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 VDD is different from the circuit optimization with long-channel and high VDD. 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 saturation, as shown in Fig.1. The other effect is that a threshold voltage of a MOSFET (VTH) 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 increases due to the carrier velocity saturation. In the low VDD 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 VTH 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].

\[ V_{DSAT} = K(V_{GS} - V_{TH})^n, I_{DSAT} = 8/2(V_{GS} - V_{TH})^n \]  

(1) \[ I_P = I_{DSAT}(1+V_{DSAT}/V_{TH}^{Sat}) \]  

(2) \[ I_P = I_{DSAT}(2 - V_{DSAT}/V_{TH}^{Sat}) \]  

(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 clarified. First, the role of the so-called current source Qs is clarified. The operation of the CMSA is not affected by changing Qs to linear operation or even to fixed voltage source at V5 as shown in Fig.3. Therefore it can be said that the role Qs is pull up the V5. Then, by setting V5 constant and equating the drain current of Q3 and Q4, the following equation is derived. Subscript 0 denotes the state where V1 and V5 are the same and A signifies the small difference from that state.

\[ I_{DSAT} = K(V_{DD} + \Delta V_1 - V_S - V_{TH})^{n(1+\Delta V_0/V_{DD}-V_0)} \]

As for V5 and VVOG, the following equation holds by equating the drain current of Q3 and Q4.

\[ \frac{\Delta V_0}{\Delta V_1} = \frac{I_{DSAT}(V_{DD} - V_S - V_{TH})^{n(1+\Delta V_0/V_{DD}-V_0)}}{I_{DSAT}(V_{DD} - V_S - V_{TH})^{n(1+\Delta V_0/V_{DD}-V_0)}} \]

(4)

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

\[ \frac{\Delta V_O}{\Delta V_1} = \frac{1}{\frac{1}{\Delta V_1} + \frac{1}{(1+n)(V_{VOG} - V_S)}} \]

(5)

Voltage gain = \[ \frac{\Delta V_O}{\Delta V_1} = \frac{1}{\Delta V_1} + \frac{1}{\Delta V_1} + \frac{1}{\Delta V_1 + \Delta V_2} \]

(6)

This expression claims that the voltage gain is increased by decreasing \(1/\Delta V_2\) and by increasing \(V_5\) and \(V_{TH}\). This effect is ascribed 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 \(V_5\) is too high, the output voltage swing is limited since \(V_O\) cannot go low enough. In this sense, \(V_5\) 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 for a large current as possible for higher speed operation within the power constraint.

The input voltage \(V_1\) should be set as low as possible for the higher gain, but since \(V_5\) is usually a bit line voltage, it can not be lowered to ensure sufficient write margin. In most cases, \(V_1\) is set equal to \(V_{DD} - V_{TH}\) because of the NMOS bit line load. Then the size of \(Q_1\) can be determined. The size of PMOS \(Q_2\) is then determined by the requirement that the initial output voltage \(V_{VOG}\) is preferably about the center of the output swing. \(b_P\) is calculated using the following formula.

\[ V_{VOG} = V_{DD} - V_{TH} - \left( \frac{b_0}{b_P} \right) \frac{1}{n(1+b_0/V_{DD} - V_{TH})} \]

It should be noted that a smaller channel length than the minimum channel length is better for use of \(Q_1\) 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 \(Q_1\) cannot 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...
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 design that more than 5-input gates should be avoided. This design rule of thumb should be changed to "avoid more than 8-input logic gates" in lower sub-micron designs.

5. Mitigating Non-Scalability of Threshold Voltage

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

\[ I_{SUB} = 10^{4} \left( \frac{V_{GS} - V_{TH}}{s} \right) \left( \frac{k_{t}}{q} \right) \ln 10 \]

s and called an s factor. s is about 110mV/decade and cannot 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].

\[ t_{pd} = \frac{C_{L} V_{DD}}{V_{DD}-V_{TH}} \left[ 1 + \frac{1 - V_{DD}/V_{TH}}{1+n} \right] \left( \frac{0.8 + 0.6 V_{DD}/V_{TH}}{2} \right) \]

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 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 VLSIs.

- Smaller body effect
- Lower junction capacitance

References

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

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

Fig. 3 CMOSA behavior with various common source implementation

Fig. 4 CMOSA behavior change for various $V_S$ voltages

Fig. 5 CMOSA behavior change for various PMOS size

Fig. 6 CMOSA 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.10 Dependence of static noise margin on velocity saturation index

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

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

Fig.13 Delay dependence on threshold voltage

Fig.14 Measured sub-threshold current vs. Vgs