Reliability and performance of optimised Schmitt trigger gates

Mihai Tache1,2, Walid Ibrahim3,4, Fekri Kharbash3, Valeriu Beiu2
1Department of Mathematics and Computer Science, Faculty of Exact Sciences, Aurel Vlaicu University of Arad, Romania
2University Politehnica of Bucharest, Romania
3Department of Computer and Network Engineering, College of Information Technology, United Arab Emirates University, United Arab Emirates
E-mail: walidibr@uaeu.ac.ae

Abstract: This study compares the performance and reliability of classical complementary metal-oxide-semiconductor (CMOS) gates with Schmitt trigger (ST) ones. The ST hysteresis, caused by the added positive feedback transistors, improves the design static noise margin (SNM) and offers noise immune operation. Hence, ST-based circuits are expected to operate more reliably than the ones implemented using classical CMOS. Although many research papers have been focused lately on using ST design concepts for implementing more reliable static random access memory (SRAM) cells, significantly less work was devoted to the application of ST concepts in the combinatorial logic domain. Moreover, available research on ST-based logic gates had only focused on the low-voltage/power applications range. The authors are going to look at the whole voltage range and performance spectrum to compare and understand not only the SNMs and the power consumption (at different frequencies and voltage levels) but also the delay and the power-delay-product of ST-based logic gates. These will be compared with classical CMOS as well as with optimally sized CMOS and ST-based logic gates. This study should give a clear picture of the potential advantages ST could offer for combinatorial logic in advanced CMOS technology nodes and of their application range.

1 Introduction

Noise and variations in digital circuits cause undesirable fluctuations of voltage levels and electrical signalling, which could cause transistors to switch inappropriately and affect the reliability of logic gates, flip-flops, and memory bit-cells. Voltage levels may be affected by different types of noises including thermal, shot, flicker, as well as crosstalk. They also suffer from variations due to materials, fabrication processes, as well as environment. Still, the effects of noises and variations on the operation of digital circuits have been of little concern in the semiconductor community till lately. This is mainly due to the high noise margins and variation immunity of the classical complementary metal-oxide-semiconductor (CMOS) gates. However, with the aggressive scaling of CMOS features down deep into the nanometre range, CMOS transistors started to face several fundamental limitations. A major one is represented by variations, and in particular by the randomness of the exact locations of doping atoms [1–4], which leads to device-to-device fluctuations/variations in key parameters, including the threshold voltage ($V_{TH}$). Such variations are in addition to the intrinsic and extrinsic noises of billions of transistors on the same integrated circuit (IC). When adding these noises (intrinsic and extrinsic) on top of variations, it looks like that reliability will be one of the greatest threats to the design of future ICs [5].

At the same time, the current and future demand for portable ultra-low-power electronics, and the increasing demand for self-powered/energy-harvesting systems (e.g. wireless sensor nodes and implantable devices) have stimulated a growing interest for near- and sub-threshold computing [6–8]. However, due to their reduced supply voltage ($V_{DD}$), near-/sub-threshold circuits are even more susceptible to noises and process variations (than circuits operating at nominal $V_{DD}$). Consequently, reliability – which includes robustness to noises as well as tolerance to variations – has been identified by the International Technology Roadmap for Semiconductors as one of the grand challenges for both present and future very large scale integration (VLSI) designs [5].

A well-known approach for improving reliability is to incorporate different redundancy schemes at the device, gate, circuit, and/or system levels. These include modular redundancy [9], cascaded modular redundancy [10], as well as von Neumann multiplexing [11], enhanced von Neumann multiplexing [12, 13], and parallel restitution [14]. It is also well established that the most effective way (i.e. one which requires the least redundancy) to enhance reliability is to apply redundancy at the lowest level in a hierarchical system [15]. That is why transistor and gate level redundancy schemes have started to stir interests. Device (transistor) level redundancy schemes have been proposed since the 1990s [16–19], including threshold logic gates [20–28], extrapolating quadded logic [29] to the device level [30–32], as well as for nanotechnologies [33–38]. All of these are alternative approaches to gate and circuit/system level redundancy schemes, and aim to achieve reliable computation by improving on the switching error probabilities of the devices (transistors) and enhancing gates’ noise immunities and their tolerance to variations. Typical recent examples include: (i) resizing transistors (e.g. unconventional sizing like, e.g. upsizling L); (ii) using arrays of transistors (e.g. multi fingers, FinFETs); (iii) different gate designs (e.g. quadded, Schmitt trigger, ST); and (iv) combinations of such approaches.

One simple way to assess the noise-robustness and tolerance to variations of a gate is to start from its static noise margin (SNM) [39]. Maximising the SNMs of the gates will not only allow them to run well at nominal $V_{DD}$ but also to operate correctly (i.e. without violating the noise margins) down to ultra-low voltages (ULVs), hence potentially allowing for saving power/energy when/if needed.

