Port Said Engineering Research Journal
Port Said University – Faculty of Engineering
Volume 16 No. 1 pp: 165:175

Speeding-up MOS Circuits Containing Stacks

Sherif M. Sharroush[1], , Ahmed A. Dessouki[2], Yasser S. Abdalla[3], and El-Sayed A. El-Badawy[4]

Abstract

MOS circuits such as NAND and NOR gates may contain stacks of NMOS and PMOS transistors especially with wide fan-in. The main problem associated with these stacks is the relatively slow response due to the relatively large RC time constant associated with charging or discharging the parasitic capacitances at the output node as well as the internal nodes of the circuit. In this paper, a proposed technique will be presented in order to reduce the time delay of these circuits. The proposed technique is analyzed quantitatively and a compact form for the percentage reduction in the discharging time delay is derived and the optimum configuration for the proposed circuit is decided on. Simulation results adopting the 0.13 µm CMOS technology reveals that about 40% of the time delay can be saved.

Key Words: MOS circuits, stack, time delay.

I. INTRODUCTION

MOS circuits such as NAND and NOR gates may contain stacks of NMOS and PMOS transistors especially with wide fan-in. The main problem associated with these stacks is the relatively slow response due to the relatively large *RC* time constant associated with charging or discharging the parasitic capacitances at the output node as well as the internal nodes of the circuit. Applications that may include a long chain of series-connected transistors are such as multi-input exclusive-*OR* gates that are required in applications such as parity-check and error-correction circuits or some built-in testing circuits. Another example is barrel-shifters [1]. In this paper, the problem of the slow response of the stacks will be addressed in detail. Previous solutions for alleviating this problem are presented. The rationale behind the proposed technique is to generate a voltage proportional to the number of activated inputs. This voltage is then used to activate a discharging transistor that is activated only when all the *n* inputs are activated. The proposed technique will be analyzed quantitatively and a compact form for the percentage reduction in the discharging time delay will be derived and the optimum configuration of the proposed circuit from the point of view of robustness to process variations will be decided on.

[1]*Dept of Elect Eng, Fac. of Eng., Port Said, Suez Canal Univ., Egypt. EM: Sherif_sharroush2003@yahoo.com*
[2]*Dept of Elect Eng, Fac. of Eng., Port Said, Suez Canal Univ., Egypt. EM: dessouki2000@yahoo.com*
[3]*Dept of Electricity, Fac. of Industrial Edu., Suez, Suez Canal Univ., Egypt. EM: yasser@alumni.uwaterloo.ca*
[4]*Alex Higher Inst. Of Eng. and Tech & Fac. of Eng., Alex. Univ., Alexandria, Egypt. EM: sbadawy@ieee.org*

The remainder of the paper is organized as follows: Section II presents the problem in detail with the impact of technology scaling on the circuits containing stacks presented in Section III. The previous solutions of this problem will be presented in Section IV while the proposed solution will be presented in Section V. The proposed solution will be analyzed quantitatively in Section VI in which a compact form for the percentage reduction in the time delay will be derived and the optimum configuration of the proposed circuit will be decided on. The proposed scheme will be simulated for the 0.13 µm CMOS technology with the simulation results presented in Section VII. Finally, the paper will be concluded in Section VIII.

II. Problem Statement

The problem associated with MOS circuits containing MOSFET transistor stacks will be discussed in detail. Refer to Fig. 1 for a NAND gate with 8 inputs implemented in the static complementary CMOS logic circuit family. Assume that the parasitic capacitance at the output node is charged to a voltage of V_{DD} through any one or a combination of the PMOS transistors of the pull-up network. Now, if all the inputs, A_1, A_2,……, A_7, and A_8, are at logic "1", then all the NMOS transistors in the stack will be activated. The parasitic capacitance, C_L, will thus be discharged through the stack.

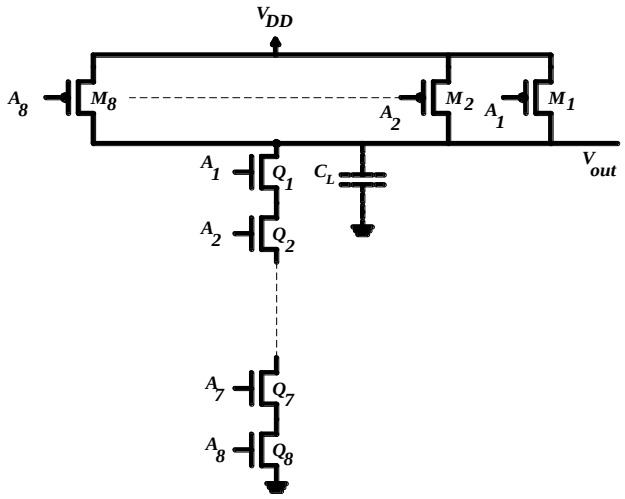

Fig. 1 A typical NAND gate with 8 inputs implemented in static complementary CMOS logic.

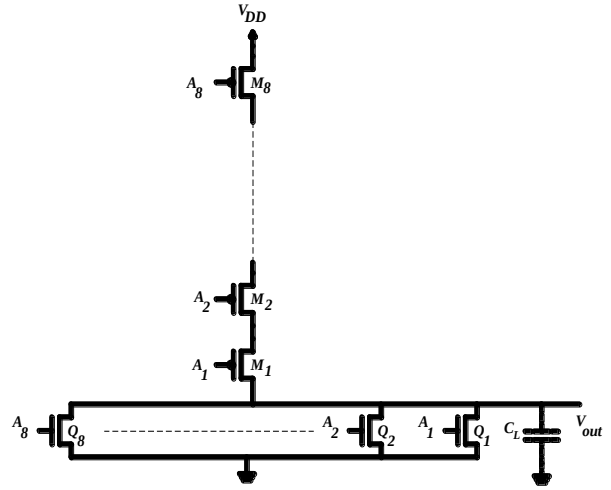

Fig. 2 A typical NOR gate with 8 inputs implemented in static complementary CMOS logic.

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

1. Simulation results show that at the beginning of the discharging process, all the NMOS transistors in the stack will operate in the triode region except the uppermost one which operates in the saturation region as its gate and drain voltages will initially be at the same value, V_{DD}. This can be explained by the reason that the voltage across the parasitic capacitance, V_{CL}, will divide across the NMOS transistors in the stack resulting in a relatively small value for the drain-to-source voltage for each transistor in the stack. The drain-to-source voltage, V_{DS}, of each transistor will thus be lower than the gate-overdrive voltage, V_{GS} - V_{thn}, of each transistor except for the uppermost one where V_{thn} is the threshold voltage of the NMOS transistor. Since the dependence of the drain current on the gate voltage is weaker in the triode region than that in the saturation region, the discharging current is expected to be relatively small. Also, due to the division of V_{CL} across the NMOS transistors in the stack, the V_{DS} voltage will be relatively small across each transistor, thus reducing the discharging current further.

