Stack Sizing Analysis and Optimization for FinFET Logic Cells and Circuits Operating in the Sub/Near-Threshold Regime

Xue Lin, Yanzhi Wang, Massoud Pedram
Department of Electrical Engineering, University of Southern California, CA USA
E-mail: {xuelin, yanzhiwa, pedram}@usc.edu

Abstract

Sub/near-threshold computing has been proposed for ultra-low power applications. FinFET devices are considered as an alternative for bulk CMOS devices due to the superior characteristics, which make FinFET an excellent candidate for ultra-low power designs. In this paper, we first present an improved analytical FinFET model covering both sub- and near-threshold regimes. This model accurately captures the drain current as a function of both the gate and drain voltages. Based on the accurate FinFET model, we provide a detailed analysis on stack sizing of FinFET logic cells, and derive the optimal stack depth in FinFET circuits. We also provide a delay optimization framework for the FinFET circuits in the sub/near-threshold region, based on the stack sizing analysis. To the best of our knowledge, this is the first work that provides in-depth analysis of the stack sizing of FinFET logic cells in the sub/near-threshold region based on the accurate FinFET modeling. Experimental results on the 32nm Predictive Technology Model for FinFET devices demonstrate the effectiveness of the proposed optimization framework.

Keywords

FinFET device, sub/near-threshold, stack sizing

1. Introduction

For some emerging applications such as sensor networks and biomedical devices, low energy consumption instead of high performance becomes the primary concern for digital designs. To minimize energy consumption, voltage scaling techniques have proved rather effective with the subthreshold design representing the endpoint of voltage scaling [1]. Comparing subthreshold designs with traditional super-threshold designs, the energy consumption is reduced by 10-fold, whereas the delay is at least three orders of magnitude larger [2]. As a result of the performance penalty of subthreshold designs, near-threshold computing (NTC), a design space where the supply voltage is approximately equal to the threshold voltage of transistors, has been proposed to improve performance while retaining much of the energy savings of the subthreshold design [3].

FinFET devices, a kind of special quasi-planar double gate (DG) devices, are promising substitutes for the bulk CMOS devices at and beyond the 32nm technology node [4]. FinFET devices have lower gate leakage, stronger gate control, reduced short-channel effects, and less performance variability compared to bulk CMOS counterparts [5][6][7]. Because of these superior characteristics, FinFET devices show significant advantages in terms of performance and energy consumption for sub/near-threshold designs [8], and allow for higher voltage scalability.

Due to the transistor-width quantization effect, i.e., wider FinFET transistors are formed by utilizing multiple parallel-connected fins with the same height, sizing of FinFET devices is quite different from that of bulk CMOS devices. FinFET device sizing in the conventional super-threshold region has been investigated in [4][9]. However, the characteristics of FinFET devices in the sub/near-threshold region are significantly different from those in the conventional strong-inversion region [10]. The drain current of a FinFET transistor in the super-threshold region follows the α-power law model [11], whereas the EKV model [12] can be applied in the sub/near-threshold region. Unfortunately, the EKV model is difficult to provide back-of-the-envelope insights and is difficult to work with analytically. An empirical transregional model is proposed for bulk CMOS devices in both sub- and near-threshold regions, based on the exponential subthreshold model [13]. This empirical transregional model matches with the simulated characteristics with high accuracy only in the case where the gate and the drain of the transistor are tied together. When the gate and drain are tied to different signals, which is the common case in logic gates, this transregional model fails to match with the simulated characteristics, especially for FinFET devices. In this paper, we improve the transregional model for FinFET devices covering both sub- and near-threshold regimes. The proposed model is able to accurately capture the drain current as a function of the gate and drain voltages, and achieve high accuracy under all input voltage combinations. The proposed improved tranregional model can also be applied to bulk CMOS transistors.

Based on the accurate FinFET modeling in the sub- and near-threshold regions, we provide an effective solution to the FinFET logic cell stack sizing problem. In some logic gates, there are several transistors connected in series forming a stack, e.g., the pull-down network of a NAND gate or the pull-up network of a NOR gate. The stack sizing problem involves determining the sizes of transistors in a stack such that the gate achieves equal rise and fall times. Different from the super-threshold regime, in the sub/near-threshold regime the stack sizing problem becomes non-trivial. A comprehensive understanding of the stack sizing problem is critical for gate sizing in digital designs. We also optimize the stack depth, i.e., the number of transistors in a stack, in a FinFET logic gate (and the circuit.) An m-input NAND/NOR gate has a stack depth of m. Based on the stack sizing analysis, we compare logic cells with different stack depths and determine the optimal stack depth of logic cells in FinFET circuits.

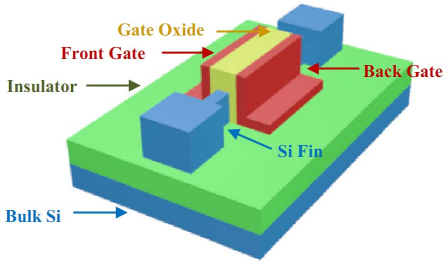

Fig. 1. Double-gate FinFET device structure.

With the comprehensive analysis of stack sizing, we conduct delay optimization for FinFET circuits in the sub- and near-threshold region. We propose a dynamic programming-based delay optimization framework to find the optimal solution in polynomial time complexity. Experimental results on HSpice simulation using 32nm Predictive Technology Model (PTM) for FinFETs [14] verify the effectiveness of the proposed optimization framework. To the best of our knowledge, this is the first work that provides an in-depth analysis of the stack sizing of FinFET logic cells in the sub/near-threshold regions based on the accurate FinFET modeling.