A well-established approach to improve on a gate’s SNM is by relying on the ST design concept [40]. STs are gates with positive feedback (i.e. loop gain>1) that are used extensively to reduce sensitivity to concept [40]. STs are gates with positive feedback (i.e. loop gain>1) that are used extensively to reduce sensitivity to noises and to stabilise the output logic, hence potentially useful for reliability enhanced (hence also ULV) designs. The name ‘ST’ originated from the thermionic trigger circuit (gate) presented in [40]. The feedback can be adjusted such that the gate holds its output logic until the input changes above/below a predefined threshold level. Therefore, the undesired noises and variations will not be sufficient to trigger a change at the output. ST gates have been used in buffers [41], sensors [42], and pulse width modulation...
circuits [43]. Lately, many papers have been looking at using ST design concepts for implementing more reliable SRAM cells [44–46]. Still, significant less work has been targeting flip-flops and combinatorial logic (i.e. logic gates), and those who did have focused only on the ultra-low-voltage/power (ULV/P) applications range [47].

Another solution for improving SNMs is to rely on unconventional transistor sizing [48]. By slightly increasing the channel lengths (L) of all transistors, followed by balancing the voltage transfer curves (VTCs), an optimal transistor (OPT) sizing approach could significantly improve CMOS gates’ SNMs even under aggressive voltage scaling [49]. It has been shown that OPT sized CMOS gates exhibit much larger SNMs than classically sized (CLS) CMOS gates. The OPT sized CMOS gates can operate correctly over the whole voltage range (i.e. both above and below $V_{TH}$). An added benefit is that OPT sized CMOS gates consume significantly less power than CLS gates.

In this paper, we are going to compare in great details the CLS CMOS gates with OPT and ST versions, as well as with ST in combination with OPT (ST-OPT) sizing. The aim is not only to evaluate the SNMs achievable by these different approaches but also to compare their delays, power and power-delay-products (PDPs) at different supply voltage levels and operating frequencies. This paper is organised as follows: Section 2 introduces the ST design concept and details theoretical SNM calculations for ST inverters (ST-INVs). The OPT sizing approach is reviewed in Section 3. Detailed SNM, power, delay, and PDP simulation results and comparisons are provided in Section 4, followed by concluding remarks in Section 5.

2 Schmitt trigger design concept

This section presents the ST design concept using an inverter (INV) as it is the simplest logic gate. Fig. 1 shows the schematic and the VTC of a CLS-INV. In a noise and variation-free environment, when logic ‘low’ ($V_{IN} < V_{THN}$) is applied at the input ($V_{IN}$) of a CMOS-INV, a logic ‘high’ is expected at the output ($V_{OUT}$). Here $V_{THN}$ is the threshold voltage at which $N_0$ starts switching ‘on’. Similarly, when logic ‘high’ ($V_{IN} > V_{DD} – V_{THP}$) is applied at the input, $V_{OUT}$ is expected to be ‘low’, where $V_{THP}$ is the threshold voltage at which $P_0$ switches ‘on’. However, in a noisy environment, the voltage applied at the input is in fact $V_{IN} = V_{IN} \pm \sigma_{NOISE}$. Similarly, variations manifest themselves as $\sigma_{V_{THN}}$ and $\sigma_{V_{THP}}$, so $V_{THN} = V_{THN} \pm \sigma_{V_{THN}}$ and $V_{THP} = V_{THP} \pm \sigma_{V_{THP}}$. Therefore, in case of logic ‘low’ input, if $V_{NOISE} \geq V_{THN} – \sigma_{V_{THN}}$, $N_0$ will switch ‘on’ pulling erroneously $V_{OUT}$ to ‘low’. A similar erroneous behaviour could occur when logic ‘high’ is applied at the input. In this case, if $V_{IN} - V_{NOISE} \leq V_{DD} – V_{THP} + \sigma_{V_{THP}}$, $P_0$ will switch ‘on’, pushing $V_{OUT}$ to logic ‘high’ (see Fig. 1b).

