Received: 29 April 2020 | *Revised:* 27 May 2020 | *Accepted:* 28 May 2020
DOI: 10.37917/ijeee.16.1.11 *Vol. 16 | Issue 1 | June 2020*

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Optimizing the Performance of MOS Stacks

Sherif M. Sharroush
Electrical-Engineering Department, Faculty of Engineering, Port Said University, Port Said, Egypt

Correspondence
*Sherif M. Sharroush
Egypt, Port Said, 262 Saad-Zaghloul Street
Email: smsharroush@gmail.com

Abstract
CMOS stack circuits find applications in multi-input exclusive-OR gates and barrel-shifters. Specifically, in wide fan-in CMOS NAND/NOR gates, the need arises to connect a relatively large number of NMOS/PMOS transistors in series in the pull-down network (PDN)/pull-up network (PUN). The resulting time delay is relatively high and the power consumption accordingly increases due to the need to deal with the various internal capacitances. The problem gets worse with increasing the number of inputs. In this paper, the performance of conventional static CMOS stack circuits is investigated quantitatively and a figure of merit expressing the performance is defined. The word "performance" includes the following three metrics; the average propagation delay, the power consumption, and the area. The optimum scaling factor corresponding to the best performance is determined. It is found that under the worst-case low-to-high transition at the output (that is, the input combination that results in the longest time delay in case of logic "1" at the output), there is an optimum value for the sizing of the PDN in order to minimize the average propagation delay. The proposed figure of merit is evaluated for different cases with the results discussed. The adopted models and the drawn conclusions are verified by comparison with simulation results adopting the 45 nm CMOS technology.

KEYWORDS: Area, CMOS stack, optimization, power consumption, time delay.

I. INTRODUCTION

CMOS circuits that can be implemented using the universal NAND or NOR gates may contain long stacks of NMOS or PMOS transistors if the fan-in is wide. The main problem associated with these circuits is the relatively slow response and high power consumption. Multi-input exclusive-OR gates that are required in applications such as parity-check and error-correction circuits or some built-in testing circuits and barrel-shifters 0 are types of applications that may include a long stack of series-connected transistors. In this paper, the problem of the slow response and high power consumption of the wide fan-in stack circuits will be discussed quantitatively along with a brief survey of the previous work related to this problem. Then, three performance metrics (the average propagation delay, the power consumption, and the area) for assessing the performance of these circuits will be discussed.

A figure of merit that includes these three performance metrics is adopted and the effect of the exponential sizing strategy on the area, the low-to-high propagation delay, the high-to-low propagation delay, and the power consumption is investigated quantitatively. The optimization of the performance of these circuits, i.e. maximizing the defined figure of merit, through the proper sizing of the stacked transistors will then be discussed with the optimum sizing determined.

The remainder of this paper is organized as follows: Section II discusses the problem of degraded performance of CMOS stack circuits in detail. Section III discusses a brief survey of the previous work related to enhancing the performance of CMOS stack circuits. The analysis performed in this paper in order to assess the performance of CMOS stack circuits will be presented in Section IV with the associated results and discussion presented in Section V. The simulation results will be presented in section VI. Finally, the paper will be concluded in Section VII with points for future work suggested in Section VIII.

II. PROBLEM STATEMENT

The problem of degraded performance associated with circuits containing MOSFET transistor stacks will be discussed in detail. Refer to Fig. 1 for a NAND gate with n inputs realized in the static complementary CMOS logic-circuit family. Assume that the parasitic capacitance at the output node, Y, is charged to a voltage of V_{DD} through any one or a combination of the PMOS transistors of the pull-up network (PUN). Now, if all the inputs, A_1, A_2,……, A_{n-1}, and A_n, are at logic "1," then all the NMOS transistors in the stack will be activated. The parasitic capacitance at the output node have to discharge through the stack.

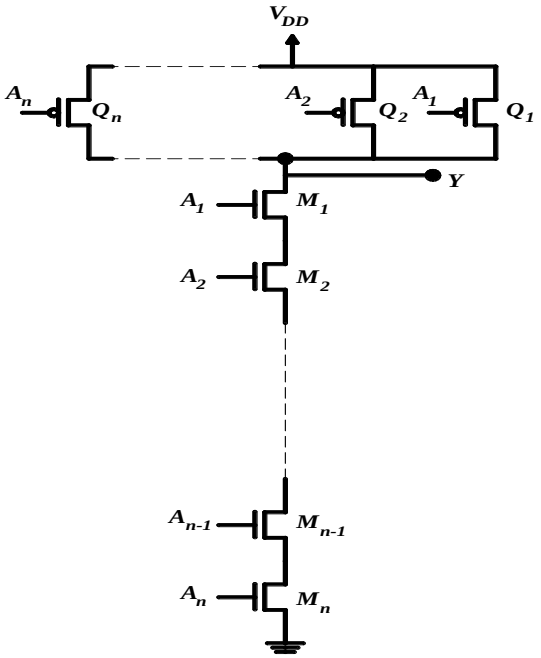

Fig. 1: A typical NAND gate with *n* inputs implemented in static complementary CMOS logic.

However, the discharging process will be slow due to the following reasons [2]:

1. Due to the voltage division between the serially connected NMOS transistors, the drain-to-source voltages of all the NMOS transistors except the uppermost one will be lower than their gate-overdrive voltages, thus they operate in the triode region during the whole discharging process. However, M_1 (refer to Fig. 1) operates in the saturation region from the onset of the discharging process to the instant of time at which its V_{DS} will be smaller than its $V_{GS} - V_{thn}$ after which it enters the triode region. Since the current in the saturation region is relatively independent of the drain-to-source voltage and since the current in the entire stack is limited by that of the uppermost device, the discharging current will be relatively independent of V_{DS} during this time interval, thus slowing down the operation. In fact, you can imagine the stack of the NMOS transistors as *n* resistors connected with each other in series. The larger the number of the transistors in the stack, the larger the total resistance will be with the result that the discharging current decreases.

2. The time interval during which the uppermost device operates in the saturation region elongates with technology scaling, thus slowing down the operation further.

3. Due to the parasitic capacitances at the internal nodes, there will be an initial voltage at the source of each transistor. So, the gate-to-source voltage of the corresponding transistor will be small, thus reducing the discharging current significantly.

4. It is obvious from the previous discussion that the source voltage of each transistor will be higher than 0 V except the lowermost one. Assuming that the substrate terminals of all the NMOS transistors in the stack are at 0 V. So, the body-source junction of each transistor except the lowermost one will be reverse-biased, resulting in an increase in the threshold voltage [3] of each transistor in the stack except the lowermost one. Increasing the threshold voltage will certainly reduce the discharging current. In fact, the increase in the threshold voltage of the transistors in the stack can be considered a double-edged weapon. Besides the reduction in the discharging current, the subthreshold-leakage current depends exponentially on the negative of the threshold voltage [4]. Thus, the subthreshold-leakage current decreases significantly through the stack.

5. Assuming the worst-case scenario, if all the inputs are activated except A_n which is connected to the lowermost transistor, then the load capacitance as well as all the internal capacitances will charge to V_{DD}. If then all the inputs are activated, all these capacitances have to discharge to ground. At the beginning of the discharging process, the V_{DS} voltages across all except the lowermost transistor will be at 0 V. So, the upper internal capacitances cannot discharge until the V_{DS} voltages across the discharging transistors rise above 0 V. The lowermost internal capacitance will thus begin discharging through M_1, then the V_{DS} voltage of M_2 will increase allowing the associated internal capacitance to discharge and so on. The upper internal capacitances at the output node will thus spend a certain time interval during which it cannot discharge, thus staying at V_{DD}. This is known as the *plateau voltage* and is shown in Fig. 2.

Certainly, increasing the size of the transistors in order to reduce the delay results in increasing the internal capacitances with the associated increase in the dynamic-switching power consumption.

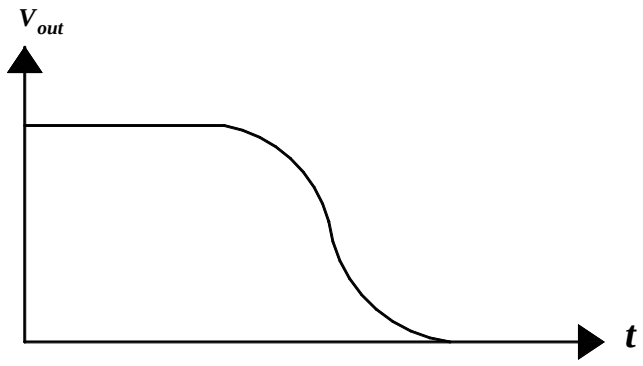

Fig. 2: The output voltage versus time illustrating what is known as the plateau voltage.

The above discussion applies equally well to the stack of PMOS transistors. The phenomenon of plateau voltage exists also in the PUN except that it is associated with a time interval during which the output voltage stays at 0 V at which the internal capacitances charge. Taking into account that the PMOS transistor requires an area larger than that of the NMOS transistor to obtain the same current due to the lower mobility of holes compared to that of free electrons, the problem of slow response will be more perceptible in

NOR gates than in NAND gates. This is due to the larger associated parasitic capacitances. Pseudo NMOS and domino logic-circuit families are not better than static-complementary family in this respect. In fact, the always activated PMOS device in the first family and the PMOS keeper in the second one dictates using a larger size for the PDN in order to combat the contention current and discharge the output node.