2. 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. In fact, one can imagine the stack of the NMOS transistors as *n* resistors connected with each other in series. The larger the number of transistors in the stack, the larger the total resistance will be with the result that the discharging current decreases.

3. 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 [2] 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 gate-to-source voltage [3]. Thus, the subthreshold-leakage current decreases significantly through the stack.

4. The parasitic capacitances at the internal nodes of the stack need to be discharged.

The above discussion applies equally well to the stack of PMOS transistors shown in the NOR gate in Fig. 2. 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, NAND gates are more widely used in implementing logic circuits than NOR gates [2].

1. Stacking NMOS Transistors in the Pseudo NMOS Logic

Due to its smaller area, lower parasitic capacitances, thus higher speed and lower dynamic-power consumption, pseudo NMOS logic can be used in which an always activated PMOS transistor is used as a load instead of the pull-up network. Refer to Fig. 3 for a pseudo NMOS logic representing an eight-input NAND gate.

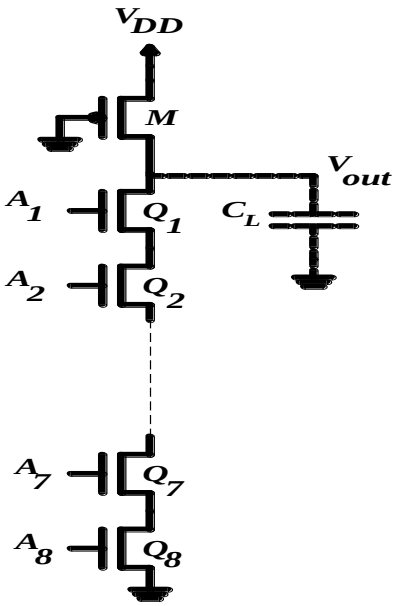

Fig. 3 An eight-input NAND gate implemented in the pseudo NMOS logic circuit family.

The price paid, however, is the static power consumption of the circuit in case of logic "0" at the output node. Also, the noise margin in this case will be smaller than that of the complementary CMOS logic circuit due to the nonzero value of the voltage on the output node. This will worsen if the pull-down network is a stack of NMOS transistors. Properly sizing the NMOS transistors of the stack is necessary in order to overcome the contention current supplied by the always activated PMOS transistor, thus pulling down the voltage at this node. This puts a lower bound on the aspect ratios of the transistors in the stack, thus losing one of the main advantages of the pseudo NMOS logic family compared to static complementary CMOS logic family. This may suggest that using pseudo NMOS logic family with these configurations for the pull-down network is not suitable.

Figure 4 shows the relationship between the discharging time delay estimated between the 90% and 10% points of the output voltage waveform and the number of series-connected NMOS transistors in the stack for static complementary CMOS logic. This curve was based on simulation results on the 0.13 µm CMOS technology with the parasitic capacitance initially charged at V_{DD} = 1.2 V.

2 Stacking NMOS Transistors in the CMOS Domino Logic

The problem of contention current from the PMOS keeper is more tactile for the case when there is a large number of serially connected NMOS transistors in the PDN. In this case, the discharging of C_L will be very slow due to the stack effect. In the following, we will discuss this problem in detail.

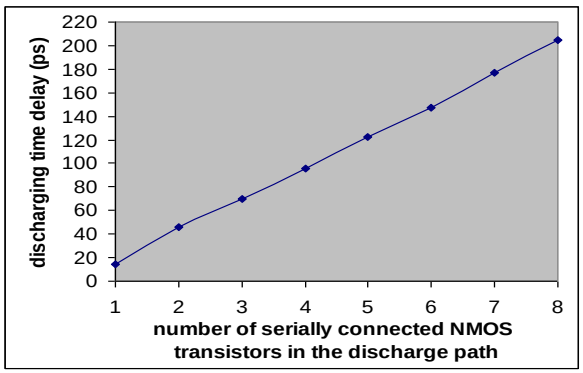

Fig. 4 The relationship between the number of serially connected NMOS transistors and the discharging time delay in the static complementary CMOS logic for the 0.13 µm CMOS technology.

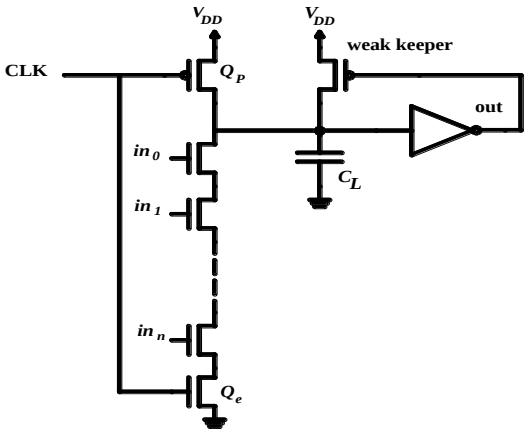

Fig. 5 Domino logic circuit schematic when there is a long chain of NMOS transistors connected in series in the PDN. The discharging process of C_L through this long chain is relatively slow.

Consider the case when there is a large number of inputs (wide fan-in) in the PDN whose corresponding NMOS transistors are connected in series. Refer to Fig. 5 for illustration. If all the inputs are activated during the evaluation phase, the discharging process of the dynamic-node capacitance, C_L, will be very slow. This is typically the case with AND gates with a large number of inputs. The discharging current in case of a large number of NMOS transistors (exceeding 8 or more) may be insufficient to compensate for the current provided by the weak PMOS keeper and thus results in either a very sluggish operation for the circuit or an erroneous output.

Of course, in order to avoid these drawbacks, the aspect ratio of the NMOS transistors of the chain of the PDN must be increased, thus increasing the required area. Also, the internal capacitances of these NMOS transistors will be increased, thus causing the dynamic-node charge to discharge further through charge sharing and degrading the noise immunity.

III. Impact of Technology Scaling

In this section, the effect of technology scaling on the performance of circuits containing NMOS and PMOS stacks will be discussed. 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 and the associated degradation in mobility. The second one is the reduction of the body effect on the threshold voltage.

First, due to the velocity saturation and mobility degradation, 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 [4].

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 [2]:

$$V_{thn} = V_{thn0} + \gamma\left(\sqrt{2\phi_f + V_{SB}} - \sqrt{2\phi_f}\right) \quad (1)$$

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

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

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.