Fig. 1c shows the schematic of ST-INV [50, 51], while Fig. 2 shows the schematic for ST-NAND-2 and ST-NOR-2. The ST-INV design consists of a classical inverter ($P_0, N_0$) and two more pairs of CMOS transistors ($N_1, N_2$) and ($P_1, P_2$). The two pairs are used to provide positive feedback when logic ‘low’ and ‘high’ are applied to the input ($V_{IN}$). Fig. 1d shows the VTCs of the ST-INV. When $V_{IN} < V_{THN}$, $V_{OUT} = V_{DD}$ and $N_2$ is ‘on’ and saturated, while $N_1$ and $N_0$ are both ‘off’. When $V \geq V_{THL}$, $N_1$ turns ‘on’, however, $N_0$ remains ‘off’ due to the feedback voltage ($V_{FB}$) at its drain terminal. The feedback voltage $V_{FB}$ depends on $V_{THN}$ and the sizing of $N_1$ and $N_2$, and can be estimated as [52]

$$V_{FB} = V_{DD} - V_{THN} - (V_{IN} - V_{THN}) \times \frac{W_N \times L_N}{W_N \times L_N + W_P \times L_P}$$

(1)

where $W$ and $L$ are the corresponding transistor's length and width. $V_{FB}$ improves the SNM of the ST-INV since a higher $\sigma_{V_{THN}}$ or larger $\sigma_{V_{THP}}$ are now required for switching $N_0$ ‘on’($V_{IN} = V_{THN} + V_{FB}$). Therefore, the ‘high’ switching threshold becomes:

$$V_{IN} = V_{THN} + (V_{DD} – V_{THN}) \times \left(1 + \frac{W_N \times L_N}{W_N \times L_N + W_P \times L_P}\right)^{-1}$$

(2)

Assuming that $N_1$ and $N_2$ are sized identically, the switching threshold becomes $(V_{DD} + V_{THN})/2$.
Similarly, when \( V_{\text{DD}} \geq V_N > V_{\text{DD}} - 1V_{\text{THP}} \), \( N_0, N_1, P_1 \) and \( P_2 \) are all 'on' and saturated while \( P_0, P_1 \) and \( N_2 \) are 'off'. For these conditions \( V_{\text{OUT}} = 0 \) and \( V_{\text{FB}} \) (the feedback voltage at the source terminal of \( P_0\)) = \( V_{\text{THP}} \). As \( V_N \) decreases, \( P_1 \) starts switching 'on' from \( V_{\text{IN}} = V_{\text{DD}} - 1V_{\text{THP}} \) and the feedback voltage \( V_{\text{FB}} \) drops to:

\[
V_{\text{FB}} = (V_{\text{DD}} - V_{\text{THP}} - V_{\text{IN}}) \times \frac{w_P \times L_{\text{eff}}}{w_P \times L_{\text{eff}}} - V_{\text{THP}} \tag{3}
\]

At this point \( V_{\text{OUT}} \) is still 0 and it will maintain a logic 'low' value until \( V_N = V_{\text{DD}} - 1V_{\text{THP}} \). Hence, the ST-IN 'low' switching threshold is:

\[
V_{\text{IN}} = V_{\text{DD}} - 1V_{\text{THP}} - (V_{\text{DD}} - 1V_{\text{THP}}) \left(1 + \frac{w_P \times L_{\text{eff}}}{w_P \times L_{\text{eff}}} \right) \tag{4}
\]

Assuming that \( p_1 \) and \( p_2 \) are sized identically, the ST-IN 'low' switching threshold is \(( V_{\text{DD}} - 1V_{\text{THP}})/2\), as shown in Fig. 1d.

Equations (1) and (3) show that \( V_{\text{THP}} \), and hence the SNM of ST-IN, depends not only on \( V_{\text{TH}} \) but also on the sizing of the feedback transistors \( N_2 \) and \( P_2 \). For example, increasing the logic 'low' switching threshold can be done by increasing \( W_{N_2} \) and/or \( L_{N_2} \).

Normally, it is very hard (in fact impossible in most cases) to control SNMs once a circuit has been designed. The ability to fine tune the switching thresholds of ST-based gates allows them to work in different environments where they tolerate different noise levels. This was suggested in 2003 by using body biasing [53]. Recently, independent tuning of the ST switching thresholds within 30% of \( V_{\text{DD}} \) was achieved by controlling the gate terminal of dual control gate-floating gate transistors [54]. A differential ST with tunable SNMs in 0.18 \( \mu \text{m} \) was proposed in [55]. By using the cross-coupled inverter pair and varying the current of the symmetrical load, the author was able to adjust the switching threshold voltage from 0.95 to 1.35 V. This range was improved even further, from 0.5 to 1.8 V in [56].

An effective way to reduce the power consumption of modern system-on-chip (SoC) designs is to reduce a significant portion of the power consumed by SRAM cells. This makes the implementation of reliable ULP SRAM cells crucial for future SoC applications. Here again, ST designs have started to be relied upon. In [44], the authors investigated the impact of transistor variations (FinFET in this particular case) on the standard 6 and 8 T ST-based SRAM cells when working in sub-threshold. Their investigation showed a significant (81%) improvement of the Read-SNM in case of 8 T ST-based SRAM cell compared to the standard 6 T one. In another study [45], a 10 T ST-based SRAM cell was proposed for ULP SRAM. The simulation results showed 1.6× higher Read-SNM, 2× higher write-trip-point, and 120 mV lower 'read' voltage when compared to the standard 6 T SRAM cell.