In the next section, a brief survey of the previous work related to enhancing the performance of CMOS stack circuits will be presented.

III. PREVIOUS WORK

The schemes used to reduce the time delay of the stack circuits can in general be classified into three categories. The first one relates to the use of additional circuits to enhance the performance of the existing circuit. The reader is referred to [5] - [6] for circuits of this type. However, using an additional circuit adds to the circuit cost and thus its usage needs to be justified in terms of the required area and power consumption. The second category is concerned with using an alternative circuit. Using alternative circuits to static CMOS stacks is usually associated with more sensitivity to process variations. A circuit of this type can be found in [7] - [9]. The third category relates to the proper sizing of the transistors in the stack. In this section, the previous work related to the sizing will be portrayed.

Changing the size of the transistors in the stack is a rudimentary solution. There are five schemes for sizing the transistors in the stack; *uniform*, *linear*, *exponential*, a combination of the uniform and exponential (or linear), and a combination of the linear and exponential sizing schemes. The *uniform*-sizing strategy simply means increasing the transistor channel widths in the stack so that all the transistor widths will be the same [10]. This sizing strategy, although decreases the transistor equivalent resistances, it causes the internal capacitances as well as the output capacitance to increase. So, the decrease in the delay caused by the reduction of the transistor equivalent resistances is partially compensated for by the associated increase in the parasitic capacitances. Thus, the net decrease of the delay will be either small or there will be no reduction at all. This scaling strategy reduces the delay significantly when the load capacitance at the output terminal is much larger than the parasitic capacitances at the source/drain junctions of the transistors in the string. However, when the load capacitance at the output terminal is on the same order of magnitude as those due to the transistors, this scaling strategy causes little or no reduction at all in the delay. This point will be confirmed in Section V.

Shoji proposed sizing the transistors in the stack such that there will be a positive gradient in the transistor channel width in the direction from the output terminal to the ground terminal in NMOS stacks. That is, in NMOS stacks, the transistor nearest the ground terminal has the largest channel width with the width decreasing upward and thus the transistor nearest the output terminal has the smallest channel width. This process is known as *tapering* [11] (refer to Fig. 3 for illustration). Transistor tapering is also appropriate for use with domino logic-circuit family if the channel width of the lowermost transistor is restricted such that the impact on the clock distribution network is minimal [12], thus avoiding the redesign of such systems.

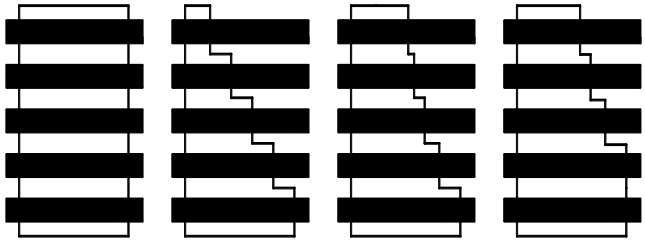

Fig. 3: Illustrating some types of tapering in MOS stacks. (a) Without tapering, (b) Linear tapering, (c) Exponential tapering, (d) Combination of unifrom and exponential tapering [13].

The transistor channel widths need thus to be scaled according to some scaling function. According to the first scaling function, the transistor channel widths are related to the width of the lowermost transistor by the relationship

$$W_1 = W_n - (n-1)(\Delta W), \tag{1}$$

where W_n is the channel width of the nth transistor with $n = 1$ representing the uppermost transistor and thus W_1 is the channel width of the transistor connected to the output node and ΔW represents the amount of reduction in the channel width. This sizing strategy is aptly called the *linear* sizing since the difference between the channel widths of any two consecutive transistors is the same as shown in Fig. 3 (b). The second sizing strategy dictates scaling down the channel widths according to this relationship [12]:

$$W_1 = \alpha^{n-1} W_n, \tag{2}$$

where α is a scaling factor that is certainly smaller than unity. This sizing strategy is aptly called the *exponential* sizing since the ratio between the channel widths of any two consecutive transistors is the same as shown in Fig. 3 (c). The third sizing strategy is a combination of the uniform and either the linear or the exponential sizing strategies. According to this strategy, some portion of the lower transistors are made to have the same size with the linear or exponential sizing applied on the transistors closest to the output terminal as shown in Fig. 3 (d). In fact, Ding et al. claimed that the last one is the best choice in reducing the time delay [13]. However, Choudhary et al. claimed that putting the largest transistor in the middle of the stack in a hill-like fashion reduces both the power dissipation and area [14]. The effects of tapering on the delay and the power consumption was investigated experimentally in [15]. In [16], the optimization of the power-delay product is optimized quantitatively using an *RC* model. However, there are no quantitative attempts reported about the effect of the

sizing of the NMOS transistors in the PDN on the low-to-high propagation delay or on the power consumption. Also, no quantitative attempts were reported about the effect of tapering on the three adopted performance metrics in a single figure of merit.

All the discussion and analysis performed in this paper are concerned with the NMOS stacks. However, they can be applied equally well to PMOS stacks with replacing NMOS by PMOS, ground by power supply, discharge by charge, high-to-low transition by low-to-high transition, and pull-down network by pull-up network.

IV. Analysis

In the analysis performed in this paper, the exponential-sizing strategy will be adopted with the scaling factor being the only design variable in varying the transistor channel widths. However, changing the transistor channel widths causes both the delay, the power consumption, and the area to change in opposite directions, thus resulting in contradicting effects [16]. As an illustration of the contradiction that may occur when acting to enhance these performance metrics simultaneously, consider this example. Increasing the transistor widths causes the time delay to decrease, however at the cost of increasing the power consumption due to the associated increase in the parasitic capacitances and increasing the area. In addition, the 50% time-delay point can be delayed for certain values of C_L and channel tapering [16] with a benefit of reducing the 90%-to-10% fall time. This causes a reduction in the time interval during which the output waveform reduces from V_{DD} - $|V_{thp}|$ to V_{thn}. Since this time interval is that during which both the NMOS and PMOS devices of the driven stage conduct, the short-circuit current will be reduced.

In this section, the derivation of the compact forms of the three performance metrics; time delay, power consumption, and area will be presented. Since these three performance metrics are preferred to be at their minimum, the adopted figure of merit for assessing the performance of the stack circuits will be defined as

$$FOM = \frac{1}{APt_p}. \quad (3)$$

The three performance metrics are involved in the denominator of the *FOM*, thus the circuit performance will be at its optimum state when the *FOM* is at its maximum. Nevertheless, the priority of a certain performance metric depends on the application at hand. For example, in portable applications, the power consumption and the area are expected to be the two most important parameters in order to reduce the weight of the product and prolong the battery lifetime. So, these two metrics must be stressed in the defined *FOM*. On the other hand, in military applications or servers, the time delay is the most important parameter. In order to emphasize the importance of a certain performance metric, the *FOM* can be modified to be in the form

$$FOM = \frac{1}{A^{a_1}P^{a_2}t_p^{\ a_3}}, \quad (4)$$

in which these parameters are raised to the exponents, a_1, a_2, and a_3. The values of these exponents are to be determined by the designer and signify the relative importance of a certain performance metric.

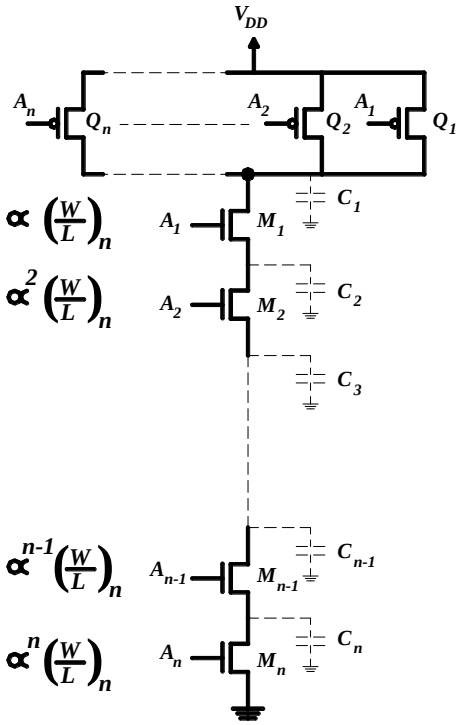

Fig. 4: The stack of Fig. 1 with the internal capacitances and sizes indicated.

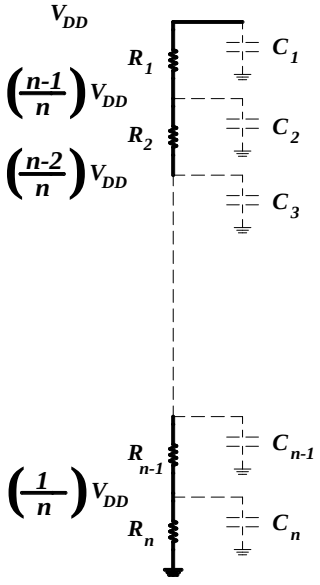

Fig. 5: Representing the PDN of Fig. 4 by an *RC* circuit. The initial voltages at the drain terminals of the transistors are also indicated.