2. FinFET modeling in the sub- and near-threshold regions

2.1 FinFET basics

FinFET devices have lower gate leakage, stronger gate control, reduced short-channel effects, and less performance variability compared to bulk CMOS counterparts [5][6][7]. A FinFET device has a double-gate structure, where each fin contains two equivalent gates: a *front gate* and a *back gate*, as shown in Figure 1. Each fin is essentially the parallel connection of the *front-gate-controlled* FET and the *back-gate-controlled* FET, both with width W equal to the height of the fin. A FinFET device can work at two possible modes: the *double-gate mode*, where both the front and the back gates of the fin are tied to the same control signal, and the *independent-gate mode*, where the front and the back gates are tied to different control signals. Due to the capacitor coupling of the front gate and back gate, the threshold voltage of the front-gate-controlled FET varies in response to the back-gate biasing voltage, and vice versa. Within a relatively small range of the back-gate biasing voltage, a linear relationship is observed between the change of the threshold voltage and the back-gate biasing voltage (suppose that we consider an N-type FinFET):

$$\frac{dV_{th}}{dV_{bs}} = -\frac{C_{oxb} \cdot C_{Si}}{C_{oxf} \cdot (C_{oxb} + C_{Si})}, \qquad (1)$$

where C_{oxb}, C_{oxf}, and C_{Si} are the back gate capacitance, front gate capacitance, and body capacitance, respectively; V_{bs} is the back gate biasing voltage of the N-type fin; and V_{th} is the threshold voltage of the front-gate-controlled FET. Eqn. (1) shows that the decrease of the back-gate biasing voltage results in the increase of V_{th} of the front-gate-controlled N-type FET and therefore an exponential decrease of the leakage power. Figure 2 shows the relationship between V_{th} of the front-gate-controlled FET and the back gate biasing voltage for an N-type FinFET and a P-type FinFET from

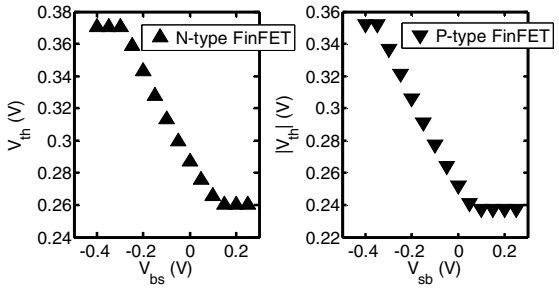

Fig. 2. Threshold voltage V_{th} of the front-gate-controlled FET v.s. the back gate biasing voltage V_{bs}.

Hspice simulation. The same relationship can be observed between V_{th} of the back-gate-controlled FET and the front gate biasing voltage.

2.2 Sub-threshold FinFET model

The drain current I_{ds}^{f} of a front-gate-controlled FET (say, the front-gate-controlled FET in an N-type fin) operating in the subthreshold regime satisfies an exponential dependency on the front gate drive voltage V_{gs}^{f} and the drain-to-source voltage V_{ds}, as given by:

$$I_{ds}^{f} = I_0 \frac{W}{L} \cdot e^{\frac{V_{gs}^{f} + \lambda V_{ds} - V_{th}(V_{gs}^{b})}{n \cdot v_T}} \cdot (1 - e^{\frac{-V_{ds}}{v_T}}), \qquad (2)$$

where I_0 is a technology-dependent parameter; W is the width of the front-gate-controlled FET in an N-type fin (equal to the height of the fin); L is the length of the fin; λ is the drain voltage dependence coefficient (similar to but much smaller than the DIBL coefficient for bulk CMOS devices); $V_{th}(V_{gs}^{b})$ is the threshold voltage of the front-gate-controlled FET as a function of the back gate biasing voltage V_{gs}^{b}; n is the subthreshold slope factor; v_T is the thermal voltage. Similarly, the drain current I_{ds}^{b} of a back-gate-controlled FET in an N-type fin is given by:

$$I_{ds}^{b} = I_0 \frac{W}{L} \cdot e^{\frac{V_{gs}^{b} + \lambda V_{ds} - V_{th}(V_{gs}^{f})}{n \cdot v_T}} \cdot (1 - e^{\frac{-V_{ds}}{v_T}}). \qquad (3)$$

2.3 Transregional FinFET model

An unified transregional model that covers both sub- and near-threshold regimes is proposed for bulk CMOS devices [13]. We generate the transregional model for FinFET devices. For the front-gate-controlled FET in an N-type fin, the transregional model for both sub- and near-threshold regimes is given as:

$$I_{ds}^{f} = I_0 \cdot \frac{W}{L} \cdot e^{\frac{\left(V_{gs}^{f} + \lambda V_{ds} - V_{th}(V_{gs}^{b})\right) - \alpha \cdot \left(V_{gs}^{f} + \lambda V_{ds} - V_{th}(V_{gs}^{b})\right)^2}{m \cdot v_T}} \cdot (1 - e^{\frac{-V_{ds}}{v_T}}), \qquad (4)$$

where α and m are empirical fitting parameters. We fit the values of I_0, α, and m from HSpice simulation results.