It is shown in [5] that in order for the MOSFET transistor device to operate properly in spite of CMOS technology scaling, the doping of the substrate, N_A, must be increased. However, the gate-oxide thickness, t_{ox}, decreases in order to reduce short-channel effects [6] with the result that the gate-oxide capacitance per unit area, C_{ox}, increases. The increase in C_{ox} is larger than that of N_A with the net result that the body-effect parameter, γ, decreases with technology scaling. So, it can be concluded that the degradation in speed due to stacking will be less with CMOS technology scaling.

If the discussion is to be extended to PMOS stacks, then one 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 [7]. This implies that the degradation of the mobility of free electrons will be larger than that of holes. So, it can be expected that the degradation in 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 [7].

In a nutshell, the problem can be stated as follows: If all the inputs are activated for the NMOS stack, how can the parasitic capacitance at the output node be discharged rapidly?

IV. Previous Solutions

In this section, some of the previously proposed schemes to speed-up the response of circuits containing MOSFET transistor stacks will be discussed.

The first solution was to properly size the NMOS transistors in the stack in order to obtain the maximum speed for a certain area [8]. This sizing process was to make the lowermost transistor the largest one with the aspect ratios of the upper NMOS transistors decreasing as we move away from bottom to top as shown in Fig. 6. The rationale behind this sizing strategy is that the values of the parasitic capacitances at the internal nodes increase with increasing the size of the connected transistors [8]. Also, these internal capacitances will be initially charged at the beginning of the discharging process. The parasitic capacitance at the internal node, D_8, will be discharged through Q_8 while the parasitic capacitance at the internal node, D_7, will be discharged through Q_7 and Q_8 and so on until one arrives at the output capacitance, C_L, which discharges through the whole stack. So, the lowermost transistor, Q_8, must be the largest one as it contributes in discharging all the internal capacitances with sizes decreasing upward. Several sizing schemes include the linear, exponential, or a combination of the two. The linear sizing scheme is the one in which the ratio between two successive MOS transistors is the same through the stack. The exponential sizing scheme is the one in which the ratio between two successive MOS transistors is not the same through the stack, instead this ratio will increase upward. It is shown in [9] that the optimum speed for a certain area can be obtained by combining the linear and exponential sizing schemes.

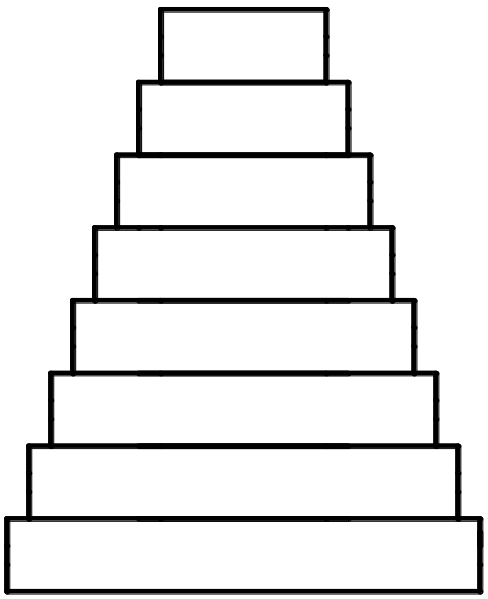

Fig..6 A simplified stick diagram for the eight-transistor layout of the pull-down network.

The second solution is decomposing the NAND gate into two- or three-input AND gates with an inverter in cascade as shown in Fig. 7. The time delay associated with the propagation of signals from one stage to the next must be taken into account.

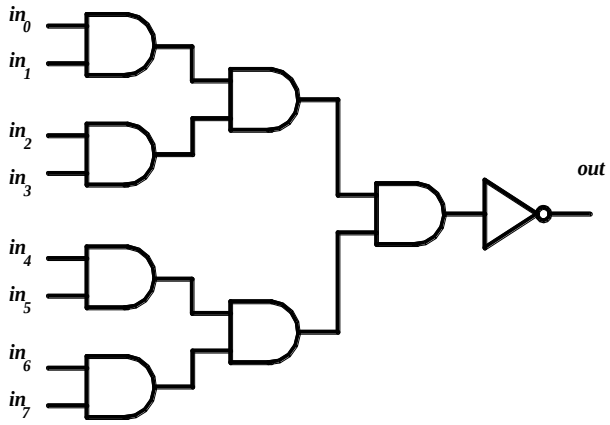

Fig. 7 The NAND gate with 8 inputs can be made with the shown cascaded AND gates with 2 inputs for each followed by an inverter.

3. The third solution is to forward bias the body-to-source junction of all the transistors in the stack in order to reduce the threshold voltage of the transistors. This can be done by connecting the body and gate terminals of all the transistors together [10]. The main concern here is not to forward bias the body-to-source junction into deep conduction in order to avoid latchup [10].

V. Proposed Solution

Circuit Evolution

In this section, the proposed solution to the problem of relatively slow response associated with stacks is discussed. Specifically, assume that it is required to speed-up the discharging process of a parasitic capacitance, C_L, for the case of n series-activated NMOS transistors, where the capacitance, C_L, is initially charged to V_{DD}. The rationale behind this technique is to generate a voltage proportional to the number of activated inputs. This voltage is then used to activate a discharging transistor that is activated only when all the n inputs are activated. Let this voltage be V_n. If the number of activated inputs is n - 1 or less, the generated output voltage, V_{n-1}, must be smaller than the threshold voltage of the discharging transistor, thus keeping the parasitic capacitance at the output node charged at V_{DD} as it must remain. The circuit shown in Fig. 8 can theoretically do the job assuming that the number of inputs, n = 8. The word theoretically will be emphasized later.

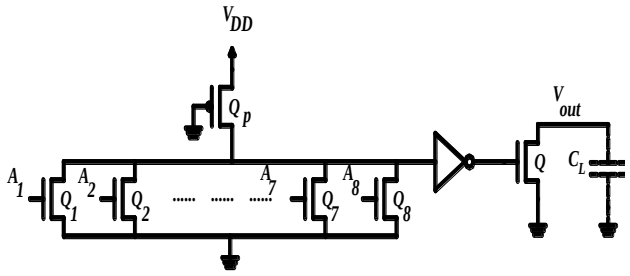

Fig. 8 The proposed scheme acting to speed-up the discharging process at the output node. The capacitance, C_L , is initially charged to V_{DD} some way or another.

