Reducing Leakage with Mixed-V_{th} (MVT)

Frank Sill[1], Frank Grassert[1], Dirk Timmermann[1]
[1]University of Rostock, Germany

Abstract

We present a new method for assignment of devices with different V_{th} in a double-V_{th}-process, whereas leakage is reduced and performance increases or is constant. A mixed-V_{th} gate type is developed, which renders new masks unnecessary. As compared with known methods, our approach achieves an additional leakage reduction of 25% while leakage reduction in raw designs is average 65%.

1. Introduction

Over the past decade mobile applications and devices became more important, as users want to communicate and work everywhere. This requires long operating times of devices, with consequently low power dissipation. Since CMOS devices scale in dimensions below 100 nm, power dissipation is not only determined by the switching power because leakage increases exponentially [1]. This is caused by aggressive reduction of threshold voltage, gate length, and physical oxide thickness, which rise short channel and quantum effects. The consequence for low power applications is the need for effective reduction of leakage power.

There are several ways to reduce leakage, like implementation of sleeping transistors, dynamic voltage scaling, or dynamic V_{th}-scaling [2][3]. Another common technique is Dual Threshold CMOS (DTCMOS), which has no area and control overhead in contrast to the methods above [4]. The idea of DTCMOS is the application of devices with different threshold voltages. Low threshold voltage (LVT) devices achieve a smaller delay than high threshold voltage (HVT), while causing increased leakage. The approaches presented in [5] and [6] detect the critical path and calculate a time slack for each gate. If time slack is long enough, gates are changed to HVT gates. A solution at transistor level is presented in [7]. After timing analysis and evaluation of the transistor time slacks, a priority is assigned to every transistor depending on delay and leakage. Transistors with the highest priority are checked first. If slack is long enough they will be replaced by high-threshold voltage transistors.

Solutions at transistor level require a lot of resources and are not useful for considerably designs, but optimization is very accurate. In contrast, solutions at gate level are less time-consuming, but fewer detailed. We propose an approach, which tries to combine advantages of both levels. This will be achieved by modified gate types and new algorithms.

This paper is organized as follows. In section 2, we introduce subthreshold current and delay model. Next, in section 3, we present evaluation of subthreshold current in stacks of different transistors. In section 4 we introduce our proposed mixed-V_{th} CMOS approach designs. In Section 5 simulation results are presented and section 6 summarizes this paper.

2. Preliminaries

Subthreshold current

Subthreshold current I_{sub}, which occurs when gate voltage is below threshold voltage V_{th}, is a main part of leakage current [2]. I_{sub} depends on different effects and voltages, which are formulated in following equations:

$$I_{sub} = I_0 \cdot e^{\frac{q}{nk_BT}\left(V_{gs}-V_{th0}+\gamma V_{bs}+\eta V_{ds}\right)}\left(1-e^{\frac{-qV_{ds}}{k_BT}}\right) \qquad (1)$$

where

$$I_0 = \mu \cdot \frac{W}{L}\sqrt{\frac{q\varepsilon_{Si}NDEP}{2\Phi_S}}\left(\frac{k_BT}{q}\right)^2 \qquad (2)$$

and

$$\eta \approx \frac{-0.5\cdot\left(ETA0+ETAB\cdot V_{bs}\right)}{cosh\left(DSUB\frac{L_{eff}}{\sqrt{\frac{\varepsilon_{Si}^{3/2}}{\varepsilon_{ox}\sqrt{q}}}\cdot\sqrt{\frac{T_{ox}}{\sqrt{NDEP}}}}\right)-1} \qquad (3)$$

and

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}}}{\varepsilon_{ox}}\cdot\sqrt{NSUB}\cdot T_{ox} \qquad (4)$$

where q is the electrical charge, T is the temperature, *n* is the subthreshold swing coefficient, k_B is the Boltzmann constant, η is the drain induced barrier lowering (DIBL) coefficient, γ is the body effect coefficient, μ is the mobility, V_{th0} is the zero-bias threshold voltage, V_{gs} is the gate-source voltage, V_{bs} is the bulk-source voltage, V_{ds} is the drain-source voltage, ε_{ox} and ε_{Si} are the gate dielectric constants of gate oxid and silicium, *NSUB* is the uniform substrate doping concentration and *NDEP* the channel doping concentration, T_{ox} is the thickness of the oxid layer, Φ_S is the surface potential, *DSUB* and *ETA0* are technology dependent DIBL coefficients, and *ETAB* is a body-bias coefficient of the BSIM4-Modell.

The delay T_d of a CMOS device can be approximated as follows

$$T_d = \frac{k' \cdot V_{dd} C_L}{(W/L)\cdot\left(V_{dd} - V_{th}\right)^{\alpha}} \tag{5}$$

where *k'* is a technology constant, C_L is the load, and α models the short channel effects [8].