Now, refer to Fig. 4 for the circuit schematic of an *n*-input NAND gate using the static complementary CMOS logic-circuit family with the proper sizing and the associated capacitances indicated. In this analysis, the exponential sizing strategy will be adopted with α representing the scaling factor. So, the size of the ith transistor is $\alpha^i(W/L)_n$. The terminology adopted in the analysis will be illustrated in the following subsection.

A. Terminology

Refer to Table I for the parameters adopted in the analysis along with their indication and adopted values for the 45 nm CMOS technology.

B. The Average Propagation Delay

There are two strategies for estimating the time delay [3]; the contamination time delay which represents the minimum or best-case scenario and the propagation time delay which represents the maximum or worst-case scenario. Toward finding these time delays, the transistor stack need to be modeled appropriately. The average propagation delay, t_p, is defined as the average of t_{PHL} and t_{PLH}. Thus,

$$t_p = \frac{t_{PHL} + t_{PLH}}{2}. \quad (5)$$

One estimation for these time delays uses the *RC*-tree computation as the primary operation in which a given network is partitioned into a spanning tree and links. Then, the signal delay is calculated and operated as the links are inadvertently added to reconstruct the original network [19]. A short-channel MOSFET model has been used in delay estimation by Sakurai et al. [20]. In [21], the time delay was estimated using empirical observations of time constants.

In the NMOS stack, the uppermost transistor, M_1, operates in the saturation region at the onset of discharging the output capacitance before entering the triode region. However, the remaining transistors operate in the triode region for the whole discharging time interval. So, for the first time interval, M_1 can be replaced by a current source and the remaining transistors can be represented by resistors [22]. To further simplify the analysis, we will represent all the NMOS transistors in the stack by their equivalent resistances. Upon representing each transistor by an equivalent resistance, the transistor stack can be represented by a simple *RC* first-order low-pass model as shown in Fig. 5, thus allowing us to easily derive compact forms for the delay.

Elmore's formula [23] of signal delay has been widely used to approximate the time taken for a signal to start from an initial value and reach half of its final value through an *RC* tree. Based on this approach, the low-to-high (or the high-to-low) propagation delay from the power supply (or ground) to the ith node is approximately given by [23] and [24]

$$t_p = (\ln 2)\sum_{k=1}^{N} C_k R_{ik}, \quad (6)$$

TABLE I

THE PARAMETERS ADOPTED IN THE ANALYSIS [17] AND [18].

Parameter	*Indication*	*Adopted Value*
n	The number of the inputs in the stack	8
W	The transistor channel width	45 nm
L	The transistor channel length	45 nm
$(W/L)_n$	The aspect ratio of NMOS devices	1
$(W/L)_p$	The aspect ratio of PMOS devices	1
k_n'	Process-transconductance parameter of the NMOS device	638 μA/V^2
k_p'	Process-transconductance parameter of the PMOS device	249 μA/V^2
V_{DD}	Power-supply voltage	1 V
V_{thn0}	NMOS threshold voltage with zero body effect	0.25 V
V_{thni}	NMOS threshold voltage of the ith transistor with body effect	0.25 V
V_{thp}	PMOS threshold voltage of any PMOS device with no body effect	-0.32 V
V_{GSi}	The gate-to-source voltage of the ith NMOS transistor	
V_{GSavg}	The average gate-to-source voltage of the n transistors in the stack	
V_{DSi}	The drain-to-source voltage of the ith NMOS transistor	
V_{SBi}	The source-to-body voltage of the ith NMOS transistor	
λ_n	Channel-length modulation effect parameter of the NMOS device	0.1 V^{-1}
λ_p	Channel-length modulation effect parameter of the PMOS devices	0.1 V^{-1}
C	The parasitic capacitance associated with each terminal of a minimum-sized NMOS device	1 fF
C_i	The internal capacitance associated with the drain terminal of the ith NMOS transistor in the stack with i = 1 and n for the uppermost and the lowermost transistors, respectively.	1 fF
C_L	The load capacitance due to the fan out of the stage	10 fF
R_i	The equivalent resistance of the ith NMOS transistor	
R_P	The equivalent resistance of any PMOS device	
α_{sw}	The switching activity	1
α	The parameter representing the scaling factor in the exponential-sizing strategy	1.2
β	The ratio by which the PMOS devices are sized relative to NMOS devices	2
f	Frequency of switching	1 GHz
t_{PHL}	The high-to-low propagation delay	
t_{PLH}	The low-to-high propagation delay	
t_p	The average propagation delay	
t_f	The fall time of the output waveform	
γ	The body-effect parameter	0.16 V$^{1/2}$
k	The linearized body-effect coefficient	0.08
P_{sw}	The dynamic-switching power consumption	
P_{sc}	The short-circuit power consumption	
P	The total estimated power consumption	
A	The total estimated area	
FOM	The figure of merit	
a_1	The factor to which the area is raised in the figure of merit	1
a_2	The factor to which the power consumption is raised in the figure of merit	1
a_3	The factor to which the average propagation delay is raised in the figure of merit	1

where R_{ik} is the resistance common to the path from the driving gate to node i and the path from the driving gate to node k. The factor (ln2) was used due to estimating the delay at the 50% point. t_{PHL} can be found using the Elmore-delay formula as follows:

$$t_{PHL} = (\ln 2)[(R_1 + R_2 + ... + R_n)C_1 + (R_2 + ... + R_n)C_2 + ...(R_{n-1} + R_n)C_{n-1} + R_nC_n] \quad (7)$$

Now, the evaluation of t_{PLH} follows. The worst-case low-to-high propagation delay occurs when all the NMOS transistors in the PDN are activated except the lowermost one, M_n, and correspondingly only one PMOS transistor is activated. In this case, all the internal capacitances are required to be charged and their charging will be relatively slow. Adopting the Elmore-delay formula, we obtain

$$t_{PLH} = (\ln 2)[R_PC_1 + (R_P + R_1)C_2 + ... + (R_P + R_1 + R_2 + ... + R_{n-1})C_n] \quad (8)$$

where R_P is the equivalent resistance of the PMOS device. Before completing the estimation of the time delays, the models adopted for the resistances and capacitances associated with the transistors in the stack will be described. In Section VI, these models will be verified by comparison with simulation results.

i. Resistance Model

There are various approaches to modeling the transistor by an equivalent resistance [25] and [26]. Shoji proposed using a factor that takes into account the discrepancies between a linear resistor and a practical four-terminal MOSFET transistor [27]. To simplify the analysis in this paper, we will represent each NMOS transistor by an equivalent resistance that is given by

$$R = \frac{V_{DSavg}}{I_{avg}}, \quad (9)$$

where V_{Dsavg} and I_{avg} are the average drain-to-source voltage and the average drain current of the corresponding NMOS transistor, respectively. The Shichman-Hodges square-law MOSFET model will be adopted [28]. Assuming that all the transistors in the stack operate in the deep-triode region, that is with $V_{DS} << 2(V_{GS} - V_{thn})$, then the term $1/2V_{DS}^2$ can be safely neglected with respect to $(V_{GS} - V_{thn})V_{DS}$. So,

$$R = \frac{V_{DS}}{k_n'\left(\frac{W}{L}\right)_n\left[(V_{GS} - V_{thn})V_{DS} - \frac{1}{2}V_{DS}^2\right]} \approx \frac{V_{DS}}{k_n'\left(\frac{W}{L}\right)_n (V_{GS} - V_{thn})V_{DS}} \approx \frac{1}{k_n'\left(\frac{W}{L}\right)_n (V_{GS} - V_{thn})} \quad (10)$$

If the initial voltage across C_1, V_{DD}, is assumed to be equally divided across the n-NMOS transistors in the stack as indicated in Fig. 5 [2], then

$$V_{GS1} = V_{DD} - \left(\frac{n-1}{n}\right)V_{DD} = \frac{V_{DD}}{n} \quad (11)$$

$$V_{GS2} = V_{DD} - \left(\frac{n-2}{n}\right)V_{DD} = \frac{2V_{DD}}{n},$$

$$....., V_{GSn-1} = V_{DD} - \left(\frac{n-(n-1)}{n}\right)V_{DD} = \frac{(n-1)V_{DD}}{n},$$

$$V_{GSn} = V_{DD} - 0 = V_{DD}.$$

The average gate-to-source voltage is thus

$$V_{GSavg} = \frac{V_{GS1} + V_{GS2} + ...+ V_{GSn-1} + V_{GSn}}{n} = V_{DD}\left(\frac{n+1}{2n}\right) \quad (12)$$

