frac{V_{D0}}{0.8V_{DD}}\ln\frac{10V_{D0}}{eV_{DD}}\right)+\frac{1}{2}\right]$$

The results are shown in Fig.13. It is seen that with a smaller n the delay dependence on V_{TH} decreases but still in 1V supply voltage, 0.1V of V_{TH} change amounts up to 50% delay change.

A leakage current of a logic gate is proportional to $\exp(-V_{TH}/s)$. V_{TH} should be as low as possible for the higher speed but the minimum V_{TH} is determined by a leakage current constraint. The s factor can be reduced by applying substrate bias. The measured substrate current dependence on V_{GS} is shown in Fig.14. If a substrate bias of -1V is applied, the s factor decreases from 110mV/decade to 91mv/decade. With keeping the leakage current constant, the decrease in s factor of this magnitude can achieve the decrease of V_{TH} by 0.1V. This is beneficial to high-speed design with low V_{DD}. The substrate bias scheme is also preferable in the following aspects to realize high-speed VLSI's.

- Smaller body effect
- Lower junction capacitance

References

[1] T.Sakurai and A.R.Newton, "Delay Analysis of Series-Connected MOSFET Circuits," IEEE J. of Solid-State Circ., Vol.26, No.2, pp.122-131, Feb.1991.

[2] T.Sakurai and A.R.Newton, "A Simple MOSFET Model for Circuit Analysis," IEEE Trans. on ED, ED-38, No.4, pp.887-894, Apr.1991.

[3] E. Seevinck, F.J.List, and J.Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," IEEE J. of Solid-State Circ., Vol.22, No.5, pp.748-754, Oct.1987.

[4] T.Sakurai and A.R.Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," IEEE J. of Solid-State Circ., Vol.25, No.2, pp.584-594, Apr.1990.

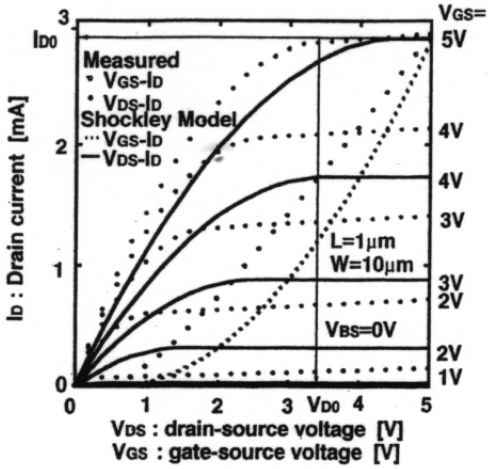


Fig.1 I_D-V_{DS} and I_D-V_{GS} characteristics of short-channel NMOS

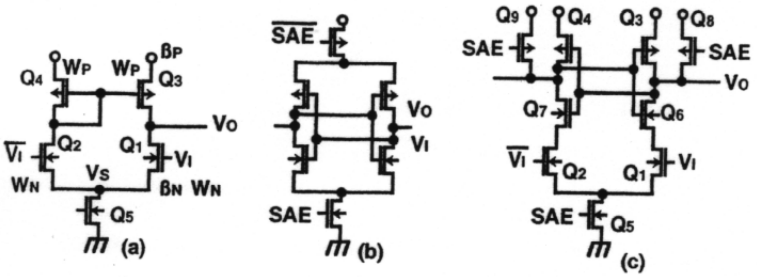


Fig.2 Various sense amplifiers. (a) Current-Mirror S/A (CMSA) (b) Voltage Latch S/A (VLSA) and (c) Current Latch S/A (CLSA)