Variation of V_{th} is a common technique to reduce leakage because I_{sub} exponentially scales with V_{th} (see equation 1). Thus, higher V_{th} results in lower leakage. However, from equation (5) follows higher V_{th} additionally results in longer delay [4]. Hence, the request is to optimize the application of low V_{th} (LVT) and high V_{th} devices (HVT).

3. Different V_{th} in a stack

It is necessary to know the behavior of transistor stacks with different V_{th} to understand the advantages of mixed-V_{th} circuits. This is shown by a comparison of a two transistor stack with equal V_{th} and a two transistor stack with different V_{th}. All transistors are equal dimensioned.

The behavior of transistor stacks is determined by the stack effect, which occurs when more than one transistor in the stack is turned off (figure 1). The intermediate node V_x has a positive voltage due to the small drain current. Hence, $V_{gs,u}$ is negative and from equation (1) follows, this exponentially reduces the leakage. Additionally, the DIBL is reduced by rising V_x. This increases V_{th} of the bottom device T_b and consequently decreases leakage.

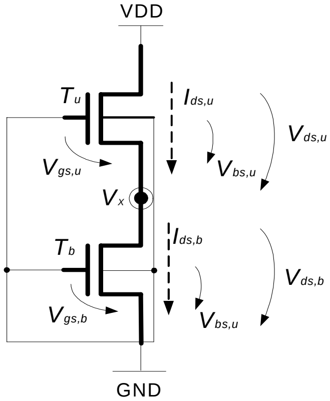

figure 1 Stack of NMOS transistors

At first, subthreshold current I_{sub} must be evaluated. This requires the determination of V_x. If we assume that subthreshold currents dominates, then from Kirchhoff's voltage law follows:

$$I_{dsu} = I_{dsb} \tag{6}$$

Combination of equation (1) and (6) in consideration of the example circuit (figure 1) results in:

$$\begin{aligned} & I_{0,u} \cdot e^{\frac{q}{nk_BT}\left(-V_x - V_{th,u} - \gamma V_x + \eta\left(V_{dd} - V_x\right)\right)}\left(1 - e^{\frac{-qV_{ds}}{k_BT}}\right) = \ldots \\ & \ldots I_{0,b} \cdot e^{\frac{q}{nk_BT}\left(-V_{th,b} + \eta V_x\right)}\left(1 - e^{\frac{-qV_{ds}}{k_BT}}\right) \end{aligned} \tag{7}$$

It can be approximated that $\frac{V_x}{V_{dd}} < 0.8$ and consequently $e^{\frac{-q}{k_BT}(V_{dd} - V_x)} \approx 0$. This results in:

$$1 - e^{\frac{-q(V_{dd} - V_x)}{k_BT}} \approx 1 \tag{8}$$

We assume that doping concentration of both devices is nearly equal, all transistors are equal dimensioned, and mobility is identical. Hence, in combination with equation (2) follows:

$$I_{0,u} \approx I_{0,b} \tag{9}$$

Combination of equations (7), (8), and (9) results in:

$$V_x = \frac{V_{th,b} - V_{th,u} + \eta V_{dd}}{\gamma + 1 + 2\eta} \tag{10}$$

That means, current I_{stack} through a stack of two transistors can approximately evaluated from:

$$I_{stack} = I_{ds,u} = I_0 \cdot e^{-V_{th,u} + \eta\frac{V_{th,b} - V_{th,u} + \eta V_{dd}}{\gamma + 1 + 2\eta}} \tag{11}$$

The ratio of subthreshold current I_{equ} in a stack of transistors with equal V_{th} versus subthreshold current I_{diff} of transistors with different V_{th} results in combination with equation (11) from:

$$\frac{I_{equ}}{I_{diff}} = \frac{I_{0,equ} \cdot e^{-V_{th}+\frac{\eta^2 V_{dd}}{\gamma+1+2\eta}}}{I_{0,diff} \cdot e^{-V_{th,u}+\eta\frac{V_{th,b}-V_{th,u}+\eta V_{dd}}{\gamma+1+2\eta}}} \approx \ldots \tag{12}$$
$$\ldots e^{V_{th,b}-V_{th}-\frac{\eta}{\gamma+1+2\eta}\cdot\left(V_{th,b}-V_{th,u}\right)}$$

If we want to show that I_{diff} is lower than I_{equ}, it is necessary that $\frac{I_{equ}}{I_{diff}} > 1$. That means:

$$V_{th,b} - V_{th} - \frac{\eta}{\gamma+1+2\eta}\cdot\left(V_{th,b} - V_{th,u}\right) > 0 \tag{13}$$

which is equal to:

$$V_{th,b} - V_{th} > \frac{1}{\frac{\gamma+1}{\eta}+2}\cdot\left(V_{th,b} - V_{th,u}\right) \tag{14}$$