Due to their enhanced SNMs, ST logic gates were also used successfully to mitigate the output level degradations of standard cell-based CMOS logic circuits [47]. The authors were able to demonstrate a fully operational sub-threshold IC with \( V_{\text{THP}} \) ranging between 62 and 84 mV in standard 0.13 \( \mu \text{m} \) bulk CMOS. They have also reported that ST logic gates show a very effective suppression of global nMOS versus pMOS process variations in corner simulations, while the ST logic gates they have analysed are still affected by transistor-to-transistor variability. Another approach for improving noise-tolerance by selectively applying feed-forward reinforcement using a modification of the standard ST circuit was presented in [57].

### 3 Unconventional sizing for enhancing reliability

The way transistors are sized in digital CMOS gates is very well established and was originally aimed at maximising the performance, while it is only lately that it has started to be revisited for trying to minimise power/energy [58–60]. Sizing has also started to be evaluated as a promising option for enhancing reliability (noise and variation tolerant gates), aiming to minimally increase area, while also reducing power/energy [61–65].

To maximise performance, VLSI designers have routinely set the channel lengths for the nMOS and pMOS transistors to the minimum (i.e. \( L_{\text{MNOS}} = L_{\text{PMOS}} = \text{min} \)). and increased the channel widths aiming to balance the rise and fall times of CMOS gates. This classical sizing method is well ingrained in the VLSI community. For an easier understanding of the concept of unconventional sizing for maximising SNMs, we start from the elementary transistors. The probability that \( V_{\text{TH}} = v \) for a transistor can be calculated as [66]

\[
P(V_{\text{TH}} = v | \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left[-\frac{(v - \mu)^2}{2\sigma^2}\right] \tag{5}
\]

For each transistor, we have estimated \( \sigma_{V_{\text{TH}}} \) as [3]

\[
\sigma_{V_{\text{TH}}} \approx 3.19 \times 10^{-4} \times \frac{T_{\text{ox}} \times N_{\text{dep}}}{\sqrt{L_{\text{eff}} \times W_{\text{eff}}}} \tag{6}
\]

\[
L_{\text{eff}} = L_{\text{drawn}} + X_L - 2 \times L_{\text{INT}} \tag{7}
\]

\[
W_{\text{eff}} = W_{\text{drawn}} + X_W - 2 \times L_{\text{INT}} \tag{8}
\]

Here, \( L_{\text{eff}} \) and \( W_{\text{eff}} \) are the effective length and width of the transistor's channel, \( T_{\text{ox}} \) is the oxide thickness, \( N_{\text{dep}} \) is the channel doping concentration at depletion edge for zero body bias, \( X_L \) is the channel length offset (due to mask/etch effect), \( L_{\text{INT}} \) is the channel length offset parameter, \( X_W \) is the channel width offset (due to mask/etch effect), \( W_{\text{INT}} \) is the channel width offset parameter. Using PTM HP v2.1 (high-k/metal gate and stress effect) [67–69] and BSIM4 v4.8 level 54 [70] and assuming that \( V_{\text{TH}} = 0 \), we could estimate \( \sigma_{V_{\text{TH}}} \) as

\[
\mu_{V_{\text{TH}}} = V_{\text{TH}} + V_{\text{TH-DBL}} \tag{9}
\]

where

\[
V_{\text{TH-DBL}} = \frac{\eta_{\text{DBL}} \times V_{\text{DS}}}{\frac{w_{\text{DBL}} \times d_{\text{DBL}}}{\phi \times q \times N_{\text{dep}}}} \tag{10}
\]

\[
L_{\text{DBL}} = \sqrt{\frac{\frac{w_{\text{DBL}} \times d_{\text{DBL}}}{\phi \times q \times N_{\text{dep}}}}{\phi}} \tag{11}
\]

\[
X_{\text{dep}} = \frac{2 \times \phi \times q}{\phi \times q \times N_{\text{dep}}} \tag{12}
\]

\[
\phi = \frac{2kT}{q} \times \ln(N_{\text{dep}}/n_i) \tag{13}
\]