For large values of n, V_{GSavg} approaches $V_{DD}/2$ as expected. This is because the source voltage of the uppermost transistor approaches V_{DD} with the result that V_{GS} of lowermost and uppermost transistors will be at 0 V and V_{DD}, respectively. Now, the threshold voltage of the ith transistor in the stack can be written as [3]

$$V_{thni} = V_{thn0} + kV_{SBi}, \quad (13)$$

where k is the linearized body-effect coefficient. So,

$$V_{thn1} = V_{thn0} + kV_{SB1} = V_{thn0} + kV_{DD}\left(\frac{n-1}{n}\right) \quad (14)$$

$$V_{thn2} = V_{thn0} + kV_{SB2} = V_{thn0} + kV_{DD}\left(\frac{n-2}{n}\right),$$

$$....., V_{thnn-1} = V_{thn0} + kV_{SBn-1} = V_{thn0} + k\left(\frac{V_{DD}}{n}\right),$$

$$V_{thn} = V_{thn0} + kV_{SBn} = V_{thn0}.$$

The threshold voltage certainly increases with moving upward due to the body effect. The gate-overdrive voltages of the NMOS transistors in the stack are thus

$$V_{GS1} - V_{thn1} = V_{DD}\left[\frac{1}{n} - k\left(\frac{n-1}{n}\right)\right] - V_{thn0} = \frac{V_{DD}}{n}\left[1 - k(n-1)\right] - V_{thn0} \quad (15)$$

$$V_{GS2} - V_{thn2} = \frac{V_{DD}}{n}\left[2 - k(n-2)\right] - V_{thn0},$$

$$V_{GSn-1} - V_{thnn-1} = \frac{V_{DD}}{n}\left[(n-1) - k\right] - V_{thn0},$$

$$V_{GSn} - V_{thnn} = V_{DD} - V_{thn0}.$$

Physically stated, the effective-gate voltage of the uppermost NMOS transistor in the stack, M_1, decreases due to the body effect of the lower $n-1$ transistors and that of M_2 is degraded due to the lower $n-2$ transistors and so on. The average gate-overdrive voltage is thus

$$(V_{GS} - V_{thn})_{avg} = \frac{V_{DD}}{n}\left[\left(\frac{n+1}{2}\right) - k\left(\frac{n-1}{2}\right)\right] - V_{thn0} \quad (16)$$

Eq. (16) makes a physical sense as increasing the body-effect coefficient of the transistors in the stack causes the effective-gate voltage to degrade. The values of the equivalent resistances are thus

$$R_1 = \frac{1}{k_n'\alpha\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}} \quad (17)$$

$$R_2 = \frac{1}{k_n'\alpha^2\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}$$

$$\ldots\ldots\ldots, R_{n-1} = \frac{1}{k_n'\alpha^{(n-1)}\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}$$

$$R_n = \frac{1}{k_n'\alpha^{n}\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}.$$

So, according to the resistance model adopted in this estimation, the difference between the values of the equivalent resistances of the transistors in the stack is assumed to be solely due to the difference of their aspect ratios not their effective-gate voltages. This is suitable for our estimation since the target is to investigate the effect of the transistor sizing. The underestimated effective-gate voltages of the lower transistors will be compensated by the overestimated effective-gate voltages of the upper transistors. So, the net total resistance of the stack will be close to the real value.

ii. Capacitance Model

Some of the transistor internal capacitances are voltage-dependent and thus performing an exact analysis is a formidable task. So, we will in this analysis adopt the simplification that each terminal of the transistor has an associated capacitance and that all the associated capacitances of a transistor are equal and directly proportional to the aspect ratio of that transistor [3]. Taking into account that the wiring as well as the load capacitance due to the fan out, C_L, appears in parallel with the parasitic capacitance associated with the drain terminal of M_1, then

$$C_1 = n\beta C + \alpha C + C_L = C_L + C(n\beta+\alpha), \quad (18)$$

$$C_2 = \alpha C + \alpha^2 C = \alpha C(1+\alpha),$$

$$\ldots\ldots\ldots\ldots, C_{n-1} = \alpha^{n-2}C + \alpha^{n-1}C = \alpha^{n-2}C(1+\alpha),$$

$$C_n = \alpha^{n-1}C + \alpha^n C = \alpha^{n-1}C(1+\alpha).$$

Substituting by the values of the equivalent resistances and the internal capacitances in Elmore-delay formula results in

$$t_{PHL} = \frac{(\ln 2)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}$$

$$\left[\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots..+\frac{1}{\alpha^n}\right)\left[C(\alpha+n\beta)+C_L\right]+\left(\frac{1}{\alpha^2}+\frac{1}{\alpha^3}+\ldots..+\frac{1}{\alpha^n}\right)\alpha C(1+\alpha)\right.$$
$$\left.+\left(\frac{1}{\alpha^3}+\frac{1}{\alpha^4}+\ldots..+\frac{1}{\alpha^n}\right)\alpha^2 C(1+\alpha)+\ldots\ldots..+\frac{1}{\alpha^n}\alpha^{n-1}C(1+\alpha)\right] \quad (19)$$

$$= \frac{(\ln 2)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[C(\alpha+n\beta)+C_L\right]\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots..+\frac{1}{\alpha^n}\right)$$
$$+\frac{(\ln 2)C(1+\alpha)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots..+\frac{1}{\alpha^{n-1}}\right)+\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots..+\frac{1}{\alpha^{n-2}}\right)\right.$$
$$\left.+\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}+\ldots..+\frac{1}{\alpha^{n-3}}\right)+\ldots\ldots..+\left(\frac{1}{\alpha}+\frac{1}{\alpha^2}\right)+\frac{1}{\alpha}\right]$$

After simple mathematical manipulations, we can state that

$$t_{PHL} = \frac{(\ln 2)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[\frac{C(\alpha+n\beta)+C_L}{\alpha}\frac{\left[1-\left(\frac{1}{\alpha}\right)^{n-1}\right]}{\left[1-\left(\frac{1}{\alpha}\right)\right]}\right]$$
$$+\frac{(\ln 2)C(1+\alpha)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[\frac{n-1}{\alpha}+\frac{n-2}{\alpha^2}+\ldots..+\frac{1}{\alpha^{n-1}}\right]$$

$$= \frac{(\ln 2)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[\frac{C(\alpha+n\beta)+C_L}{\alpha}\frac{\left[1-\left(\frac{1}{\alpha}\right)^{n-1}\right]}{\left[1-\left(\frac{1}{\alpha}\right)\right]}\right] \quad (20)$$
$$+\frac{(\ln 2)C(1+\alpha)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\sum_{i=1}^{n-1}\left(\frac{n-i}{\alpha^i}\right)$$

The series in the second term can be manipulated by a change of variables to obtain [29]

$$\sum_{i=1}^{n-1}\left(\frac{n-i}{\alpha^i}\right)=\sum_{r=1}^{n-1}\left(\frac{r}{\alpha^{n-r}}\right)=\alpha^{-n}\frac{\alpha\left[1-n\alpha^{n-1}+(n-1)\alpha^n\right]}{(1-\alpha)^2} \quad (21)$$

The summation can be substituted from Eq. (21) into Eq. (20) to obtain

$$t_{PHL} = \frac{(\ln 2)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\left[\frac{C(\alpha+n\beta)+C_L}{\alpha}\frac{\left[1-\left(\frac{1}{\alpha}\right)^{n-1}\right]}{\left[1-\left(\frac{1}{\alpha}\right)\right]}\right]$$
$$+\frac{(\ln 2)C(1+\alpha)}{k_n'\left(\frac{W}{L}\right)_n (V_{GS}-V_{thn})_{avg}}\alpha^{-n}\frac{\alpha\left[1-n\alpha^{n-1}+(n-1)\alpha^n\right]}{(1-\alpha)^2} \quad (22)$$

To find the equivalent resistance of the PMOS device, R_P, the regions of operation of this device need to be determined. During the charging process of the internal and load capacitances, any PMOS device operates in the saturation region for a certain time interval then in the triode region. So, the equivalent resistance of any PMOS device, RP, will be estimated as the average between the equivalent

resistances in the saturation and triode regions, R_{Psat} and $R_{Ptriode}$. R_{Psat} and $R_{Ptriode}$ are given by

$$R_{Psat} = \frac{V_{SDavgsat}}{I_{avgsat}} \tag{23}$$

and

$$R_{Ptriode} = \frac{V_{SDavgtriode}}{I_{avgtriode}}. \tag{24}$$

$V_{SDavgsat}$ and $V_{SDavgtriode}$ are the averages of the V_{SD} voltages across the PMOS device in the saturation and the triode regions, respectively, and are given by $0.5(V_{DD} + V_{DD} - |V_{thp}|)$ and $0.5(V_{DD} - |V_{thp}|)$. I_{avgsat} and $I_{avgtriode}$ are the corresponding average currents. After substituting for the equivalent resistances and internal capacitances, performing simple mathematical manipulations, and using these two identities [29]

$$\sum_{i=1}^{n-1}\alpha^i = \alpha\frac{(1-\alpha^{n-1})}{(1-\alpha)} \tag{25}$$

and

$$\sum_{i=1}^{n-1} i\alpha^i = \alpha\left[\frac{1-n\alpha^{n-1}+(n-1)\alpha^n}{(1-\alpha)^2}\right], \tag{26}$$

we can write

$$t_{PLH} = (\ln 2)R_P[C(\alpha+n\beta)+C_L] + (\ln 2)R_P C(1+\alpha)\alpha\frac{(1-\alpha^{n-1})}{(1-\alpha)} + \frac{(\ln 2)C(1+\alpha)}{k_n'\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}}\left[\frac{n(1-\alpha^{n-1})}{(1-\alpha)} - \frac{1-n\alpha^{n-1}+(n-1)\alpha^n}{(1-\alpha)^2}\right] \tag{27}$$