In this case, the eight transistors will instead be connected in parallel rather than in series as shown in the figure. Refer to Fig. 9 for the static input-output characteristics of the inverter. The inverter and the NMOS transistor, Q, are designed such that if seven inputs or less are activated simultaneously, then the inverter output voltage will be smaller than the threshold voltage of the NMOS transistor, V_{thn}.

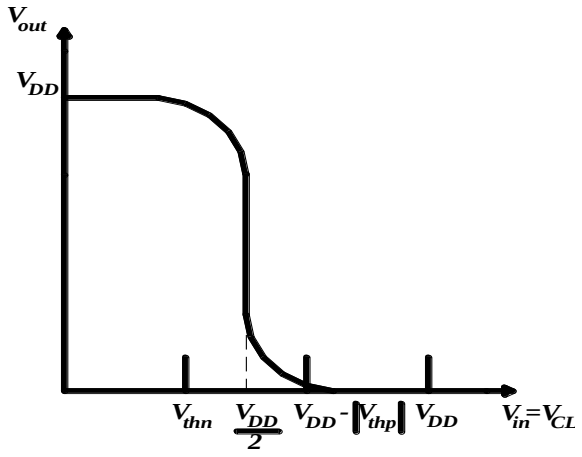

Fig. 9 The static input-output characteristics of the ideal CMOS inverter.

So, it will not conduct as it must do. On the other hand, if all the eight inputs are activated, then the inverter output voltage will be larger than the threshold voltage of the NMOS transistor, Q, such that it will conduct, thus discharging the parasitic capacitance at the output node as it must do. Refer to Fig. 10 for the relationship between the number of activated inputs, n, and the output voltage of the inverter. As the number of activated inputs increases, the current through the parallel connection will increase, thus decreasing the voltage at the input of the inverter and increasing the inverter output voltage accordingly. If the two voltages, V_{n-1} and V_n, are chosen to lie in the vicinity of the knee of the inverter characteristics, the voltage difference V_n - V_{n-1} will be as large as possible.

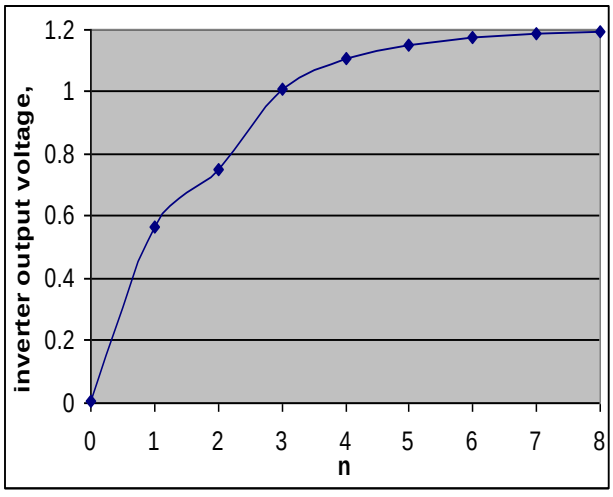

Fig. 10 The relationship between the number of activated inputs, n, and the output voltage of the inverter, V_{out}.

Now, it is obvious from the operation of the proposed scheme that the threshold voltage of the NMOS transistor, Q, must be chosen to be larger than V_{n-1} and smaller than V_n. This can be achieved by shifting the curve of the

static input-output characteristics of the inverter (shown in Fig. 9) to the left. This will make the range which we can choose V_{thn} of the discharging transistor, Q, to lie within to be as large as possible, thus making the circuit more robust to process variations.

Note the relatively large subthreshold-leakage current associated with the use of n parallel-connected NMOS transistors compared to the stack. However, this poses no problem as the PMOS transistor, Q_p, will compensate for this leakage current.

Changing the threshold voltage of the inverter, V_{thinv}, can be done by varying the aspect ratio of the NMOS or PMOS transistors of the inverter. However, simulation results show that the voltage difference, V_8 - V_7, increases very slightly with the increase in $(W/L)_n$ of the inverter. Another method is to weaken the PMOS transistor of the inverter. This can be done either by reducing its aspect ratio $(W/L)_p$ or increasing its threshold voltage $|V_{thp}|$ by ion implantation or by reverse-body biasing. The first solution is not possible, however, if the fabricated transistors are minimum-sized. Another idea to weaken the PMOS transistor is to replace the PMOS transistor by a number of series-connected PMOS transistors. It can be proved that the aspect ratio of the equivalent MOS transistor of n MOS transistors connected in series with an aspect ratio of (W/L) for each is $(1/n)(W/L)$ [1].

However, the voltage difference between that generated in case of 8 activated inputs and 7 activated inputs is so small that the circuit may malfunction. Specifically, due to process variations, the threshold voltage of the discharging transistor will change within a certain range that is determined by the process technology. So, a modification like that shown in Fig. 11 is required in order to enlarge the voltage difference between V_8 and V_7 where the circuit is shown for the case of eight inputs.

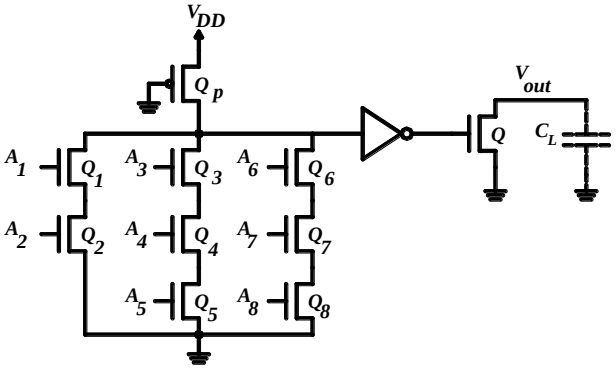

Fig. 11 The circuit shown in Fig. 8 after modification. The capacitance, C_L, is initially charged to V_{DD}.

The circuit of Fig. 11 operates as follows: If none of the inputs is activated, then the parasitic capacitance at the inverter input will be maintained charged at V_{DD} by virtue of the PMOS transistor, Q_P. The inverter output voltage will thus be at 0 V and the NMOS transistor, Q, will be deactivated, thus maintaining the charge of C_L. If only one or two paths of the discharging paths of Fig. 11 are activated, then the parasitic capacitance at the input of the inverter will be discharged. The PMOS transistor, Q_P, and the parallel conducting paths will form a voltage divider. The threshold voltage of the inverter, V_{thinv}, is adjusted such that the input voltage of the inverter for this case is larger than V_{thinv} with the result that the inverter output voltage will be at almost 0 V and the NMOS transistor, Q, will be deactivated. The output capacitance, C_L, will thus be maintained charged at V_{DD}.