Here \( V_{\text{TH-DBL}} \) is the long channel threshold voltage at \( V_{\text{TH}} = 0 \), \( \eta_{\text{DBL}} \) is the DBL (dual induced barrier lowering) coefficient in sub-threshold, \( D_{\text{SUB}} \) is the DBL coefficient exponent in sub-threshold, \( L_{\text{DBL}} \) is the characteristic length, \( n_i \) is permittivity of silicon, \( T_{\text{ox}} \) is the oxide thickness, \( X_{\text{dep}} \) is the depletion width, \( \epsilon_{\text{ox}} \) is the permittivity of the oxide, \( \phi \) is the surface potential, \( q \) is the electron charge, \( T \) is the temperature, \( N_{\text{dep}} \) is the channel doping concentration at depletion edge for zero body bias, and \( n_i \) is the intrinsic carrier concentration in the channel region. In fact, \( V_{\text{TH}} \) can be calculated much more accurately as

\[
\mu_{V_{\text{TH}}} = V_{\text{TH-BE}} + V_{\text{TH-DBL}} - V_{\text{TH-DBL}} \tag{14}
\]

Equation (14) is used in this paper for calculating \( V_{\text{TH}} \) as precisely as possible.

---

This is an open access article published by the IET under the Creative Commons Attribution -NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/).
Fig. 3 shows that a slight increase of \( L_{\text{drawn}} \) above minimum \( (L_{\text{min}} = 22 \text{ nm}) \) causes a steep modification of \( \mu_{VTH} \). It also shows a reduction of \( \sigma_{VTH} \) as \( L_{\text{drawn}} \) increases (±\( \sigma_{VTH} \) given by (6) are shown as bars). On the other hand, increasing \( W_{\text{drawn}} \) above minimum \( (W_{\text{min}} = 22 \text{ nm}) \) reduces \( \sigma_{VTH} \) but it does not affect \( \mu_{VTH} \). Fig. 4 shows the effect of increasing \( W_{\text{drawn}} \) and \( L_{\text{drawn}} \) on the \( VTH \) distribution of the nMOS transistors. The figure reveals several important issues. First, the \( VTH \) distribution is not balanced. The probability of having \( V_{TH} = 0 \) is much higher than the probability of having \( V_{TH} = 0.8 \) (nominal voltage). Increasing \( W_{\text{drawn}} \) from 22 to 44 nm have significantly reduces \( P(V_{TH} = 0.8) \) further. However, it has a very modest effect on \( P(V_{TH} = 0) \). On the other hand, increasing \( L_{\text{drawn}} \) significantly reduces \( P(V_{TH} = 0) \) while have almost no effect on \( P(V_{TH} = 0.8) \). This means that increasing \( L_{\text{drawn}} \) could have a better effect of the switching probability of the transistor than increasing \( W_{\text{drawn}} \).

A switching failure occurs when the variation of \( VTH \) and the noise at the gate signal causes the transistor to switch incorrectly ON or OFF. For example, in the case of nMOS transistors, when logic low is applied at the gate, the probability that the transistor will switch incorrectly ON can be calculated as

\[
\text{PF}_{\text{nMOS}(\text{high})} = 0.5 \times \left( 1 + \text{erf} \left( \frac{V_{IN} + V_{\text{NOISE}} - \mu_{VTH}}{\sigma_{VTH} \times \sqrt{2}} \right) \right)
\]

While \( \text{PF}_{\text{nMOS}(\text{high})} \) can be calculated as

\[
\text{PF}_{\text{nMOS}(\text{high})} = 0.5 \times \text{erfc} \left( \frac{V_{IN} - V_{\text{NOISE}} - \mu_{VTH}}{\sigma_{VTH} \times \sqrt{2}} \right)
\]

Similarly, in the case of pMOS, the switching probability of failure (PF) can be calculated as

\[
\text{PF}_{\text{pMOS}(\text{low})} = 0.5 \times \left( 1 + \text{erf} \left( \frac{V_{DD} + (V_{IN} + V_{\text{NOISE}}) - \mu_{VTH}}{\sigma_{VTH} \times \sqrt{2}} \right) \right)
\]

\[
\text{PF}_{\text{pMOS}(\text{high})} = 0.5 \times \text{erfc} \left( \frac{-V_{DD} + (V_{IN} - V_{\text{NOISE}}) + |\mu_{VTH}|}{\sigma_{VTH} \times \sqrt{2}} \right)
\]

Using (5) in conjunction with (6) and (14–19), the probability that a transistor exhibits a switching error (PF) was estimated for the \( V_{IN} \) at both logic high and low. In this paper, the probability that the transistor switches incorrectly is called the PF. Table 1 presents the effect of transistor sizing on switching PF of bulk CMOS transistors assuming \( V_{NOISE} = 50\text{ mV} \). It reveals a noticeable difference between the PF of nMOS and pMOS transistors, and also between the PF of the same transistor at different input voltages (‘high’ or ‘low’). For instance, Table 1 shows that, in the case of minimally sized transistors \([20 \times 20 \text{ nm}] \), when logic low is applied at the gate terminal \( (V_{IN} = 0 + 50\text{ mV}) \), the pMOS transistor has a probability \( = 5.32 \times 10^{-8} \) to switch incorrectly OFF, while the nMOS transistor has a probability \( = 6.2 \times 10^{-3} \) to switch incorrectly ON. Similarly, when logic high is applied at the gate terminal \( (V_{IN} = V_{DD} = -50\text{ mV}) \), the pMOS transistor has a probability \( = 2.77 \times 10^{-2} \) to switch incorrectly ON, while the nMOS transistor has a probability \( = 4.74 \times 10^{-6} \) to switch incorrectly OFF.