Figure 3 plots the simulated curve of the drain current of a front-gate-controlled N-type FET as a function of V_{DD} (we set $V_{gs}^{f} = V_{ds} = V_{DD}$) on a semi-logarithmic scale. We sweep

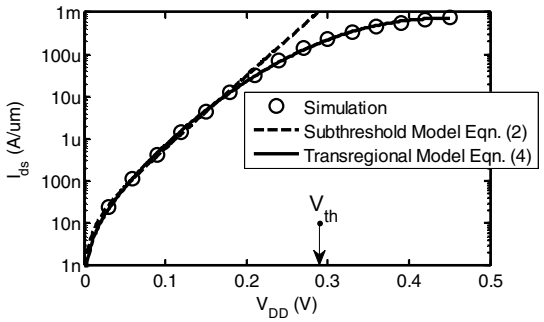

Fig. 3. I_{ds} v.s. V_{DD} from simulation, subthreshold model and transregional model of the front-gate-controlled N-type FET.

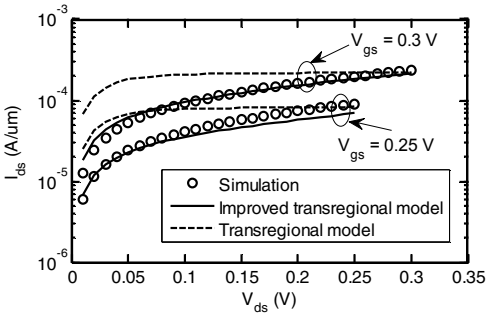

Fig. 4. I_{ds} v.s. V_{ds} from simulation, transregional model and the improved transregional model of the front-gate-controlled N-type FET.

the V_{DD} value in the sub- and near-threshold regimes, from 0~0.45 V. The threshold voltage of the front-gate-controlled N-type FET is about 0.29 V. Figure 3 also plots the subthreshold model Eqn. (2) and the transregional model Eqn. (4). We can observe that the subthresheld model is only accurate in the subthreshold region, whereas the transregional model is accurate for sub/near-threshold regimes.

2.4 Improved transregional FinFET model

The transregional model matches with simulation results with high accuracy for the case that the gate and the drain are tied together as shown in Figure 3. However, for the case that the gate and the drain are tied to different voltage levels, which is the common case in logic cells, the transregional model fails to match with simulation results. As can be seen from Figure 4, where I_{ds} is plotted as a function of V_{ds} at different V_{gs} values, the transregional model deviates from the simulation result significantly. To enhance the accuracy of the transregional model, we propose the improved transregional FinFET model for the sub- and near-threshold regimes, as given by:

$$I_{ds}^{f} = I_0 \cdot \frac{W}{L} \cdot e^{\frac{\left(v_{gs}^{f}+\lambda V_{ds}-V_{th}\left(v_{gs}^{b}\right)\right)-\alpha\cdot\left(v_{gs}^{f}+\lambda V_{ds}-V_{th}\left(v_{gs}^{b}\right)\right)^2}{m\cdot v_T}} \cdot \left(\beta\cdot\left(1-e^{\frac{-V_{ds}}{v_T}}\right)+\gamma\cdot V_{ds}\right), \quad (5)$$

where β and γ are new fitting parameters. Given the parameters I_0, α, and m from the transregional model Eqn. (4), we can fit the values of β and γ using the HSpice simulation results. From Figure 4, we can observe that the improved transregional model achieves much higher accuracy than the transregional model described in Section 2.3: the average error of the latter model is as high as 98.9%, whereas the average error of the improved transregional model is only 12.4%. Although the improved transregional model is proposed for FinFET devices, it can also be applied to bulk CMOS devices.

3. Stack sizing of FinFET logic cells

In order to design FinFET logic cells for sub/near-threshold computing, we need to solve the stack sizing problem. In some logic gates, there are several transistors connected in series forming a *stack*, e.g., the pull-down network of a NAND gate or the pull-up network of a NOR gate. The stack sizing problem involves determining the transistor sizes in a stack such that the gate achieves equal rise and fall times. We use the 2-input NAND gate as an example. Figure 5 shows an template inverter and a 2-input NAND gate, where FinFET devices operate in the *double-gate mode* (i.e., the front and back gates are tied together) and the number on top of a FinFET transistor symbol denotes the number of parallel connected fins. The template inverter has $W_P : W_N \approx 2$ for achieving equal rise and fall times in the sub/near-threshold region. We denote the stack sizing factor in an m-input NAND gate by $\rho_{N,m}$, where the subscript N denotes N-type FinFET devices. Similarly, the stack sizing factor in an m-input NOR gate is denoted by $\rho_{P,m}$. The stack sizing factor $\rho_{N,2}$ of the 2-input NAND gate is defined as the ratio of the number of N-type fins connected to an input signal in the 2-input NAND gate to that in the template inverter, such that the pull down network of the NAND gate has the same current driving strength as that in the template inverter. In the super-threshold region, $\rho_{N,2}$ can be simply set as 2. However, in the sub/near-threshold region the stack sizing problem becomes non-trivial due to the fact that the drain current has an exponential dependency on the terminal voltages.

The worst-case fall delay of the NAND gate can be obtained when input A is high and input B makes a transition from low to high. The discharging process can be separated into two phases. During the first phase, V_x drops to a relatively low value $V_{x,tar}$, whereas V_{out} remains near to V_{DD}. This is because at the beginning of phase one, I_B is much larger than I_A, because $V_{gs,A} \approx 0$ while $V_{gs,B} \approx V_{DD}$. In the sub/near-threshold regime, a larger gate voltage can increase the current exponentially. As V_x is decreasing, I_A increases exponentially and I_B decreases slowly. Therefore,

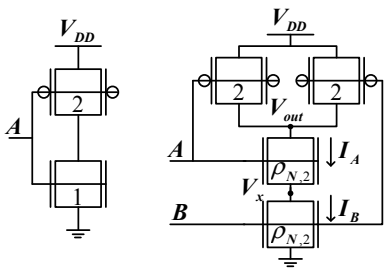

Fig. 5. Illustration of stack sizing for a 2-input NAND gate.