Now, if all the inputs are activated, then there will be three discharging paths for the parasitic capacitance at the inverter input instead of two or one. The threshold voltage of the inverter, V_{thinv}, is adjusted such that the inverter input voltage for this case will be less than V_{thinv} with the result that the inverter output voltage will be at almost V_{DD} and the NMOS transistor, Q, will be activated. The output capacitance, C_L, will thus discharge as it must do.

The previous technique can also be applied to the stack of PMOS transistors. Assume that there is a stack of 8 PMOS transistors as shown in Fig. 2 that acts to charge the parasitic capacitance, C_L, in case all the inputs are at logic "0". In fact, the charging process will even be slower than the case of NMOS transistor stack because of the relatively small mobility of holes compared to that of free electrons [2], thus forcing the designer to choose a larger value for the aspect ratios of the PMOS transistors in the stack. The circuit shown in Fig. 12 acts to speed-up the charging process of C_L in case all the inputs are at logic "0".

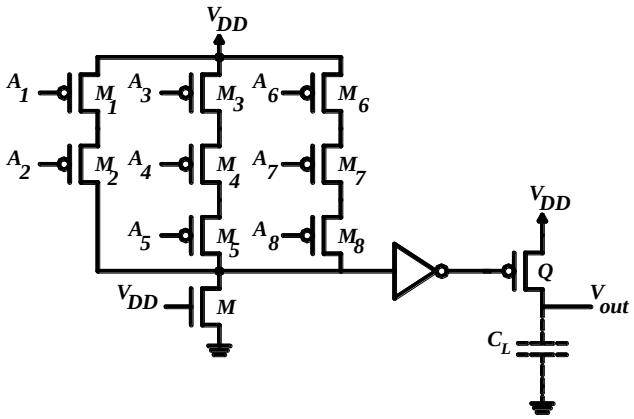

Fig. 12 The proposed circuit for speeding-up the charging process of C_L instead of charging it through the stack of PMOS transistors.

The circuit of Fig. 12 operates in a manner analogous to that of Fig. 11. Specifically, if all the inputs are at logic "1", then the parasitic capacitance at the inverter input will be maintained discharged at 0 V by virtue of the NMOS transistor, Q. The inverter output voltage will thus be at V_{DD} and the PMOS transistor, M, will be deactivated, thus maintaining C_L discharged. If only one or two paths of the charging paths of Fig. 12 are activated, then the parasitic capacitance at the input of the inverter will be charged. The NMOS transistor, M, and the parallel conducting paths will form a voltage divider. The threshold voltage of the inverter, V_{thinv}, is adjusted such that the input voltage of the inverter for this case is smaller than V_{thinv} with the result that the inverter output

voltage will be at almost V_{DD} and the PMOS transistor, *Q*, will be deactivated. The output capacitance, C_L, will thus be maintained discharged at 0 V.

Now, if all the inputs are activated, then there will be three charging paths for the parasitic capacitance at the inverter input instead of two or one. The threshold voltage of the inverter, V_{thinv}, is adjusted such that the inverter input voltage for this case will be larger than V_{thinv} with the result that the inverter output voltage will be at almost 0 V and the PMOS transistor, *Q*, will be activated. The output capacitance, C_L, will thus charge as it must do.

VI. Quantitative Analysis

We will in this section analyze the proposed scheme quantitatively from two aspects; the first one is finding the percentage reduction in the discharging time delay with using the proposed scheme. The second aspect is to decide on the configuration of the NMOS transistors in the pull-down network of the proposed scheme so as to obtain the best robustness with respect to process variations, i.e. to obtain the widest range within which we can choose the value of the threshold voltage of the inverter.

1. Percentage Reduction in the Discharging Time Delay
1.1 Estimating the Discharging Time Delay According to the Conventional Method

Refer to Fig. 1 for the conventional 8-input NAND gate where the parasitic capacitance, C_L, is to discharge through the stack of 8-NMOS transistors. This analysis was already performed in [11] for a number of transistors larger than that in the current case by one. So,

$$t_{dc} = t_{d1} + t_{d2} = \frac{C_L\left[V_{thn} + \alpha(V_{DD} - V_{thn})\right]\left[1 + WC_{ox}R_{total}(1+K)v_{sat}\right]}{WC_{ox}(V_{DD} - V_{thn0})v_{sat}} + 2.3(n)RC_L \quad (3)$$

where *R* is the resistance of each of the NMOS transistors and

$$R_{total} = \frac{(n-1)}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}.$$

1.2 Estimating the Discharging Time Delay According to the Proposed Method

The time delay according to the proposed scheme consists of three components; the first one is the time delay associated with discharging the capacitance at the inverter input through the series-parallel connected NMOS transistors, td1p, the second part is the time delay associated with the low-to-high propagation delay of the inverter, t_{d2p}, and the third and last one is the time delay associated with discharging the parasitic capacitance at the output node, C_L, through the activated NMOS transistor. To evaluate the first component, t_{d1p}, we will divide this time delay into two parts, t_{d11p} and t_{d12p}. The first component, t_{d11p}, corresponds to the time interval during which the topmost transistor in each leg of the pull-down network operates in saturation. Toward that end, we must first estimate the current through each leg, then multiply this by 3 to obtain the total discharging current during this time interval. The three legs, however, are not identical and the current through the leg with two NMOS transistors will be larger than that with the legs with three NMOS transistors. Thus this analysis, although simple, overestimates the time delay according to the proposed scheme.

The current, *i*, through each leg is [11]:

$$i = \frac{WC_{ox}(V_{DD} - V_{thno})v_{sat}}{\left[1 + \frac{2WC_{ox}(1+K)v_{sat}}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}\right]} \quad (4)$$

The total discharging current through the pull-down network will thus be

$$3i = \frac{3WC_{ox}(V_{DD} - V_{thno})v_{sat}}{\left[1 + \frac{2WC_{ox}(1+K)v_{sat}}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}\right]}. \quad (5)$$

The discharging time delay, t_{d11p}, will be

$$t_{d11p} = \frac{C_{in}\Delta V_{in}}{3i} \quad (6)$$

where ΔV_{in} is the range of the inverter input voltage during which the topmost transistor in each leg operates in the saturation region where

$$\Delta V_{in} = V_{thn} + \alpha(V_{DD} - V_{thn}). \quad (7)$$

So,

$$t_{d11p} = \frac{C_{in}\left[V_{thn} + \alpha(V_{DD} - V_{thn})\right]}{\left[1 + \frac{2WC_{ox}(1+K)v_{sat}}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}\right]}$$

$$= \frac{C_{in}\left[V_{thn} + \alpha(V_{DD} - V_{thn})\right]\left[1 + \frac{2WC_{ox}(1+K)v_{sat}}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}\right]}{3WC_{ox}(V_{DD} - V_{thno})v_{sat}} \quad (8)$$