One way to reduce the effects of \( VTH \) variations, and improve PF, is to increase the channel width of the transistor \( (W_{\text{drawn}}) \) as customarily done in the semiconductor industry. Fig. 4 shows the effect of increasing \( W_{\text{drawn}} \) and \( L_{\text{drawn}} \) on the probability and the mean value of \( V_{TH}(\mu_{VTH}) \). It shows that increasing \( W_{\text{drawn}} \) has no effect on \( \mu_{VTH} \) (the line drawn at the bottom of the figure). Table 1 shows that increasing \( W_{\text{drawn}} \) from 22 to 44 nm, and then to 66 significantly reduces \( V_{TH} \) variations \( (\sigma_{VTH}) \) and hence the switching PF. Increasing \( W_{\text{drawn}} \) from 22 to 44 nm reduces \( \sigma_{VTH} \) from 0.101 to 0.060 V, while \( \mu_{VTH} \) remains unchanged at 0.303 V. This consequently reduces \( \text{PF}_{\text{nMOS(low)}} \) to \( 1.28 \times 10^{-5} \), while \( \text{PF}_{\text{nMOS(high)}} \) is significantly reduced to \( 4.51 \times 10^{-14} \).

Table 1 also confirms that increasing \( L_{\text{drawn}} \) has a better effect on the transistor overall switching reliability compared to increasing \( W_{\text{drawn}} \), while the minimum sized nMOS transistor has a worst case \( 6.2 \times 10^{-3} \) when logic is applied, increasing to 44 nm reduces the worst case of \( 1.28 \times 10^{-5} \) which still occurs when a logic is applied \( (4.51 \times 10^{-14}) \). On the other hand, increasing to 44
shows that slightly increasing the minimum transistor size is used (i.e. 22 × 22 nm). Fig. 5 illustrates the effect of tuning Ldrawn is increased is more than two orders of magnitude better than when Ldrawn is increased. Similar behaviour is also observed in case of the pMOS transistor.

Since CMOS gates use both nMOS and pMOS transistors, reducing the probability that a gate switches incorrectly should be done in two steps:

i. Making the switching probabilities of failure equal at logic low and logic high for each and every transistor.

ii. Reducing σVTH while also aiming to make PFnMOS ≈ PFpMOS, i.e. balancing the nMOS and pMOS transistors as trying to make σVTHnMOS = σVTHpMOS.

The first step means that VTH of both transistors should be as close as possible to VDD/2, i.e. symmetrical and as far as possible from both VDD and GND (i.e. VTH = |VTHp| = VDD/2). This equalises (balances) the PF of a given transistor at GND and VDD. Obviously, VTH can be modified by tuning L as shown in (14) and Table 1. Equation (14) is used to determine the Ldrawn for which VTH = VDD/2. L drawn for these ‘tuned’ transistors are then calculated using (7). We call Ldrawn, for which VTH = VDD/2, the optimum channel length (Lopt).

Extensive simulations for 22 nm PTM HP v2.1 have identified Lopt,nMOS = 25.3 nm, and Lopt,pMOS = 29.8 nm [49, 65]. Sizing in very small increments (e.g. even below 1 nm) could be envisaged by relying on the optical proximity correction as suggested in [71, 72].

Fig. 5 illustrates the effect of tuning Ldrawn and Wdrawn on the switching PF as a function of VNOISE. Fig. 5a shows the PF when the minimum transistor size is used (i.e. 22 × 22 nm). Fig. 5b shows the PF of the transistors when using Lopt and WMIN. It shows that slightly increasing L from Lmin to Lopt reduces PFnMOS(low) by more than two orders of magnitude to 2.65 × 10^{-5}, while PFpMOS(high) is reduced to 4.0 × 10^{-7} (i.e. by more than four orders of magnitude).