On the other hand, the best-case low-to-high propagation delay occurs when the n PMOS devices are activated and consequently, all the NMOS devices are deactivated with the result that C_1 only will be charged. So,

$$t_{PLH} = (\ln 2)\frac{R_P}{n}C_1 = (\ln 2)\frac{R_P}{n}[C(\alpha+n\beta)+C_L] \tag{28}$$

C. The Power Consumption

The power consumption of an integrated circuit includes the leakage, the contention, the switching, and the short-circuit components. The first two can be safely neglected compared with the third and fourth components in static CMOS. The switching-power consumption includes the power required to charge the parasitic capacitances associated with the gate terminals of the NMOS and PMOS transistors, P_{sw1}, added to that required to charge the internal capacitances, P_{sw2}. P_{sw1} and P_{sw2} are given by

$$P_{sw1} = \alpha_{sw} f V_{DD}^2\left[\alpha C + \alpha^2 C + ...+ \alpha^n C\right] = \alpha_{sw} f C V_{DD}^2 \alpha\frac{(1-\alpha^n)}{(1-\alpha)} \tag{29}$$

$$P_{sw2} = \alpha_{sw} f V_{DD}^2 C_L + \alpha_{sw} f V_{DD}^2 C\left[(n\beta+\alpha)+\alpha(1+\alpha)\frac{(1-\alpha^{n-1})}{(1-\alpha)}\right] \tag{30}$$

The short-circuit power consumption in the driven stage can be computed as follows: Assuming that the output waveform of the NAND gate under investigation is as shown in Fig. 6 in which it is approximated by a linear straight line. t_f is the fall time between the two instants of time at which V_{out} is at V_{DD} - $|V_{thp}|$ and at V_{thn}, respectively.

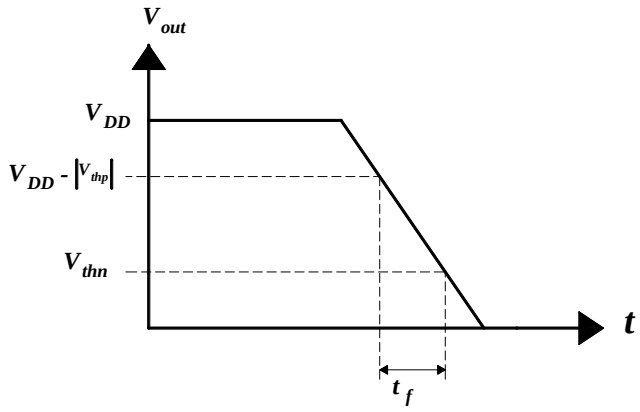

Fig. 6: The approximation of the output waveform of the NAND gate under study by a linear straight line.

The fall time, t_f, can be estimated from the amount of charge drained from C_1 during this time interval in two ways. The first one is through

$$I_{avg}t_f = C_1\Delta V_{out}, \tag{31}$$

where ΔV_{out} is the change of the output voltage during this time interval which is equal to V_{DD} - $|V_{thp}|$ - V_{thn} and I_{avg} is the average discharging current of C_1. The second way is to approximate all the transistors in the PDN by the same resistance value as follows: During this time interval and for typical values of V_{thn} and V_{thp}, all the transistors in the NMOS stack including the uppermost one operate in the deep-triode region in which V_{DS} is much smaller than (V_{GS} - V_{thn}). So, each transistor can be replaced by an equivalent resistance which is equal to

$$R \approx \frac{1}{k_n'\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})} \tag{32}$$

where V_{DS} was neglected compared to $2(V_{GS} - V_{thn})$ in Eq. (32). Substituting by the average gate-overdrive voltage into Eq. (32) and combining these resistances into a single one lead to

$$R_{total} = \frac{1}{k_n'\alpha\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}} + \frac{1}{k_n'\alpha^2\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}} + ...+ \frac{1}{k_n'\alpha^n\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}}$$
$$= \frac{1}{k_n'\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}}\left[\frac{1}{\alpha}+\frac{1}{\alpha^2}+.....\frac{1}{\alpha^n}\right] = \frac{1}{k_n'\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}}\left[\frac{\left(\frac{1}{\alpha}\right)\left(1-\left(\frac{1}{\alpha}\right)^n\right)}{1-\left(\frac{1}{\alpha}\right)}\right]$$
$$= \frac{1}{k_n'\left(\frac{W}{L}\right)_n(V_{GS}-V_{thn})_{avg}}\left[\frac{\left(1-\left(\frac{1}{\alpha}\right)^n\right)}{\alpha-1}\right] \tag{33}$$

Assume that the total capacitance at the output node, C_1, is much larger than the parasitic capacitances at the internal nodes. The last assumption makes sense due to two reasons. The first one is that the output node is connected to all the parallel PMOS devices of the PUN, thus each PMOS device adds its own parasitic capacitance to the output node. The second reason is that the short-circuit power consumption is significant only when the rise or fall times of the output waveform are relatively large which is the case with large fan out.

Now, neglecting the parasitic capacitances at the internal nodes leads to the equivalent circuit shown in Fig. 7 where the serially connected resistances are combined into a single resistance, R_{total}.

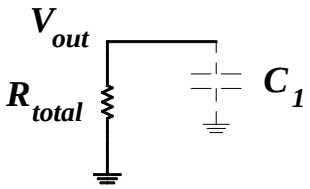

Fig. 7: The approximate equivalent circuit to the NMOS stack where the internal capacitances are neglected and the NMOS stack is replaced by a single resistance, R_{total}.

The current discharging C_1 can be written as

$$i = -C_1 \frac{dV_{out}}{dt} = \frac{V_{out}}{R_{total}} \tag{34}$$

Integrating both sides of Eq. (34) with the integration limits of 0 and t with respect to time and V_{DD} - $|V_{thp}|$ and $V_{out}(t)$ with respect to V_{out} results in

$$V_{out}(t) = \left(V_{DD} - |V_{thp}|\right) e^{\frac{-t}{R_{total}C_1}}. \tag{35}$$

V_{out} equals V_{thn} when t is equal to t_f. So,

$$t_f = R_{total} C_1 \ln\left[\frac{\left(V_{DD} - |V_{thp}|\right)}{V_{thn}}\right]. \tag{36}$$

Assuming that the output of the stack circuit is connected to a symmetric CMOS inverter, then the short-circuit power consumption in this inverter is given by [24]

$$P_{sc} = \frac{\alpha_{sw} k_n' \left(\frac{W}{L}\right)_n t_f \left(V_{DD} - V_{thn} - |V_{thp}|\right)^3}{12}, \tag{37}$$

where t_f is given by Eq. (36).

D. The Area

In estimating the total area of the stack, we will adopt the approximation that the area of a certain transistor is equal to its channel area [30]. The area is thus

$$A = WL\left[n\beta + \alpha + \alpha^2 + ... + \alpha^n\right]$$

$$A = WL\left[n\beta + \alpha\left(1 + \alpha + \alpha^2 + ... + \alpha^{n-1}\right)\right]$$

$$A = WL\left[n\beta + \alpha\left(\frac{1-\alpha^n}{1-\alpha}\right)\right]. \tag{38}$$

V. Results and Discussion

In this section, the various performance metrics and the proposed figure of merit, *FOM*, will be plotted versus the scaling factor, α, with the results discussed. Also, the impact of technology scaling will be discussed.

A. The Three Performance Metrics

Refer to Figs. 8, 9, 10, and 11 for the plots of the dynamic-switching power, the short-circuit power, the total power, and the plots of the average worst- and best-case propagation delays versus α. Increasing α causes the dynamic-switching power consumption, P_{sw}, to increase due to the associated increase in the internal capacitances. On the other hand, increasing α causes the discharging process of C_1 to speed-up, thus reducing the fall time, t_f. The reduction of P_{sc} with increasing α is thus expected. However, the decrease of P_{sc} is more than compensated by the increase of P_{sw} with the net result that the total power increases with α. The monotonic decrease of the average best-case propagation delay with α is obvious in Fig. 11. This is expected as increasing α decreases the high-to-low propagation delay and does not affect the best-case low-to-high propagation delay. On the other hand, increasing α causes the internal capacitances to increase which in turn increases the worst-case low-to high propagation delay due to the need to charge the internal capacitances of the PDN. An optimum behavior for the worst-case average propagation delay is thus expected as shown in Fig. 11.