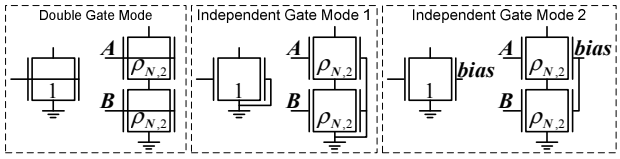

Fig. 6. Illustration of stack sizing in the pull-down network of a 2-input NAND gate: the double gate mode and the independent gate modes.

before the time when I_A becomes equal to I_B, the node x loses more charge than the node out. In addition, the capacitance at node x is typically smaller than that at node out. As a result, V_x drops to a relatively small value $V_{x,tar}$ rapidly, while V_{out} remains near to V_{DD} during phase one. Then the second phase starts when I_A reaches I_B, and V_{out} starts to drop in this phase. Since the second phase dominates the entire discharge process, we consider the second phase for calculating the stack sizing factor.

3.1 Stack sizing factor calculation

The stack sizing factor is calculated from

$$I^f_{ds,INV} + I^b_{ds,INV} = I^f_{ds,B} + I^b_{ds,B}. \quad (6)$$

In Eqn. (6), $I^f_{ds,INV}$ ($I^b_{ds,INV}$) is the drain current of the front(back)-gate-controlled N-type FET in the template inverter, which is given by:

$$I^f_{ds,INV} = I^b_{ds,INV} = I_0 \cdot \frac{W}{L} \cdot e^{\frac{\left(V_{DD}+\lambda V_{DD}-V_{th}(V_{DD})\right)-\alpha\cdot\left(V_{DD}+\lambda V_{DD}-V_{th}(V_{DD})\right)^2}{m\cdot v_T}} \cdot \left(\beta\cdot\left(1-e^{\frac{-V_{DD}}{v_T}}\right)+\gamma\cdot V_{DD}\right). \quad (7)$$

Moreover, $I^f_{ds,B}$ ($I^b_{ds,B}$) is the drain current of the front(back)-gate-controlled N-type FET connected to input B in the NAND gate, which is given by:

$$I^f_{ds,B} = I^b_{ds,B} = I_0 \cdot \frac{\rho_{N,2}\cdot W}{L} \cdot e^{\frac{\left(V_{DD}+\lambda V_{x,tar}-V_{th}(V_{DD})\right)-\alpha\cdot\left(V_{DD}+\lambda V_{x,tar}-V_{th}(V_{DD})\right)^2}{m\cdot v_T}} \cdot \left(\beta\cdot\left(1-e^{\frac{-V_{x,tar}}{v_T}}\right)+\gamma\cdot V_{x,tar}\right). \quad (8)$$

Because $\lambda \approx 0$ for FinFET devices, we calculate $\rho_{N,2}$ in the following way:

$$\rho_{N,2} = \frac{\beta\cdot\left(1-e^{\frac{-V_{DD}}{v_T}}\right)+\gamma\cdot V_{DD}}{\beta\cdot\left(1-e^{\frac{-V_{x,tar}}{v_T}}\right)+\gamma\cdot V_{x,tar}}. \quad (9)$$

We need to derive $V_{x,tar}$ before we can calculate $\rho_{N,2}$ from Eqn. (9). When V_x drops to $V_{x,tar}$, we have

$$I^f_{ds,A} + I^b_{ds,A} = I^f_{ds,B} + I^b_{ds,B}, \quad (10)$$

where $I^f_{ds,A}$ ($I^b_{ds,A}$) is the drain current of the front(back)-gate-controlled N-type FET connected to input A in the NAND gate, given by

TABLE I. Stack sizing factors for 2-input NAND/NOR gates.

$V_{DD} = 0.3\ V$ Near-threshold regime						
	Double gate		Independent gate mode 1		Independent gate mode 2	
	$\rho_{N,2}$	$\rho_{P,2}$	$\rho_{N,2}$	$\rho_{P,2}$	$\rho_{N,2}$	$\rho_{P,2}$
calculation	3.20	2.70	3.80	3.01	3.55	2.72
verification	3.05	2.70	4.10	3.20	3.65	2.85
$V_{DD} = 0.25\ V$ Subthreshold regime						
	Double gate		Independent gate mode 1		Independent gate mode 2	
	$\rho_{N,2}$	$\rho_{P,2}$	$\rho_{N,2}$	$\rho_{P,2}$	$\rho_{N,2}$	$\rho_{P,2}$
calculation	3.28	2.73	3.83	3.05	3.61	2.74
verification	3.35	2.85	4.30	3.35	3.95	2.95

TABLE II. Stack sizing factors for 3-input and 4-input NAND/NOR gates.