The verification of this inequation requires at first the solution of term $\frac{\gamma+1}{\eta}$, which results from:

$$\frac{\gamma+1}{\eta} \approx \left(\frac{\sqrt{2q\varepsilon_{Si}}}{\varepsilon_{ox}}\cdot\sqrt{NSUB}\cdot T_{ox}+1\right)\cdot\left[0.5\left(ETA0+ETAB\cdot V_{bs}\right)\right]^{-1}\cdot\ldots \tag{15}$$
$$\ldots\left(\frac{1}{2}\left(\frac{DSUB\cdot L_{eff}\cdot NDEP^{1/4}}{\sqrt{\frac{\varepsilon_{Si}^{3/2}}{\varepsilon_{ox}\sqrt{q}}}\cdot\sqrt{T_{ox}}}\right)^2+\frac{1}{24}\left(\frac{DSUB\cdot L_{eff}\cdot NDEP^{1/4}}{\sqrt{\frac{\varepsilon_{Si}^{3/2}}{\varepsilon_{ox}\sqrt{q}}}\cdot\sqrt{T_{ox}}}\right)^4\right)$$

The solution of this equation with approximated values from current technologies results in:

$$-10^4 < \frac{\gamma+1}{\eta} < -10^1 \tag{16}$$

As example, we got for applied BPTM models $\frac{\gamma+1}{\eta} = -3.12\cdot 10^3$.

Next step is determination of threshold voltage ratio in both stacks. The constraint is that delay of both stacks is equal. Hence, the maximum delay of both stacks can be approximated as follows.

$$T_{d,stacked} = \frac{k'\cdot 2\cdot V_{dd}C_L}{(W/L)\cdot\left(V_{dd}-V_{th}\right)^{\alpha}}$$
$$= \frac{k'\cdot V_{dd}C_L}{(W/L)\cdot\left(V_{dd}-V_{th,u}\right)^{\alpha}} + \frac{k'\cdot V_{dd}C_L}{(W/L)\cdot\left(V_{dd}-V_{th,b}\right)^{\alpha}} \tag{17}$$

where V_{th} is the threshold voltage of equal transistors. Because $\alpha \approx 1.3$ [8], it can be approximated:

$$\left|V_{th} - V_{th,u}\right| \approx \left|V_{th,b} - V_{th}\right| \tag{18}$$

The last step of verification of statement in (14) is combination of (14) and (18), which results in:

$$V_{th,b} - V_{th} > \ldots$$
$$\ldots\frac{1}{\frac{\gamma+1}{\eta}+2}\left(\left[V_{th,b}-V_{th}\right]+\left[V_{th}-V_{th,u}\right]\right) \tag{19}$$

If $V_{th,b} > V_{th} > V_{th,u}$, then (19) results in:

$$1 > \frac{2}{2+\frac{\gamma+1}{\eta}} \tag{20}$$

From (16) follows $2+\frac{\gamma+1}{\eta} < 0$. In combination with (20) follows:

$$2+\frac{\gamma+1}{\eta} < 2 \tag{21}$$

and consequently:

$$\frac{\gamma+1}{\eta} < 0 \tag{22}$$

From (16) follows, statement in (22) is true and consequently (14) is true. That means:

$$I_{equ} > I_{diff} \tag{23}$$

Hence, at equal delay leakage in a stack of equal transistors is higher than in a mixed-V_{th} stack, where transistors have different V_{th}. However, the difference of both subthreshold currents, which mainly depends on difference of threshold voltages, varies from 3 to 10%.

But, if signal probability of each input is known, leakage can be strongly reduced because order of input signals is irrelevant in stacks. Leakage in mixed-V_{th} stacks is significant lower as in equal-V_{th} stacks when only one transistor is blocking and this is a low-V_{th} transistor. Consequently, signals with highest '*0*'-probability (NMOS-stack) or highest '*1*'-probability (PMOS-stack) should connect to high-V_{th} transistors via pin-reordering.

4. Mixed-V_{th} (MVT) CMOS Circuits

The ambition of mixed-V_{th} technique is reduction of leakage within a Dual Threshold Voltage (DVT) design, without decreasing performance and without rising mask costs. This will be achieved by optimization of fast LVT gates and by generating a new gate type. In [9] we presented two new gate types for standard cell processes.

MLVT gates

MLVT gates have the same maximum delay as LVT gates, but reduced leakage power dissipation. Only the transistors in the slower path within a gate have a low V_{th}.