As mentioned above, the aim of the second step is to improve PF even further by reducing σVTH, and at the same time balancing the nMOS and pMOS transistors (i.e. PFnMOS ≈ PFpMOS). Reducing σVTH can be done by increasing W and/or L as shown by (6). It could also be done by using arrays of identically sized transistors (e.g. FinFETs). Since modifying Lopt would undo the advantages already obtained during the first step, in the second step the circuit designers are left only with the option of increasing W’s. Increasing the WnMOS from 22 to 27.9 nm reduces σVTHnMOS from 0.0864 to 0.07075, which ≈ σVTHp (0.0708). Consequently reduces PFnMOS further to 4.0 × 10^{-5} (see Fig. 5c). Obviously, OPT sized transistors will have better PFs than CLS sized transistors (at the same W) and will compose gates which are more reliable, as having better noise immunity and higher tolerance to variations.

### 4 Classical, optimised, and ST gates with optimal sizing

Section 3 describes two steps to improve the PFs of the nMOS and pMOS transistors. In this section, a third step is used to improve the SNM/reliability further. The third step moves from the transistor level to the gate level aiming at maximising SNM by balancing the gate’s VTC. This balancing is not trivial as each gate could have more than one VTC depending on the applied input vector. It is impossible to make all these VTCs cross at VDD/2. That is why the problem has to be restated as a joint balancing of all the VTCs of the CMOS gate to cross as close as possible and evenly distributed around VDD/2 by tuning only the channel width of the transistors (W’s) [49]. Inverters are obviously the simplest gates, and they can be perfectly balanced as their VTCs have only one VDD/2 crossing.

In the case of ST gates, optimising the transistor sizing is even more challenging as the number of transistors doubles. In this research, we simultaneously optimise all possible input transitions, aiming to maximise the hysteresis between the worst possible transitions. To do that all possible the input combinations were represented by slow (1 ms) ramp signals, all the possible input conditions. Sizing is then done such that the worst outputs are spaced symmetrically and as far as possible from VDD/2. Table 2

| Size (W × L), nm | 22 × 22 | 44 × 2 | 66 × 22 | 22 × 44 |
|-----------------|--------|--------|--------|--------|
| nMOS PF (low)   | 6.20 × 10^{-3} | 1.28 × 10^{-5} | 3.30 × 10^{-6} | 4.44 × 10^{-16} |
| nMOS PF (high)  | 4.74 × 10^{-6} | 4.51 × 10^{-14} | 5.50 × 10^{-22} | 7.37 × 10^{-7} |
| pMOS PF (low)   | -0.235 | -0.235 | -0.235 | -0.444 |
| pMOS PF (high)  | 0.097 | 0.058 | 0.045 | 0.052 |

This is an open access article published by the IET under the Creative Commons Attribution -NonCommercial License (http://creativecommons.org/licenses/by-nc/3.0/)

---

**Table 1** Effect of sizing on switching PF of bulk CMOS transistors, assuming VNOISE = 50 mV

---

**Fig. 5** Switching PF for CMOS transistors (W × L): (a) Minimum (22 × 22 nm), (b) Optimal (nMOS: 22 × 25.3 nm, pMOS: 22 × 29.8 nm), (c) Balanced (nMOS: 27.9 × 25.3, pMOS: 22 × 29.8 nm)
shows the CLS and OPT sizing and their effect on \( V_{TH} \) and \( \sigma_{VTH} \) variation for INV, NAND-2, and NOR-2, while Table 3 shows that same data for the ST version.

5 Simulation results

Spice simulations were used to compare the performances, not only SNMs but also delay, power, and PDPs of INV, NAND-2, and NOR-2 gates implemented both as CMOS gates as well as ST gates, while the transistors are sized both classically (CLS) and optimally (OPT). All the simulations have used 22 nm PTM HP v2.1 (high-k/metal gate and stress effect). The basic setting for power and delay simulations was for one gate to drive four identical gates, was varied between 200 and 800 mV in increments of 100 mV, while decreasing the frequency from 1 GHz (1 ns) down to 1 kHz (1 ms) as a power of 10.

The first experiment investigated the effects of using the OPT and the ST approaches on improving the SNMs. The SNMs for all gates under investigation were calculated from their VTCs using the maximum square method [73]. Since SNM is usually calculated relative to the operating \( V_{DD} \). Table 4 shows the SNM of the INV, NAND-2, and NOR-2 gates measured as a percentage of the operating \( V_{DD} \) when the four design methods are used. At 800 mV the OPT sizing method improved the SNM by 10% over CLS-INV, while ST-CLS and ST-OPT show improvements of 29% and 32% over CLS-INV respectively. This trend is sustained in sub-threshold (200 mV); however, the SNM improvements over the classical INV are reduced to 7%, 17%, and 21% in case of OPT, ST-CLS, and ST-OPT, respectively. Similar improvement was also measured for the NOR-2 and NAND-2 gates as illustrated in the table. It is evident that the effect of the ST method on SNM is significantly higher than the effect of the OPT method.