	Double gate mode		Independent gate mode 1		Independent gate mode 2	
	$\rho_{N,3}$	$\rho_{P,3}$	$\rho_{N,3}$	$\rho_{P,3}$	$\rho_{N,3}$	$\rho_{P,3}$
$V_{DD} = 0.3\ V$	5.35	4.40	7.70	5.65	6.70	4.80
$V_{DD} = 0.25\ V$	5.90	4.70	8.10	5.90	7.30	5.05
	$\rho_{N,4}$	$\rho_{P,4}$	$\rho_{N,4}$	$\rho_{P,4}$	$\rho_{N,4}$	$\rho_{P,4}$
$V_{DD} = 0.3\ V$	7.70	6.15	11.45	8.10	9.90	6.80
$V_{DD} = 0.25\ V$	8.50	6.60	12.15	8.50	10.80	7.10

$$I^f_{ds,A} = I^b_{ds,A} = I_0 \cdot \frac{\rho_{N,2}\cdot W}{L} \cdot e^{\frac{\left(V_a+\lambda V_a-V_{th}(V_a)\right)-\alpha\cdot\left(V_a+\lambda V_a-V_{th}(V_a)\right)^2}{m\cdot v_T}} \cdot \left(\beta\cdot\left(1-e^{\frac{-V_a}{v_T}}\right)+\gamma\cdot V_a\right). \quad (11)$$

In Eqn. (11), $V_a = V_{DD} - V_{x,tar}$. We find $V_{x,tar}$ from Eqns. (8), (10), (11) using numerical method. Then we substitute the value of $V_{x,tar}$ into Eqn. (9) to directly calculate $\rho_{N,2}$.

We have derived the stack sizing factor for the 2-input NAND gate in the double gate mode. Similarly, we can also derive the stack sizing factor for NAND gate in the independent gate mode. Figure 6 illustrate the differences between the double gate mode and the independent gate modes: in the double gate mode both the front and back gates of a FinFET device are tied to an input signal, in the independent gate mode 1 the back gate is tied to the ground, and in the independent gate mode 2 the back gate is tied to a controlled biasing voltage. Our stack sizing method is valid for both sub- and near-threshold regions, since it is based on the proposed improved transregional model, which is accurate for both sub- and near-threshold regimes. A comprehensive understanding of the stack sizing problem is critical for gate sizing in digital designs.

3.2 Stack sizing results

We summarize in Table I the stack sizing factor $\rho_{N,2}$ ($\rho_{P,2}$) for a 2-input NAND (NOR) gate in the double and independent gate modes. For the independent gate mode 2, a forward biasing voltage of 0.1 V is used, which is equivalent to $V_{bias} = 0.1\ V$ for N-type FinFET devices and $V_{bias} = V_{DD} - 0.1\ V$ for P-type FinFET devices. The stack sizing factors for both sub- and near-threshold regions are

calculated. In order to validate our stack sizing method, we also summarize the stack sizing factor values obtained from the HSpice simulation in Table I. We observe that our stack sizing method is accurate for both sub- and near-threshold regions and for different (independent or double) gate modes, which demonstrates the accuracy of our improved transregional model in both sub- and near-threshold regions. The average errors of our stack sizing method are 6.1% for N-type FinFET devices and 5.1% for P-type FinFET devices. Due to the transistor width quantization effect, the stack sizing factors may need to be rounded into integer numbers. In this sense, our stack sizing method can give almost the same results as those from the HSpice verification. Also, please note that the stack sizing factor for 2-input FinFET logic cells in the sub/near-threshold region are larger than 2, which is the typical stack sizing factor in the super-threshold regime.

We further calculate the stack sizing factors for 3-input and 4-input NAND and NOR gates at the sub/near-threshold supply voltage as shown in Table II. We can observe that (i) we need to heavily upsize the FinFET transistors in a stack in order to achieve the same delay as the template inverter; (ii) the stack sizing factors of logic cells in the independent gate modes is larger than those in the double gate mode; (iii) the stack sizing factors in the subthreshold regime ($V_{DD} = 0.25\,V$) are larger than the corresponding values in the near-threshold regime ($V_{DD} = 0.3\,V$). In fact, the stack sizing factors for FinFET devices are larger than the corresponding values for the bulk CMOS devices, because (i) the higher dependency of the FinFET driving current on V_{ds} due to the $\gamma \cdot V_{ds}$ term in Eqn. (5), and (ii) the dependency of the threshold voltage of the front(back)-gate-controlled FET on the voltage applied on the back (front) gate.

4. Investigation on the optimal stack depth

Based on the stack sizing analysis for different kinds of logic cells in different (double gate or independent gate) design modes, now we will investigate the optimal stack depth problem. An m-input NAND/NOR gate has a stack depth of m, since there is a stack comprised of m series-connected transistors in that gate. Usually, a logic function can be implemented using logic cells with different stack depths. Let us take the AND function as an example. The AND function is typically implemented using the NAND-NOR structure. Suppose we need a 16-input AND function. It can be implemented by a two-stage circuit comprised of four 4-input NAND gates in the first stage connected to one 4-input NOR gate in the second stage. Equivalently, it can be implemented by a four-stage circuit consisting of ten 2-input NAND gates (eight in the first stage and two in the third stage) and five 2-input NOR gates (four in the second stage and one in the last stage.) Selections of logic cells with different stack depths can affect the delay and area of the circuits. In order to investigate the optimal stack depth problem, we implement 16-input, 256-input and 4096-input AND functions using NAND and NOR gates with different stack depths. The NAND and NOR gates are sized such that both the pull-up network and the pull-down network have the same driving strength as the template inverter.

First, we use FinFET NAND and NOR gates in the double gate mode. The results are summarized in Table III. For an AND function, we design it using NAND and NOR gates with different stack depths and measure the corresponding area, delay, and area-delay product (ADP) values. The area is estimated in terms of the total number of fins in the design. In addition, we use FinFET NAND and NOR gates in the independent gate mode 1. The results are shown in Table IV.