To find the time delay component, t_{d12p}, we will substitute the discharging time constant, τ, by $\frac{1}{3}R_{leg}C_L$ as the resistor combination of the PDN will be approximately $\frac{1}{3}R_{leg}$ during this time interval, where R_{leg} is the equivalent resistance of each leg of the pull-down network. Evaluating the time delay, t_{d12p}, to the instant of time at which the voltage reaches 0.1 of its initial value

results in

$$2.3\tau = 2.3\frac{1}{3}R_{leg}C_{in} = 2.3\frac{1}{3}3RC_{in} = 2.3RC_{in} \quad (9)$$

The first component of the discharging time delay of the proposed method, t_{d1p}, will thus be

$$t_{d1p} = t_{d11p} + t_{d12p}$$
$$= \frac{C_{in}\left[V_{thn} + \alpha\left(V_{DD} - V_{thn}\right)\right]\left[1 + \frac{2WC_{ox}(1+K)v_{sat}}{k_n'\left(\frac{W}{L}\right)\left(\frac{V_{DD}}{2} - V_{thn}\right)}\right]}{3WC_{ox}\left(V_{DD} - V_{thno}\right)v_{sat}} + 2.3RC_{in} \quad (10)$$

The second component of the discharging time delay is the low-to-high propagation delay of the inverter which is [2]

$$t_{d2p} = \frac{2C_{out}}{k_p'\left(\frac{W}{L}\right)_p\left(V_{DD} - V_{thn}\right)}\left[\frac{V_{thn}}{V_{DD} - V_{thn}} + \frac{1}{2}\ln\left(\frac{3V_{DD} - V_{thn}}{V_{DD}}\right)\right] \quad (11)$$

where C_{out} is the parasitic capacitance at the inverter output. An approximate method for evaluating the parasitic capacitances at the inverter input and output terminal can be found in [12] in which the internal capacitances of the NMOS and PMOS transistors of the preceding and subsequent stages and the wiring capacitance are taken into account.

The last component of the time delay, t_{d3p}, is the high-to-low propagation delay associated with the discharging of C_L through the activated transistor, Q. This is equal to [2]

$$t_{d3p} = \frac{2C_L}{k_n'\left(\frac{W}{L}\right)_n\left(V_{DD} - V_{thn}\right)}\left[\frac{V_{thn}}{V_{DD} - V_{thn}} + \frac{1}{2}\ln\left(\frac{3V_{DD} - V_{thn}}{V_{DD}}\right)\right] \quad (12)$$

where C_L is the parasitic capacitance at the output node. The total time delay according to the proposed method is

$$t_{dp} = t_{d1p} + t_{d2p} + t_{d3p}. \quad (13)$$

It must be kept in mind that this evaluation for t_{dp} is overestimated as there is some overlap between the discharging of C_{in} through the pull-down network and the charging of C_{out} through the PMOS transistor of the inverter. This is because the charging process begins when the voltage across C_{in} reduces to V_{DD} - $|V_{thp}|$. Also, there will be some overlap between the charging of C_{out} and the discharging of C_L. This is because the discharging of C_L begins when the voltage across C_{out} rises to V_{thn} with the activation of the discharging transistor, Q. The percentage reduction in the discharging time delay can be found from

$$\%reduction = \frac{t_{dc} - t_{dp}}{t_{dc}} x100\% \quad (14)$$

after substituting n = 8 in Eq. (3). Note that the time delay associated with charging the capacitance at the inverter input, C_{in}, was not taken into account. This is because it is equivalent to the time required to charge the parasitic capacitance, C_L, by the pull-down network in the conventional circuit.

2. Optimum Configuration of the Pull-Down Network of the Proposed Scheme

Toward that end, assume that the pull-down network of the NMOS transistors according to the proposed scheme is as shown in Fig. 11. Assume for simplicity that each transistor of these ones can be represented by a resistor. Although this assumption is not valid for all the transistors as not all the transistors operate in the deep triode region for the whole discharging time interval, we will adopt this assumption as our main target is to decide on the optimum configuration of the PDN. Assuming that all these transistors are identical in size, then each one of these transistors can be represented by the same resistance value, R. Also, the always activated PMOS transistor, M_P, can be replaced by a suitable resistor. Assuming that all the NMOS transistors in the stack as well as the PMOS transistor, M_P, are minimum-sized and assuming that the mobility of the free electrons is 2.5 times that of the holes [2], so the resistance value representing M_P will be approximately 2.5R. The equivalent circuit will thus be as shown in Fig. 13.

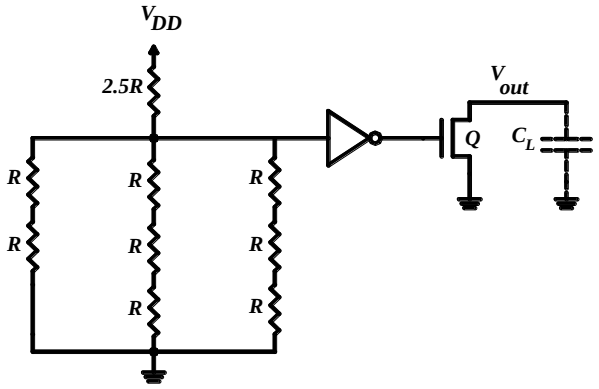

Fig. 13 The equivalent circuit of Fig. 11 after representing each transistor with a suitable resistance.

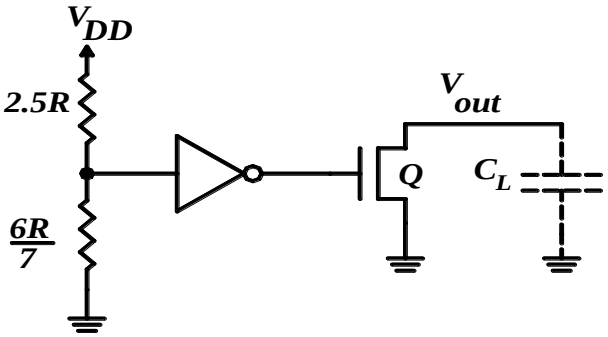

Fig. 14 The circuit of Fig. 13 after combining the resistors representing the NMOS transistors in the pull-down network.