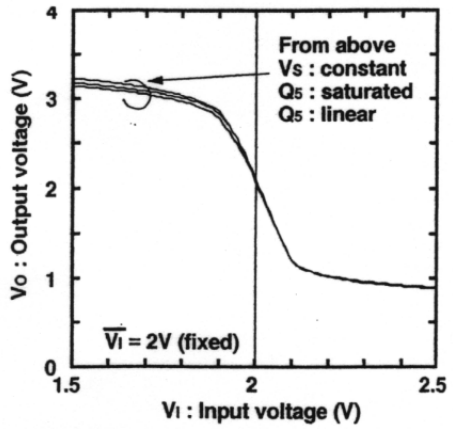


Fig.3 CMSA behavior with various common source implementation

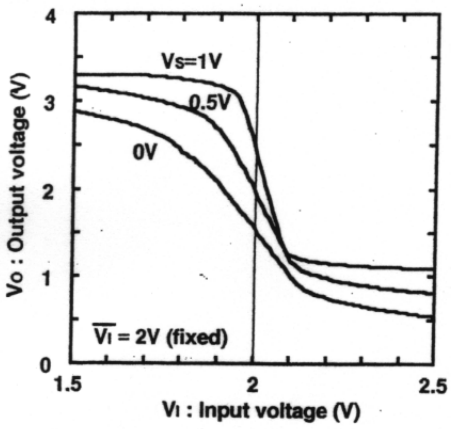


Fig.4 CMSA bahavior change for various V_S voltages

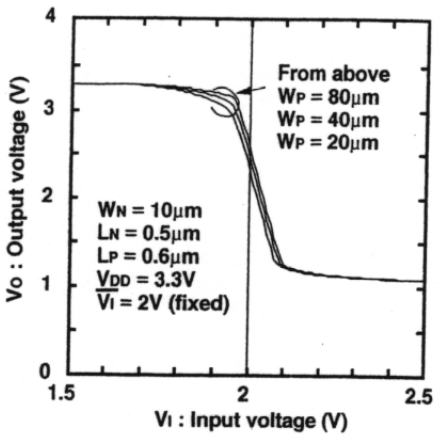


Fig.5 CMSA behavior change for various PMOS size

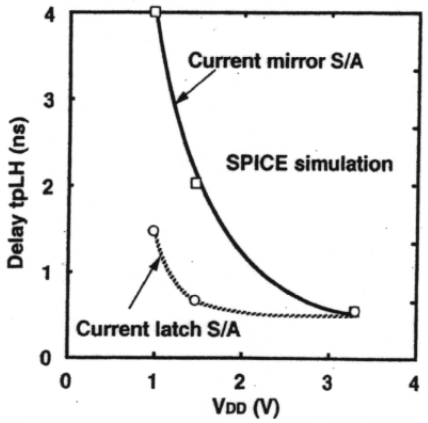
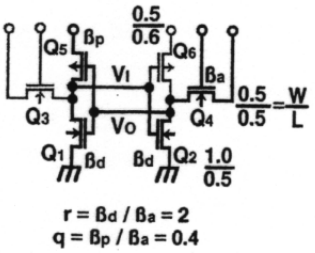