TABLE III. Comparison of different stack depths in designs using the double gate mode.

V_{DD}		16-input AND		256-input AND		4096-input AND		
	Stack depth	2	4	2	4	2	3	4
0.3V	Area	160	212	2720	3604	43680	47697	57876
	Delay (ns)	0.0463	0.0480	0.0958	0.1031	0.1452	0.1496	0.1578
	ADP	7	10	261	372	6342	7135	9133
0.25V	Area	170	232	2890	3944	46410	52308	63336
	Delay (ns)	0.0787	0.0803	0.1647	0.1753	0.2503	0.2527	0.2702
	ADP	13	19	476	691	11616	13218	17113

TABLE IV. Comparison of different stack depths in designs using the independent gate mode 1.

V_{DD}		16-input AND		256-input AND		4096-input AND		
	Stack depth	2	4	2	4	2	3	4
0.3V	Area	140	220	2380	3740	38220	49209	60060
	Delay (ns)	0.0657	0.0745	0.1372	0.1632	0.2088	0.2266	0.2520
	ADP	9	16	327	610	7980	11151	15135
0.25V	Area	140	240	2380	4080	38220	50751	65520
	Delay (ns)	0.1313	0.1507	0.2740	0.3269	0.4176	0.4522	0.5035
	ADP	18	36	652	1334	15961	22950	32989

It can be observed from Table III and Table IV that for the AND function, a larger stack depth results in both a larger area and a larger delay in the sub- and near-threshold regimes, which is quite different from the super-threshold regime. For example, in Table IV for the 4096-input AND function, it has been designed with stack depths of 2, 3 and 4. When the stack depth is 2, it is implemented by a 12-stage circuit consisting of 2730 2-input NAND gates and 1365 2-input NOR gates; when the stack depth is 3, it is implemented by a 8-stage circuit consisting of 1537 3-input NAND gates and 514 3-input NOR gates; when the stack depth is 4, it is implemented by a 6-stage circuit consisting of 1092 4-input NAND gates and 273 4-input NOR gates. For the 4096-input AND function at $V_{DD} = 0.3\,V$, when the stack depths are 2, 3, and 4, the areas are 38220, 49209, and 60060, respectively, and the delay values are 0.2088 ns, 0.2266 ns, and 0.2520 ns, respectively. Comparing designs with stack depths of 2, 3 and 4, even though more stages and larger number of gates are needed in the design with stack depth of 2, it still has both smaller area and smaller delay. One can further observe from Table III and Table IV that this effect is more phenomenal when (i) the independent gate control technique is utilized and (ii) the circuit operates in the subthreshold regime compared with the near-threshold regime. Besides, FinFET logic cells with a small stack depth of 2 will have better soft-error immunity and larger noise margin with respect to the high impact of process variations in the sub/near-threshold regime, compared with those with

larger stack depth values. Therefore the stack depth of 2 is highly preferred for FinFET circuit designs in the sub/near-threshold region.

5. Delay optimization of FinFET circuits based on stack sizing

Based on the in-depth stack sizing analysis, we develop a delay optimization framework for FinFET circuits in the sub- and near-threshold regimes. The delay of a FinFET logic cell can be estimated with its sizing (determining the driving strength) and the sizing of its fanout gates (determining the output load capacitance), as given by:

$$d_i^r = f_i^r\left(x_i, y_i, c_i^{fanout}\right), \tag{12}$$

where d_i^r is the rise delay of the i-th gate along a circuit path; x_i and y_i are numbers of N-type and P-type fins connected to an input signal in the i-th gate; c_i^{fanout} is the output load capacitance of the i-th gate (i.e., c_i^{fanout} is given in the form of the total number of fins connected to the output of i-th gate); and the delay estimation function f_i^r can be characterized based on the type and connection mode (double gate mode or independent gate mode) of the i-th gate that are given in prior. Similarly, we can estimate the fall delay of the i-th gate i.e., d_i^f. Please note that the delay function (12) is general and can be realized using the logical effort method, the Elmore delay calculation, look-up tables, and so on. The effect of leakage current on the rise/fall delay, which is phenomenal in sub/near-threshold circuits, can also be accounted for in this function.

The delay optimization problem for a path in the FinFET circuit can be formulated as: ***Given*** the input capacitance C_{in} and the output load capacitance C_{out} of a path consisting of K gates, ***Find*** the optimal values of x_i and y_i for $1 \le i \le K$, to ***Minimize*** the delay of this path. We develop a dynamic programming-based algorithm to find the optimal solution in polynomial time complexity. We can derive a general optimization problem of discrete gate sizing based on iteratively executing the path delay optimization on the critical paths of a circuit [15]. However, this is out of the scope of the current paper because the main target of this paper is to shed some light on the stacking effect on FinFET gate sizing in the sub/near-threshold regime. Hence, we will use the delay optimization method in some symmetric circuits to analyze the stacking effect.