MVT gates

Common Dual-V_{th} (DVT) designs consist of two gate types LVT and HVT. Hence, the gate-level optimization has only two degrees of freedom, which can result in high concentration of LVT gates. An approach to reduce the number of LVT gates is an additional gate type, which has smaller delay as HVT gates and lower leakage as LVT gates. This gate type could consist of transistors with normal-V_{th}, which is lower than high-V_{th} and higher than low-V_{th}. But this requires extra masks for the manufacturing process.

Our proposed approach is the mix of low-V_{th} and high-V_{th} transistors within stacks, which mostly form the critical path. As shown in section 3, mixed stacks have equal delay and lower leakage as stacks of transistor with normal-V_{th}. As explained, probability analysis and pin-reordering additionally reduces leakage. Furthermore, noncritical paths of these mixed-V_{th} (MVT) gates consist of high-V_{th} transistors.

5. Simulation Results

We created a library of LVT, MLVT, HVT, and MVT 2-input gates with the modified BPTM transistor models [10]. Next, we generated DVT, MVT with pin-reordering, and MVT without pin-reordering implementations of ISCAS'85 [11]. Subsequently, we simulated generated designs and a LVT version of each design. We set the signal probability of every design input to 50%.

Leakage power dissipation decreased by average 65% with proposed mixed-V_{th} technique compared to the LVT implementation. Thereby, main part of optimization is based on insertion of HVT gates in noncritical paths. Compared to DVT implementations the leakage power could be reduced by average 25% (figure 2).

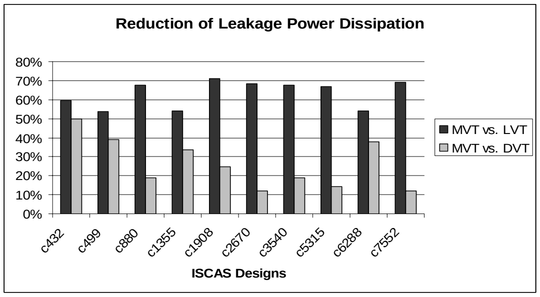

figure 2 Reduction of leakage with MVT

6. Conclusion

We proposed a mixed-V_{th} (MVT) CMOS design technique to reduce leakage of a design. Thereby, the performance of the design stays constant or increases. A new mixed-V_{th} gate type was generated, which can use beside LVT and HVT gates. Fortunately, no additionally masks for the manufacturing process are necessary. Simulation results indicate that mixed-V_{th} reduces leakage by average 65% in contrast to non optimized designs and by average 25% by usually used DVT designs.

REFERENCES

[1] N.S. Kim, et.al, *Leakage Current: Moore's Law Meets Static Power*, in IEEE Computer, p. 68, no. 12 (2003).

[2] M. Anis and M. Elmasry, *Multi-Threshold CMOS Digital Circuits*, Kluwer Academic Publishers (2003).

[3] J.Chang and M. Pedram, *Energy minimization using multiple supply voltages,* in IEEE Transactions on Very Large Scale Integration Systems, volume 4, pages 436–443, December (1997).

[4] J.K. Kao and A. Chandrakasan, *Dual-Threshold Voltage Techniques for Low-Power Digital Circuits,* in IEEE Journal of Solid State Circuits, p. 1009, no. 35 (2000).

[5] L.Wei, K. Roy, and C. Koh, *Power Minimization by Simultanous Dual-Vth Assignment and Gate-sizing,* in Proceedings of the IEEE Custom Integrated Circuits Conference, pp. 413-416 (2000).

[6] V.Sundararajan and K.Parhi, *Low Power Synthesis of Dual Threshold Voltage CMOS VLSI Circuits*, in Proceedings of the IEEE ISLPED, pp. 139-144 (1999).

[7] L.Wei, Z.Chen, and K.Roy, *Mixed-v_{th} (MVT) CMOS Circuit Design Methodology for Low Power Applications*, in Proceedings of the 36th Design Automation Conference, pp. 430-435 (1999).

[8] T. Sakurai and A. Newton, *Alpha-Power Law MOSFET Model and its Application to CMOS Inverter Delay and other Formulas,* in IEEE Journal of Solid-State Circuits, pp. 584-594, no. 2 (1990).

[9] F. Sill, F.Grassert, D. Timmermann, *Low Power Gate-level Design with Mixed-Vth (MVT) Techniques,* SBCCI 2004.

[10] Berkeley Predictive Technology Model, *www-device.eecs.berkeley.edu/~ptm* (2002).

[11] M. Hansen, H. Yalcin, and J. P. Hayes, *Unveiling the ISCAS-85 Benchmarks: A Case Study in Reverse Engineering*, in IEEE Design and Test, p. 72, no. 16 (1999).