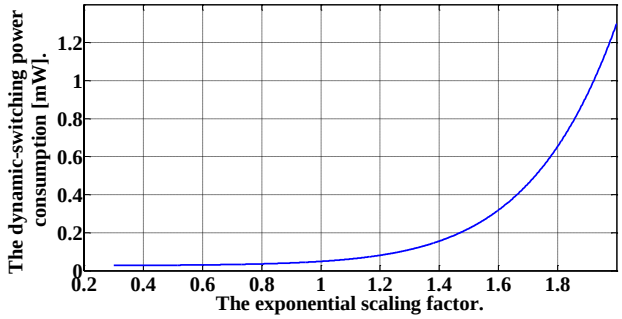

Fig. 8 The plot of the dynamic-switching power versus α.

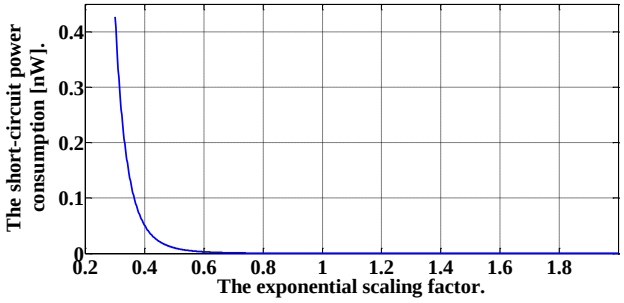

Fig. 9: The plot of the short-circuit power versus α.

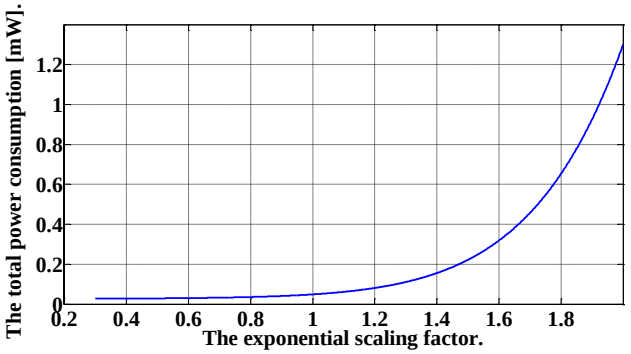

Fig. 10: The plot of the total power versus α.

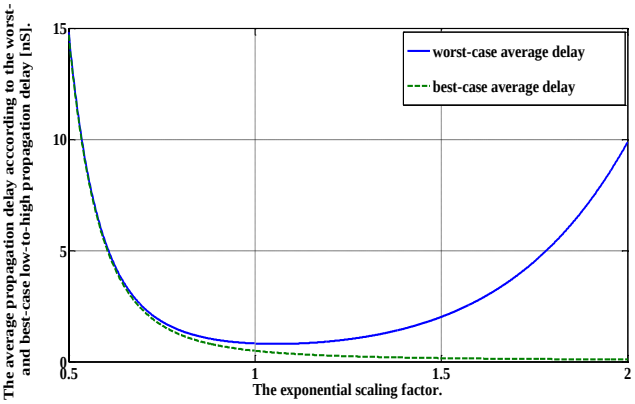

Fig. 11: The plots of the average worst- and best-case propagation delays versus α.

B. The Figure of Merit

Refer to Fig. 12 for the plot of the *FOM* (adopting the best-case low-to-high propagation delay) versus α. As expected, this curve displays an optimum behavior. Its optimum value occurs at α = 1.001. Had we adopted the worst-case low-to-high propagation delay, the optimum value of the *FOM* would have occurred at 0.9009 which is less than 1. This is not unexpected as there will be a need to charge the parasitic capacitances associated with the PDN as stated before. Increasing α leads to several contradicting effects. The first one is increasing the current-driving capability of the discharging transistors, thus speeding-up the discharging process. However, the increase in α increases the internal capacitances, thus hindering the rapid discharge along with the associated increase in the power consumption. Also, the area certainly increases with α.

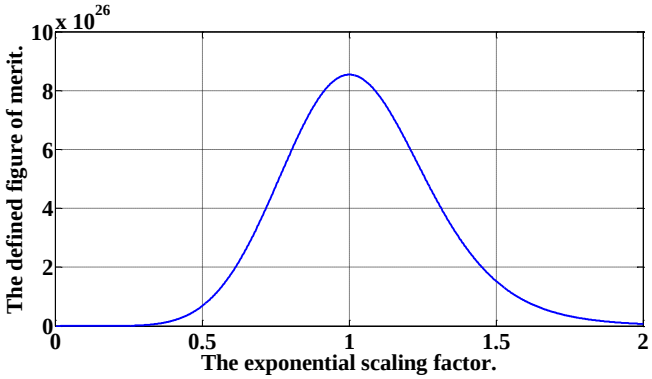

Fig. 12: The plot of the proposed *FOM* versus α.

In all the discussions that follow, the worst-case propagation delay will be adopted. When increasing the weight of the area in the *FOM*, a_1, it is expected that the optimum value of α, α_{opt}, will decrease monotonically because increasing the area of the PDN will be considered an area wasting. To confirm this, refer to the plot of α_{opt} versus a_1 in Fig. 13. The plots of α_{opt} versus a_2 and a_3 are shown also in this figure. As obvious, the decrease of α_{opt} with a_2 is larger than that with a_1. However, α_{opt} increases with increasing a_3 due to the emphasis of the delay importance with increasing a_3 compared with that of the area, thus the need arises to increase the size of the PDN in order to speed-up the operation.

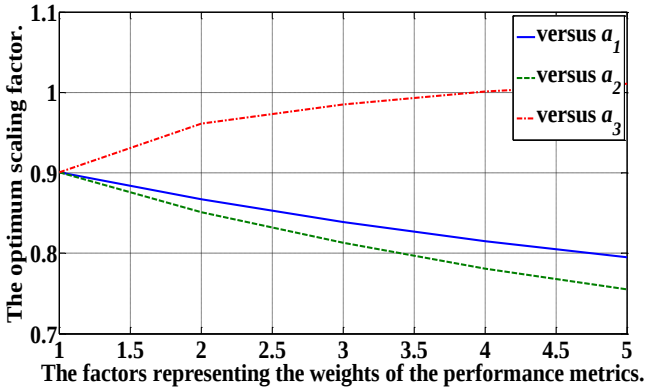

Fig. 13: The plots of α_{opt} versus a_1, a_2, and a_3.

The effects of f, n, and C_L on the *FOM* are shown in Figs. 14, 15, and 16. As expected, the three curves show a monotonic decrease. Increasing the frequency of operation causes the power consumption to increase; however, it has no effect on the area or the time delay. Increasing n causes the three performance metrics to degrade and thus a rapid decrease in the *FOM* compared with that due to increasing f or C_L. Finally, increasing C_L causes both the time delay and the power consumption to increase.

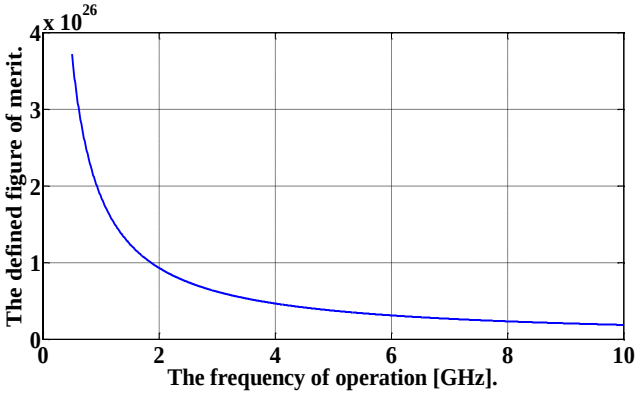

Fig. 14: The relationship between the frequency of operation and the defined *FOM*.

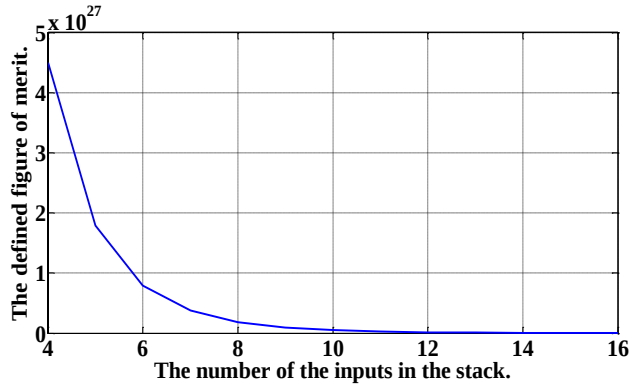

Fig. 15: The relationship between the number of the inputs in the stack and the defined *FOM*.

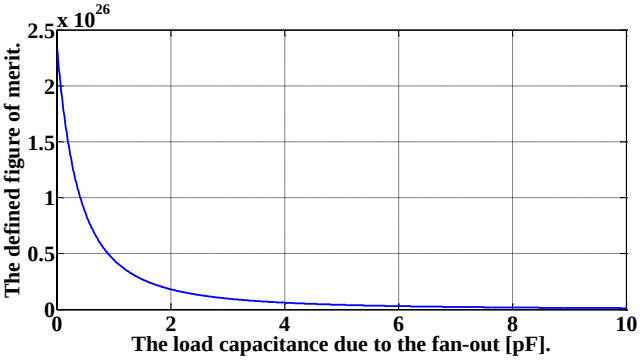

Fig. 16: The relationship between the load capacitance due to the fan-out and the defined *FOM*.