Let us use the AND function as an illustration. In the sizing optimization process, we maintain matrices $\boldsymbol{D}^r$, $\boldsymbol{D}^f$, $\boldsymbol{X}$, and $\boldsymbol{Y}$. For example, $\boldsymbol{D}^r(i, z_{i+1})$ strores the optimal rise delay from the input of the first gate to the output of the i-th gate when the input capacitance of the $(i+1)$-st gate is z_{i+1}, i.e., a total number of z_{i+1} fins including both N-type and P-type connected to the output of the i-th gate. Suppose that we want to calculate $\boldsymbol{D}^r(i, z_{i+1})$. We already know the values of $\boldsymbol{D}^r(i-1, z_i)$ (and also $\boldsymbol{D}^f(i-1, z_i)$) for all the z_i values. Therefore, we only need to find the optimal values of x_i and y_i in this step (we have $x_i + y_i = z_i$). We provide some details in the following. When x_i, y_i, and z_{i+1} are given, the delay from the circuit input to the output of the i-th gate is calculated by:

$$\max\begin{Bmatrix}\boldsymbol{D}^r(i-1, x_i+y_i) + f_i^f(x_i, y_i, z_{i+1})\\ \boldsymbol{D}^f(i-1, x_i+y_i) + f_i^r(x_i, y_i, z_{i+1})\end{Bmatrix}. \tag{13}$$

We need to find the optimal x_i^{opt} and y_i^{opt}, such that Eqn. (13) is minimized. And store the values of $\boldsymbol{D}^r(i, z_{i+1})$, $\boldsymbol{D}^f(i, z_{i+1})$, $\boldsymbol{X}(i, z_{i+1})$, and $\boldsymbol{Y}(i, z_{i+1})$ as:

$$\boldsymbol{D}^r(i, z_{i+1}) = \boldsymbol{D}^f\left(i-1, x_i^{opt}+y_i^{opt}\right) + f_i^r\left(x_i^{opt}, y_i^{opt}, z_{i+1}\right), \tag{14}$$

$$\boldsymbol{D}^f(i, z_{i+1}) = \boldsymbol{D}^r\left(i-1, x_i^{opt}+y_i^{opt}\right) + f_i^f\left(x_i^{opt}, y_i^{opt}, z_{i+1}\right), \tag{15}$$

$$\boldsymbol{X}(i, z_{i+1}) = x_i^{opt}, \tag{16}$$

$$\boldsymbol{Y}(i, z_{i+1}) = y_i^{opt}. \tag{17}$$

We perform *trace back* to find the optimal size of each FinFET gate after we have derived $\boldsymbol{D}^r(K, C_{out})$ and $\boldsymbol{D}^f(K, C_{out})$ at the end of this path. More details about the algorithm are shown below.

Please note that the optimal x_i and y_i values derived from this algorithm may result in unbalanced rise and fall delays for the i-th gate, which may be actually preferred in delay minimization for FinFET circuits in the sub/near-threshold regime. For example, considering a 2-input NOR gate in the circuit, if we want to achieve equal rise and fall delays in the subthreshold region ($V_{DD} = 0.25\,V$), it should be sized as $x_i = 1$ and $y_i = 6$. If y_i is reduced, it will result in a smaller input capacitance of this gate, and therefore a smaller delay of the previous gate.

We use the proposed delay optimization framework to size the gates in AND functions, only using the NAND and NOR gates with stack depth of 2. We compare the results from the proposed delay optimization method to those from

---

Algorithm: Dynamic Programming Algorithm for Delay Optimization

---

Input: the input capacitance C_{in} and the output load capacitance C_{out} of a path consisting of K gates.

Output: the number of N-type fins x_i and the number of P-type fins y_i connected to an input signal in the i-th gate $(1 \le i \le K)$.

Initialize the values of $\boldsymbol{D}^r(0, z_1)$ and $\boldsymbol{D}^f(0, z_1)$ to zeros.

For i from 1 to K:

  For each possible z_{i+1} value:

    Find the x_i^{opt} and y_i^{opt} values such that the objective function (13) is minimized.

    Store the x_i^{opt} and y_i^{opt} values in $\boldsymbol{X}(i, z_{i+1})$ and $\boldsymbol{Y}(i, z_{i+1})$, respectively.

    Calculate $\boldsymbol{D}^r(i, z_{i+1})$ and $\boldsymbol{D}^f(i, z_{i+1})$ according to (14) and (15), respectively.

  End

End

Perform trace back and find the optimal values of x_i and y_i for each gate in the path.

TABLE V. Comparison of proposed delay optimization method and the baseline methods at $V_{DD} = 0.30\ V$.

		256-input AND function			4096-input AND function		
C_{in}	C_{out}	Proposed (ps)	Baseline1 (ps)	Baseline2 (ps)	Proposed (ps)	Baseline1 (ps)	Baseline2 (ps)
5	500	173.5	325.4	194.0	134.4	228.6	168.6
5	1000	182.8	380.8	203.4	135.1	257.2	179.0
5	2000	194.9	465.3	222.1	136.4	280.5	177.4
10	500	168.8	250.6	180.7	121.8	247.1	161.8
10	1000	176.7	267.7	191.0	122.4	266.9	167.0
10	2000	188.5	351.2	202.7	123.5	303.4	172.4

TABLE VI. Comparison of proposed delay optimization method and the baseline methods at $V_{DD} = 0.25\ V$.

		256-input AND function			4096-input AND function		
C_{in}	C_{out}	Proposed (ps)	Baseline1 (ps)	Baseline2 (ps)	Proposed (ps)	Baseline1 (ps)	Baseline2 (ps)
5	500	300.2	539.7	344.9	216.9	383.9	287.9
5	1000	315.7	638.1	372.5	217.9	413.6	317.9
5	2000	348.4	779.8	386.6	218.6	456.6	315.4
10	500	294.8	612.0	316.8	218.5	430.9	283.3
10	1000	311.5	549.7	337.4	218.8	475.4	292.6
10	2000	346.6	654.8	367.7	221.3	539.6	310.8

the baseline methods. The baseline methods size the gates to have equal rise and fall delays, and also the logic gates along a path are gradually sized up according to $(x_{i+1} + y_{i+1})/(x_i + y_i) = (C_{out}/C_{in})^{1/stage\_num}$. In the baseline method 1, NAND and NOR gates with stack depth of 4 are used, and in the baseline method 2, gates with stack depth of 2 are used.