Simply combining the values of the resistances in the pull-down network, we obtain

$$\frac{1}{R_{total}} = \frac{1}{2R} + \frac{1}{3R} + \frac{1}{3R} = \frac{3+2+2}{6R} = \frac{7}{6R} \quad (15)$$

$$R_{total} = \frac{6R}{7}, \quad (16)$$

and the simplified circuit will be as shown in Fig. 14. Thus, the inverter input voltage can be simply found by voltage division as follows (there is no current drawn on the inverter input):

$$V_{in} = V_{DD} \frac{\frac{6R}{7}}{\frac{6R}{7} + 2.5R} = 0.25V_{DD} \quad . \quad (17)$$

Now, the threshold voltage of the inverter must be chosen so that the discharging transistor, *Q*, will be activated in case all the inputs are at logic "1" and not activated in case all the inputs are not at logic "1". The worst case occurs when the voltage at the input of the inverter is as close as possible to that in the case of all activated inputs. This will happen when the discharging current through the pull-down network decreases slightly and this corresponds to the case shown in Fig. 15.

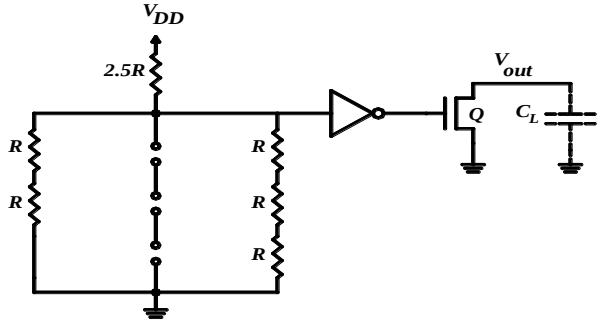

Fig. 15 The circuit of the proposed scheme for the case A_3, A_4, or A_5 is at logic "0". This is done in order to choose the threshold voltage of the inverter assuming the worst case.

Proceeding with the same analysis steps as before, we will obtain the equivalent resistance of the pull-down network as

$$R_{total} = \frac{3R(2R)}{3R+2R} = \frac{6R^2}{5R} = \frac{6R}{5} \quad . \quad (18)$$

So, the inverter input voltage will be

$$V_{in} = V_{DD} \frac{\frac{6R}{5}}{\frac{6R}{5} + 2.5R} = 0.324V_{DD} \quad . \quad (19)$$

So, from Eqs. (17) and (19), we will get the following range within which we can choose the threshold voltage of the inverter:

$$0.324V_{DD} - 0.25V_{DD} = 0.074V_{DD} \quad . \quad (20)$$

Had we assumed that the other branch that has a total resistance of 2*R* is deactivated, then the equivalent resistance of the pull-down network will be 1.5*R* and the inverter input voltage will thus be

$$V_{in} = V_{DD} \frac{1.5R}{1.5R + 2.5R} = 0.375V_{DD} \quad . \quad (21)$$

The corresponding range within which we can choose the threshold voltage of the inverter will be

$$0.375V_{DD} - 0.25V_{DD} = 0.125V_{DD}, \quad (22)$$

which is larger than that obtained for the first analysis. Thus, it does not represent the worst case.

Now, assume that the pull-down network is modified to contain 4 NMOS transistors in each branch as shown in Fig. 16. In case all the inputs are activated, the equivalent circuit will be as shown in Fig. 17. The equivalent resistance of the pull-down network will thus be

$$R_{total} = 2R \quad . \quad (23)$$

So, by means of the voltage division between the 2.5*R* and the 2*R* resistors, the inverter input voltage will be (Fig. 18)

$$V_{in} = V_{DD} \frac{2R}{2R + 2.5R} = 0.444V_{DD} \quad . \quad (24)$$

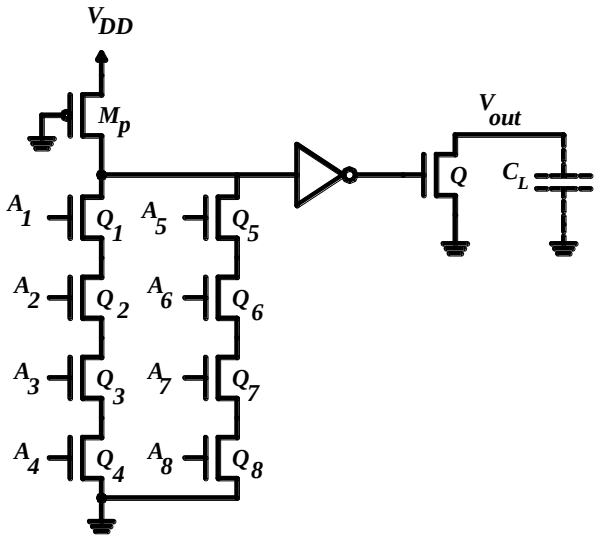

Fig. 16 An alternative configuration for the NMOS transistors in the pull-down network according to the proposed scheme.

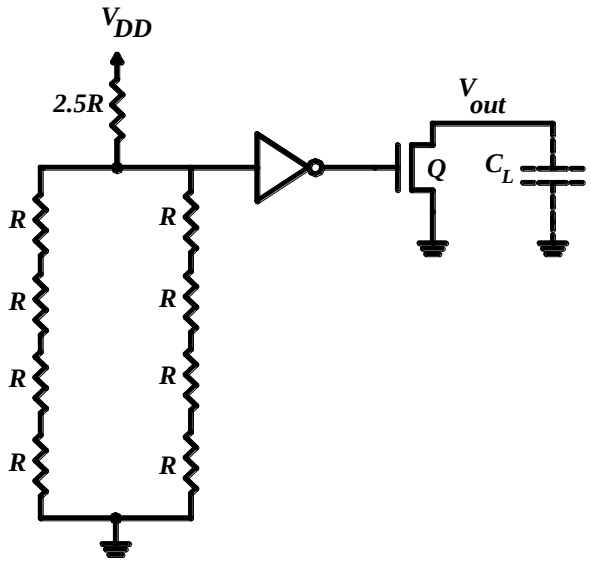

Fig. 17 The equivalent circuit of the proposed scheme for the alternative configuration.

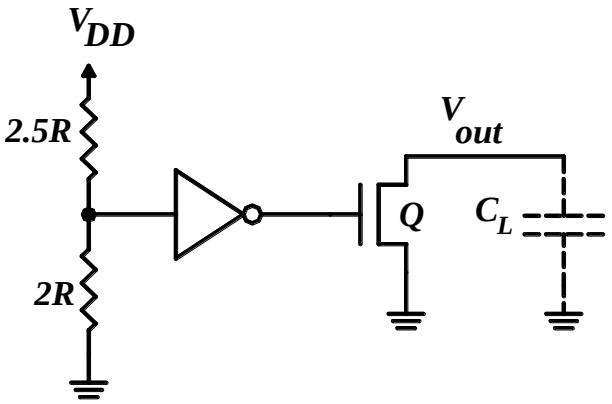

Fig. 18 The equivalent circuit of Fig. 17 after resistor combination.