C. Effect of the Power-Supply Voltage

Fig. 17 shows the variation of the figure of merit with the power-supply voltage. The *FOM* exhibits a local maximum versus V_{DD}. This is certainly due to the historical well known tradeoff between the increase in the power consumption and the decrease in the time delay with the increase in the power-supply voltage.

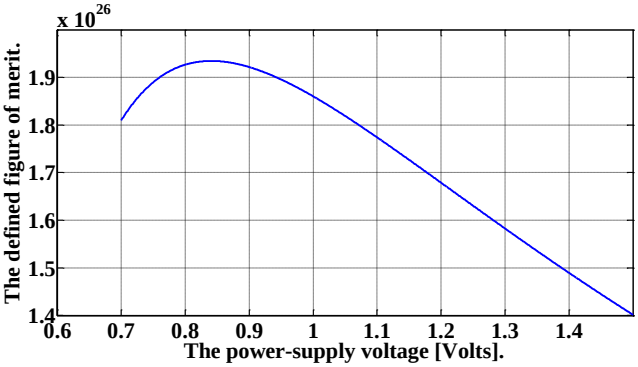

Fig. 17: The plot of the *FOM* versus the power-supply voltage.

D. Effect of the Load Capacitance

Refer to Fig. 18 for the plot of the optimum value of α, α_{opt}, versus the load capacitance, C_L. It is obvious that the optimum value of the scaling factor increases with increasing C_L. This is due to the fact that the increase of C_L is associated with an increase in both the time delay and the power consumption but not in the area as this capacitance is elated to the next stage and the wiring and interconnect capacitance. Thus, in order to obtain the best performance, it is required to increase the size of the PDN in order to speed-up the discharging process. At C_L = 87.50 fF, α_{opt} is equal to 1. For C_L larger than this value, α_{opt} will be larger than 1. When C_L is much larger than the parasitic capacitance at the output node, $(\alpha + n\beta)C$, it is preferred to use a value of α that is larger than 1 (that is, use tapering) in order to speed-up the discharging of C_L.

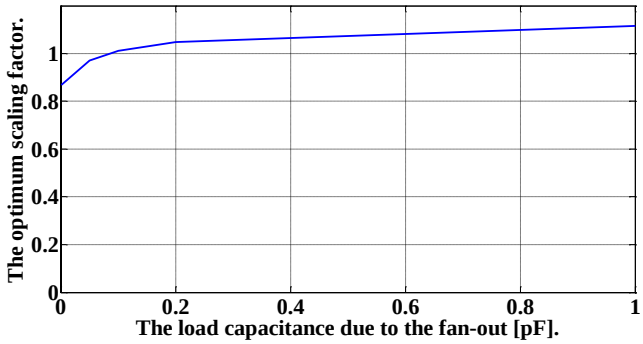

Fig. 18: The plot of α_{opt} versus C_L.

E. Effect of Technology Scaling

In this subsection, the effect of technology scaling on the performance of circuits containing NMOS and PMOS stacks will be discussed in light of the optimization process. Specifically, there are two important effects associated with short-channel MOSFET transistors that can affect the stack performance. The first one is the velocity saturation. The second one is the reduction of the body effect on the threshold voltage.

First, due to the velocity saturation, the dependence of I_D on V_{GS} will be weaker. So, the degradation in the discharging current due to stacking is thus expected to be less than that in the case of long-channel devices [31]. If the discussion is to be extended to PMOS stacks, then we must take into account the fact that the drift velocity of the free electrons saturates at an electric field of typically 3 V/µm compared to 10 V/µm for holes [32]. This implies that the effect of the velocity saturation is more perceptible in NMOS devices compared to that in PMOS devices. So, it can be expected that the degradation in the charging speed associated with PMOS stacks will be less than that in the discharging speed associated with NMOS stacks. Also, it is expected that the sizing constraint imposed on the PMOS transistors will be mitigated [32]. Thus, β is expected to decrease with technology scaling with the result that its loading on the output node is alleviated. The performance enhancement with the reduction of β is confirmed by the increase in the *FOM* as shown in Fig. 19.

Second, the body-effect changes the threshold voltage of the MOSFET transistor, thus affecting its current-driving capability. The threshold voltage will change with the source-to-substrate voltage, V_{SB}, according to the following familiar relationship [30]:

$$V_{thn} = V_{thn0} + \gamma\left(\sqrt{2\phi_f + V_{SB}} - \sqrt{2\phi_f}\right) \quad (39)$$

where V_{thn0} is the threshold voltage at V_{SB} = 0, $2\phi_f$ is a physical parameter related to the energy-band diagram, and γ is a fabrication-process parameter given by

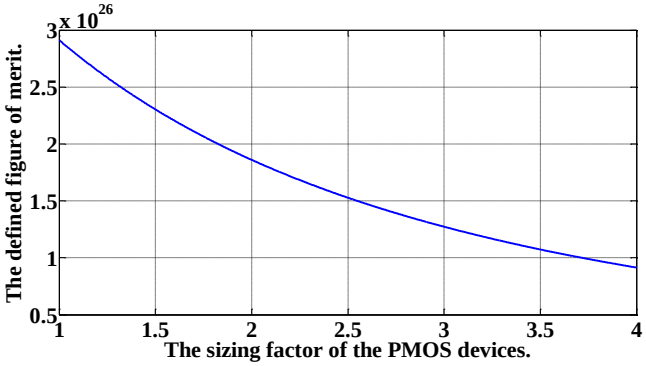

Fig. 19: The relationship between the sizing factor of the PMOS devices, β, and the defined *FOM*.

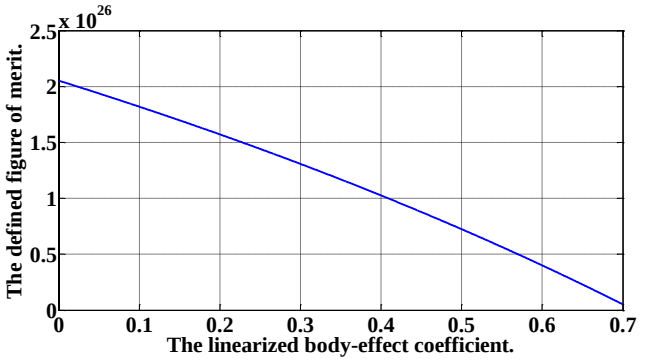

Fig. 20: The relationship between the linearized body-effect coefficient, *k*, and the defined *FOM*.

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_s}}{C_{ox}} \quad (40)$$

where q is the electronic charge, N_A is the doping concentration of the p-type substrate, ε_s is the electric permittivity of silicon ($1.04*10^{-12}$ F/cm), and C_{ox} is the gate-oxide capacitance per unit area. If the voltage, V_{SB}, is relatively small with respect to $4\phi_f$, then the following approximation can be used:

$$V_{thn} = V_{thn0} + kV_{SB},$$

where $k = \gamma/2\sqrt{2\phi_f}$. In order for the MOSFET transistor to operate properly in spite of CMOS technology scaling, the doping of the substrate, N_A, must be increased in order to reduce the thickness of the depletion regions associated with the source/drain and substrate junctions. However, the gate-oxide thickness, t_{ox}, must be decreased in order to increase C_{ox} and reduce the short-channel effects. The increase in C_{ox} is larger than that of N_A with the net result that the body-effect parameters, γ and k, decrease with technology scaling. So, it can be concluded that the degradation in speed due to stacking will be less with CMOS technology scaling. This is confirmed by the increase of the *FOM* with the decrease of *k* as shown in Fig. 20.

VI. Simulation Results

In this section, the resistance and capacitance models adopted in representing the PDN will be confirmed by comparison with the simulation results adopting the 45 nm CMOS technology. Refer to Figs. 21 and 22 for the plots of the low-to-high propagation delay of a single PMOS device and the high-to-low propagation delay, both versus C_L, according to the adopted models and to the simulation results using α = 1.

The simulation results showing the average propagation delay according to the best-case and the worst-case are presented in Fig. 23. As indicated, there is a minimum value for the worst-case average propagation delay while that according to the best case decreases monotonically with increasing the scaling factor; the same conclusion drawn in the previous section.

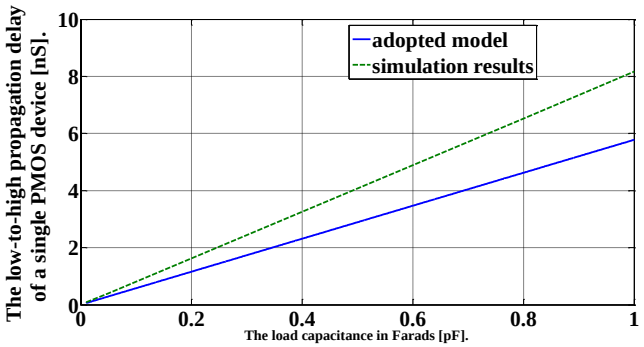

Fig. 21: The low-to-high propagation delay of a single PMOS device versus C_L according to the adopted model and the simulation.