The results from the proposed method and the baseline methods for near-threshold regime and the subthreshold regime are summarized in Tables V and VI, respectively. C_{in} and C_{out} are given in the form of the fin numbers. As can be observed from the tables, our proposed method can achieve delay values up to 59.3% smaller than the baseline method 1 and 30.8% smaller than baseline method 2. Comparing the proposed method with the baseline method 2, we can find that unbalanced sizing for individual gates can reduce the overall delay of a path. We define the unbalance factor of a path as

$$avg_{1 \le i \le K}\left(\left|1 - \frac{S_i^{up}}{S_i^{down}}\right|\right), \qquad (18)$$

where S_i^{up} is the driving strength of the pull-up network in the i-th gate, and S_i^{down} is the driving strength of the pull-down network in the i-th gate. For a perfectly balanced path, Eqn. (18) is zero. In Table VI, for the 4096-input AND function when $C_{in} = 5$ and $C_{out} = 2000$, the unbalance factor from the proposed method is 0.83 while the unbalance factor from the baseline method 2 is 0.39. Please note that even though baseline method 2 sizes the gates to have equal rise and fall delays, the unbalance factor from the baseline method 2 is larger than zero due to the transistor-width quantization effect. The rounding process that is required in the baseline methods will also deteriorate the performance of the baseline methods when no optimization is performed. Comparing baseline method 1 to baseline method 2, we again prove that a stack depth of 2 is highly preferred in the sub/near-threshold region.

6. Conclusion

In this paper, we investigate the gate sizing problem of FinFET circuits in the sub/near-threshold region. We start from an accurate FinFET modeling method. We present an improved empirical FinFET model that is accurate in both sub- and near-threshold regimes. The model accurately captures the drain current as a function of terminal voltages. After that, we conduct an in-depth analysis on the stack sizing problem of FinFET logic cells, which is the basis for FinFET circuit gate sizing. In the end, we develop a delay optimization framework for FinFET circuits operating in the sub/near-threshold region. To the best of our knowledge, this is the first work that provides an in-depth analysis of the stack sizing of FinFET logic cells in the sub/near-threshold region based on an accurate FinFET modeling.

7. Acknowledgements

This research is sponsored in part by grants from the PERFECT program of the Defense Advanced Research Projects Agency and the Software and Hardware Foundations of the National Science Foundation.

8. References

[1] B.H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, Sep 2005.

[2] D. Markovic, C. Wang, L. Alarcon, T. Liu, and J. Rabaey, "Ultralow-power design in near-threshold region," *Proc. IEEE*, vol. 98, no. 2, Feb 2010.

[3] R.G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: reclaiming Moore's law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, no. 2, Feb 2010.

[4] P. Mishra, A. Muttreja, and N. K. Jha, "FinFET circuit design," *Nanoelectronic Circuit Design*, Springer, 2011.

[5] T. J. King, "FinFETs for nanoscale CMOS digital integrated circuits," *Proc. of Int'l Conference on Computer-Aided Design (ICCAD)*, 2005.

[6] T. Matsukawa, S. O'uchi, K. Endo, Y. Ishikawa, H. Yamauchi, Y. X. Liu, J. Tsukada, K. Sakamoto, and M. Masahara, "Comprehensive analysis of variability sources of FinFET characteristics," *Proc. of Symposium on VLSI Technology*, 2009.

[7] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20nm FinFET design with predictive technology models," *Proc. of the 49th Design Automation Conf. (DAC)*, Jun 2012.

[8] F. Crupi, M. Alioto, J. Franco, P. Magnone, M. Togo, N. Horiguchi, and G. Groeseneken, "Understanding the basic advantages of bulk FinFETs for sub- and near-threshold logic circuits from device measurements," *IEEE Trans. on Cicuits and Systems II*, vol. 59, no. 7, Jul 2012.

[9] B. Swahn, and S. Hassoun, "Gate sizing: FinFETsvs

32nm bulk MOSFETs," *Proc. of the 43rd Design Automation Conf. (DAC)*, Jun 2006.

[10] X. Lin, Y. Wang, and M. Pedram, "Joint sizing and adaptive independent gate control for FinFET circuits operating in multiple voltage regimes using the logical effort method," *Proc. of Int'l Conf. on Computer Aided Design (ICCAD)*, Nov. 2013.

[11] T. Sakurai, and R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, Apr 1990.

[12] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current application," *Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, Jul 1995.

[13] D.M. Harris, B. Keller, J. Karl, and S. Keller, "A transregional model for near-threshold circuits with application to minimum-energy operation," *Proc. of Int'l Conference on Microelectronics (ICM)*, Dec 2010.

[14] W. Zhao, and Y. Cao, "New generation of predictive technology model for sub-45nm early design exploration," *IEEE Trans. on Electronic Devices*, vol. 53, no. 11, Nov 2006.

[15] O. Coudert, R. Haddad, and S. Manne, "New algorithms for gate sizing: a comparative study," in Proc. of *DAC*, 1996.