In case any one of the inputs is deactivated, the corresponding branch will be open-circuited. So, the inverter input voltage will be

$$V_{in} = V_{DD}\frac{R}{R+2.5R} = 0.285V_{DD} \quad (25)$$

So, the range within which the threshold voltage of the inverter can be chosen will be

$$0.444V_{DD} - 0.285V_{DD} = 0.16V_{DD} \quad (26)$$

which is larger than the previous case. The robustness to process variations will thus be better, however, at the expense of increasing area in order to obtain the same speed. The increased area is expected due to the longer stack length and the smaller number of the discharging paths. The circuit designer must decide on the optimum shape for the pull-down network according to the application requirements.

VII. Simulation Results

The proposed scheme is simulated for the 0.13 µm CMOS technology with the power-supply voltage, V_{DD}, put equal to 1.2 V and using Multisim 11. All the transistors are minimum-sized according to the conventional and proposed schemes except for the strong NMOS transistor which has an aspect ratio of 5. The simulation results are shown in Fig. 19 for the case of 8 series-connected NMOS transistors in the chain. The modified circuit is chosen to have two branches with 4 NMOS transistors in each. The fall time is measured between the time instant at which the inputs are activated to the time instant at which the voltage across the output parasitic capacitance reduces to 10% of its initial value. According to the simulation results, the fall times of the conventional and proposed schemes will be 395 ps and 240 ps, respectively, thus saving approximately 40% of the time delay.

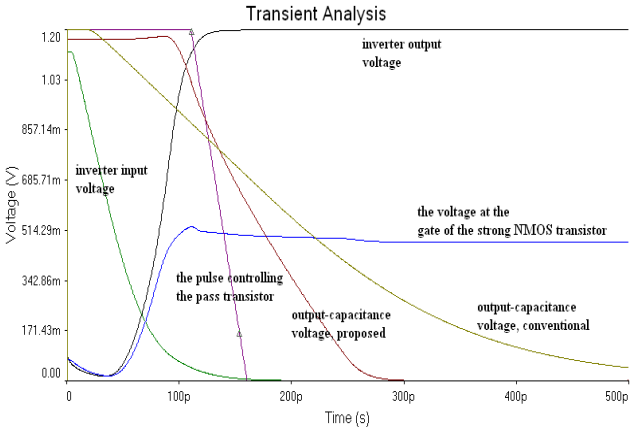

Fig. 19 The simulation results of the conventional and proposed schemes for the case of all activated inputs.

If the aspect ratio of the PMOS transistor of the inverter in the proposed circuit increases, the inverter threshold voltage will increase. This in turn speeds-up the appearance of logic "1" at the gate of the discharging transistor, thus speeding-up the operation. Refer to Fig. 20 for a parameter-sweep analysis showing the effect of varying the aspect ratio of this PMOS transistor on the voltage across the output capacitance. It can be shown that increasing the (W/L) of the inverter PMOS transistor from 1 to 3 reduces the discharging time delay from 240 ps to 173 ps. Further increase in the size of this transistor will not result in perceptible improvement.

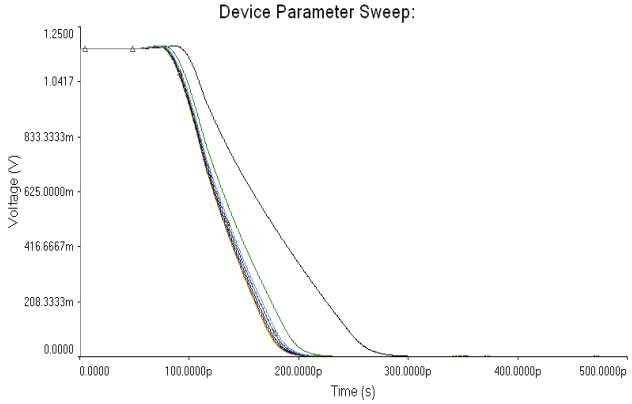

Fig. 20 Parameter-sweep analysis showing the effect of varying the aspect ratio of the PMOS transistor of the inverter used according to the proposed scheme.

VIII. Conclusions

In applications that include gates with wide fan-in,

there may be a long chain of MOS transistors that result in a relatively slow performance. In this paper, this problem was discussed and a solution was proposed. The proposed technique depends on using a circuit whose output voltage is proportional to the number of activated inputs. If all the inputs except one are activated, then the output voltage will be smaller than the threshold voltage of a strong NMOS transistor whose job is to discharge the output parasitic capacitance. If all the inputs are activated, the opposite occurs and the strong NMOS transistor will be activated, thus discharging the parasitic capacitance. The method was simulated for the 0.13 µm CMOS technology for a circuit that contains a stack of 8 NMOS transistors. The time delay can be reduced by about 40%. The proposed method was investigated quantitatively in order to determine the percentage reduction in the discharging time delay and to decide on the optimum configuration of the proposed circuit from the point of view of robustness to process variations.

REFERENCES

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

[2] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, Fourth Edition, Oxford University Press, 1998.

[3] D. A. Neaman, *Semiconductors Physics and Devices: Basic Principles*, Second Edition, Irwin, 1997.

[4] 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.

[5] J. P. Uyemura, *CMOS Logic Circuit Design*, Kluwer Academic Publishers, 2002.

[6] V. Kursen et al., *Multi-Voltage CMOS Circuit Design*, John Wiley & Sons Ltd., 2006.

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

[8] S. M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits, Analysis and Design*, Second Edition, McGraw-Hill, 1999.

[9] L. Ding and P. Mazumder, "On Optimal Tapering of FET Chains in High-Speed CMOS Circuits," IEEE Transactions on Circuits and Systems, vol. 48, no. 12, pp. 1099–1109, December 2001.

[10] C. Zhang, "Techniques for Low Power Analog, Digital, and Mixed Signal CMOS Integrated Circuit Design," Doctor of Philosophy Thesis, Louisiana State University, May 2005.

[11] S. M. Sharroush, Y. S. Abdalla, A. A. Dessouki, and E. A. El-Badawy, "A Novel Domino CMOS Technique for Speeding up Circuits Containing a Long Chain of NMOS Transistors", International Conference on Electronic Design, Malaysia, 2008.

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