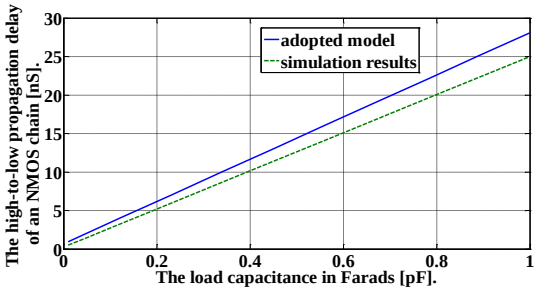

Fig. 22: The high-to-low propagation delay of an NMOS stack versus C_L according to the adopted model and the simulation.

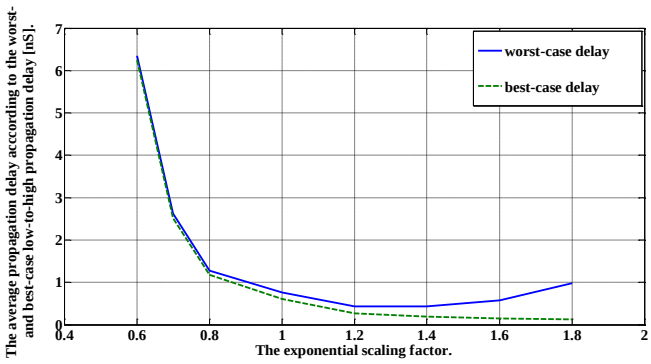

Fig. 23: The plots of the average worst- and best-case propagation delays versus α according to the simulation results.

VII. Conclusions

In this paper, the problem of the slow response and high power consumption of long stacks of MOS devices along with the optimization of these circuits were discussed. A proposed figure of merit that includes the average propagation delay, power consumption, and area was utilized to assess the performance of such circuits. The main point is that under the worst-case low-to-high transition at the output, the preferred sizing strategy is to decrease the sizing downward for the NMOS stacks. This is due to the reduction of the internal capacitances with the scaling factor. This is in contrast to the best-case scenario at which the sizing is preferred to be increased. This is due to that, in the best-case charging, all the NMOS devices are deactivated and thus their sizes do not affect the charging process. On the other hand, increasing their sizes speeds-up the discharging process, thus reducing the average propagation delay. Also, it was found that the performance degrades with the increase in the frequency of switching, the number of the inputs, and the load capacitance. The larger rate of degradation of the performance was found to be due to increasing the number of the inputs. The performance was evaluated for different weights of the time delay, the power consumption, and the area. Finally, the performance enhancement with CMOS technology scaling was confirmed.

VIII. Future Work

The analysis in this paper was performed adopting the exponential sizing strategy. Therefore, the scaling factor was only the adjustable variable and thus allowing the search for the optimal solution in only one-dimensional subspace of the whole variable space. However, this analysis can be repeated adopting the other sizing strategies or a combination of them with the results compared with those of this paper. By this way, more than one factor will be allowed to vary, thus allowing for more flexibility in searching for the optimal solution.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] K. Martin, *Digital Integrated Circuit Design*, Oxford University Press, 2000.

[2] S. M. Sharroush, Design Techniques for High Performance MOS Digital Integrated Circuits, Doctor of Philosophy Thesis, Port Said University, Egypt, 2011.

[3] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, Fourth Edition, Addison-Wesley, 2011.

[4] K. K. Kim, Y. B. Kim, M. Choi, and N. Park, "Leakage minimization technique for nanoscale CMOS VLSI based on macro-cell modeling," IEEE Transactions on Design & Test of Computers, Dec. 13, 2006.

[5] S. M. Sharroush, Y. S. Abdalla, A. A. Dessouki, and E. A. El-Badawy, "A novel technique for speeding up domino CMOS circuits containing a long chain of NMOS transistors," International Conference on Electronic Design (ICED), Penang, 1- 3 Dec., 2008.

[6] H. Mostafa, M. Anis, and M. Elmasry, "Novel timing yield improvement circuits for high-performance low-power wide fan-in dynamic OR gates," IEEE Transactions on Circuits and Systems I: Regular Papers, Vol. 58, Issue: 8, Pages: 1785 – 1797, Aug. 2011.

[7] S. M. Sharroush, "A novel high-speed CMOS circuit based on a gang of capacitors," International Journal of Electronics, 2017.

[8] S. M. Sharroush, "A novel high-performance time-balanced wide fan-in CMOS circuit," Alexandria Engineering Journal, Vol. 55, Issue: 3, Pages: 2565 – 2582, 2016.

[9] M. M. Khellah and M. I. Elmasry, "Use of charge sharing to reduce energy consumption in wide fan-in gates," IEEE International Symposium on Circuits and Systems (ISCAS), Monterey, CA, USA, 31 May - 3 Jun. 1998.

[10] M. Shoji and N. J. Warren, "Apparatus for increasing the speed of a circuit having a string of IGFETs," U. S. Patent: 4430583, Feb. 7, 1984.

[11] B. S. Cherkauer and E. G. Friedman, "The effects of channel width tapering on the power dissipation of serially connected MOSFETs," IEEE International Symposium on Circuits and Systems, Vol. 3, Pages: 2110 – 2113, Chicago, IL, 3 - 6 May 1993.

[12] R. H. Krambeck, C. M. Lee, and H. F. S. Law, "High-speed compact circuits with CMOS," IEEE Journal of Solid-State Circuits, Vol. SC – 17, Pages: 614 – 619, Jun. 1982.

[13] L. Ding and P. Mazumder, "On optimal tapering of FET chains in high-speed CMOS circuits," IEEE Transactions on Circuits and Systems – II: Analog and Digital Signal Processing, Vol. 48, No. 12, Dec. 2001.

[14] S. Choudhary and S. Qureshi, "Power aware channel width tapering of serially connected MOSFETs," International Conference on Microelectronics, Pages: 399 – 402, Cairo, 29 – 31 Dec. 2007.

[15] B. S. Cherkauer and E. G. Friedman, "Channel width tapering of serially connected MOSFET's with emphasis on power dissipation," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 2, Issue: 1, Pages: 100 – 114, Mar. 1994.

[16] J. Yuan and C. Svensson, "Principle of CMOS circuit power-delay optimization with transistor sizing," IEEE International Symposium on Circuits and Systems, Vol. 1, Pages: 637 – 640, Atlanta, GA, 12 - 15 May 1996.

[17] W. Kuzmicz, *Leakage Physics and Modeling Exercises*, Warsaw University of Technology.

[18] M. V. Dunga, X. Xi, J. He, W. Liu, K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, and C. Hu, *BSIM4.6.0 MOSFET Model: User's Manual*, University of California, Berkeley, 2006.

[19] P. K. Chan and K. Karplus, "Computing signal delay in general RC networks by tree/link partitioning," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 9 , Issue: 8, Pages: 898 – 902, Aug. 1990.

[20] T. Sakurai, Member and A. R. Newton, "Delay analysis of series-connected MOSFET circuits,", IEEE Journal of Solid-State Circuits, Vol. 26, Issue: 2, Pages: 122 – 131, Feb. 1991.

[21] S. S. Bizzan, G. A. Jullien and W. C. Miller, "Analytical approach to sizing nFET chains," Electronics Letters, Vol. 28, Issue: 14, Pages: 1334 – 1335, 2 Jul. 1992.

[22] S. A. Spanoche, "Long chain digital MOS transistor optimal design," Proceedings of the International Semiconductor Conference, Pages: 371 – 374, Sinaia, 11 - 14 Oct. 1995.

[23] W. C. Elmore, "The transient response of damped linear networks with particular regard to wideband amplifiers," Journal of Applied Physics, Vol. 19, No. 1, Pages: 55 – 63, Jan. 1948.

[24] J. E. Ayers, *Digital Integrated Circuits: Analysis and Design*, CRC Press, 2005.

[25] J. K. Oueterhout, "Switch-level delay models for digital MOS VLSI," 21st Conference on Design Automation, Pages: 542 – 548, 25-27 Jun. 1984.

[26] J. P. Uyemura, *Chip Design for Submicron VLSI: CMOS Layout and Simulation*, Thomson, First Edition, 2006.

[27] M. Shoji, "FFT scaling in domino CMOS gates," IEEE Journal of Solid-State Circuits, Vol. 20, Issue: 5, Pages: 1067 – 1071, Oct. 1985.

[28] H. Shichman and D. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuit," IEEE Journal of Solid-State Circuits, Vol. sc-13, No. 3, Pages: 285 - 289, Sep. 1968.

[29] D. Zwillinger, *Table of Integrals, Series, and Products*, Academic Press, Eighth Edition, 2014.

[30] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, Seventh Edition, New York: Oxford, 2015.

[31] D. A. Hodges, H. G. Jackson, and R. A. Saleh, *Analysis and Design of Digital Integrated Circuits: in Deep Submicron Technology*, McGraw-Hill, Third Edition, 2003.

[32] M. W. Allam, New Methodologies for Low-Power High-Performance Digital VLSI Design, Doctor of Philosophy Thesis, Waterloo, Ontario, Canada, 1999.