Fig.6 CMSA and CLSA delay with low supply voltage



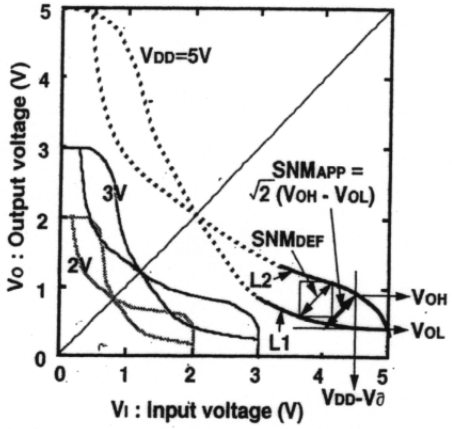


Fig.7 Static noise margin (SNM) of full CMOS SRAM cell

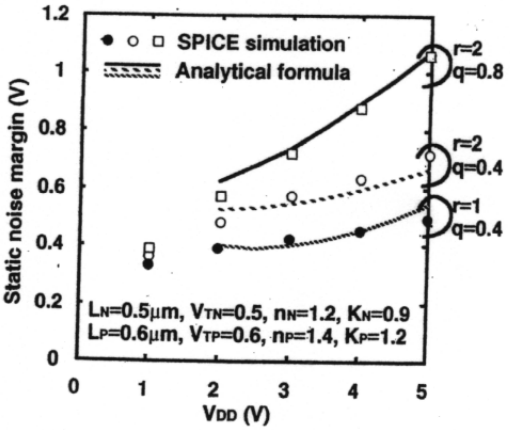


Fig.8 V_{DD} dependence of static noise margin

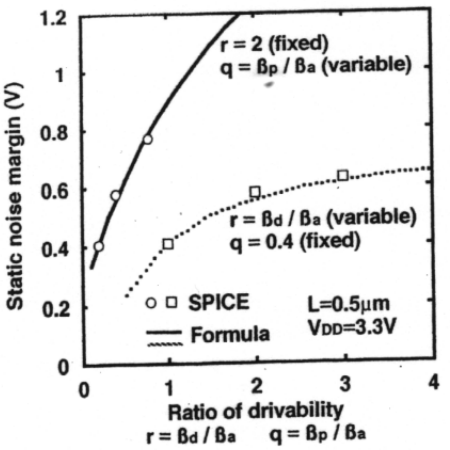


Fig.9 Ratio dependence of static noise margin

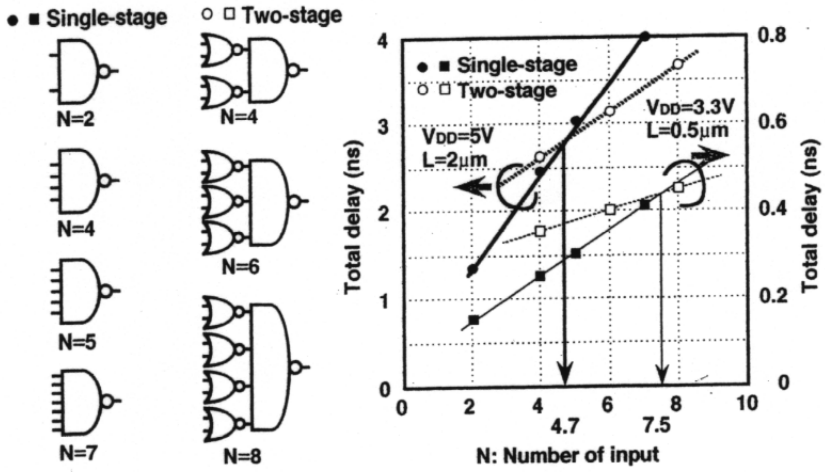


Fig.12 Delay dependence on number of input of logic gate

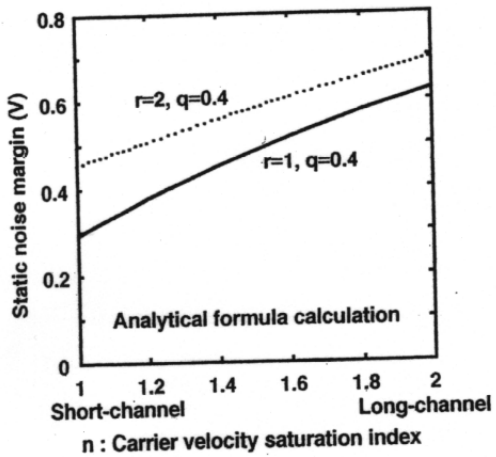


Fig.10 Dependence of static noise margin on velocity saturation index

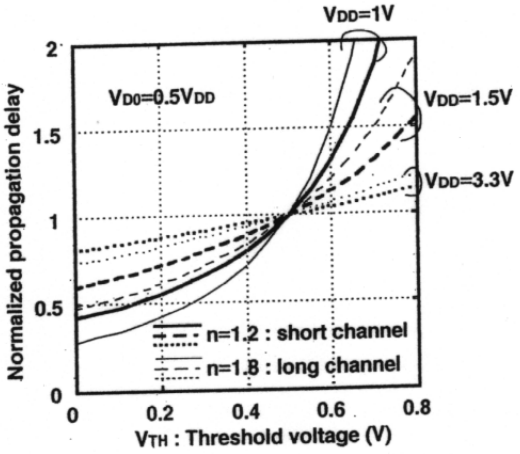


Fig.13 Delay dependence on threshold voltage

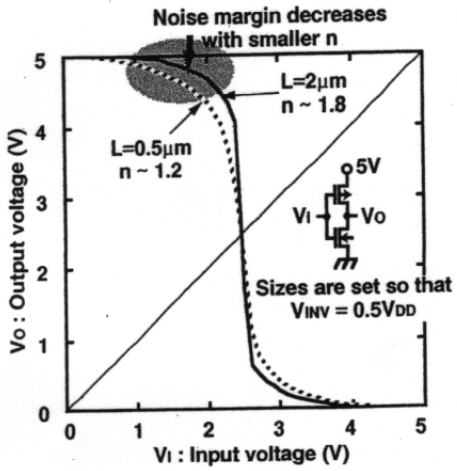


Fig.11 Noise margin of inverter with long and short channel devices

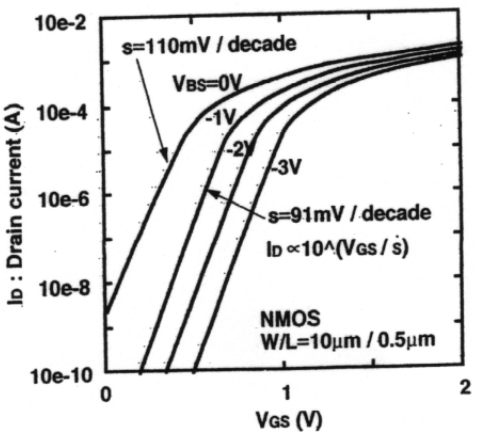


Fig.14 Measured sub-threshold current vs. V_{GS}