The second experiment is used to characterise and compare the performance of the four design methods when operating at different \( V_{DD} \). During this experiment in addition to SNM measured in the first experiment, the delay, power, and PDP of each gate were measured and recorded for each design method when operating at different \( V_{DD} \) (200–800 mV). To make the comparison clearer, normalised parameters were used instead of the absolute ones. The normalised parameters for each gate are calculated with respect to the best measured values. For example, the normalised power for method \( i \) of gate \( j \) (\( nPWR_{ij} \)) is calculated as

\[
nPWR_{ij} = PWR_{i,j} / PWR_{0,j}
\]

Table 2 CLS and OPT sizing (\( W \times L \)) in nm, and their effects on \( V_{TH} \) and \( \sigma_{VTH} \) of CMOS gates

|       | CLS | OPT |
|-------|-----|-----|
| VTH   |     |     |
| \( \sigma_{VTH} \) |     |     |
| W x L |     |     |
| 66 x 22 | -235 | 44.8 |
| 44 x 22 | 303 | 60 |
| 55 x 22 | -235 | 50 |
| 44 x 22 | 303 | 60 |
| 47 x 22 | 303 | 57.5 |
| 44 x 22 | 303 | 60 |

Table 3 CLS and OPT sizing (\( W \times L \)) in nm, and their effects on \( V_{TH} \) and \( \sigma_{VTH} \) of ST gates

|       | CLS | OPT | ST | ST-OPT |
|-------|-----|-----|----|--------|
| VTH   |     |     |    |        |
| \( \sigma_{VTH} \) |     |     |    |        |
| W x L |     |     |    |        |
| 87 x 22 | -235 | 37.5 |
| 66 x 22 | -235 | 44.8 |
| 66 x 22 | -235 | 44.8 |
| 66 x 22 | -235 | 55.2 |
| 55 x 22 | -235 | 303 |
| 44 x 22 | 399 | 51.3 |
| 44 x 22 | 399 | 51.3 |
| 55 x 22 | 50 | 183 |
| 77 x 42 | 43 | 66 |
| 44 x 22 | 603 | 44 |

Table 4 SNM as a percentage of operating \( V_{DD} \) for INV, NAND-2, and NOR-2

|       | CLS | OPT | ST | ST-OPT |
|-------|-----|-----|----|--------|
| VTH   |     |     |    |        |
| \( \sigma_{VTH} \) |     |     |    |        |
| W x L |     |     |    |        |
| 200 | 23.0% | 30.0% | 39.5% |
| 300 | 27.7% | 34.7% | 48.3% |
| 400 | 29.8% | 37.3% | 53.0% |
| 500 | 30.8% | 38.4% | 55.6% |
| 600 | 31.2% | 39.2% | 57.0% |
| 700 | 30.9% | 39.4% | 58.0% |
| 800 | 29.5% | 39.3% | 58.5% |

\( nPWR_{ij} = PWR_{i,j} / PWR_{0,j} \)
shows that the CLS design method exhibits the best normalised delay occurs at nominal V_{DD} = 540 mV. At nominal V_{DD}, the CLS design method becomes the favourable method, while ST-OPT can be used for application running in near threshold.

Our simulation results also show that both INV and NOR-2 gates exhibit similar behaviours as the NAND-2 gate as illustrated in Table 5. The italic values mark the design methods and the operating voltages that have the best performance (e.g. ST-OPT and 800 mV in the case of SNM). On the other hand, the bold values mark the designs and operating voltages with the worst performance (e.g. CLS operating at 200 mV in the case of SNM).

Third experiment aims to evaluate the advantages of applying the different design methods for different applications. To simplify the comparison, we used a figure of merit (FoM) to capture the relative improvements in SNM, power, delay, and pdp of a certain method. Basically, the FoM is a weighted sum of the normalised parameters described above. Four different weight sets have been considered to illustrate four different application modes:

- normal applications:
  FoM1 = 40% nSNM + 30% nPWR + 30% nDLY
- low-power applications:
  FoM2 = 40% nSNM + 50% nPWR + 10% nDLY
- high-performance applications:
  FoM3 = 40% nSNM + 10% nPWR + 50% nDLY
- low-energy applications:
  FoM4 = 40% nSNM + 60% nPDP.

Fig. 7 shows the performance of NAND-2 gate with respect to the four FoMs when each of the fours methods is applied. It is interesting to see that the FoM shapes change significantly and are strongly influenced by the weighting factors. This gives the circuit designers the opportunity to select the option that best fit their applications. Fig. 7a shows that the OPT method is the preferred design method for normal applications running in subthreshold. However, it becomes the worst method when the gate is operating at nominal V_{DD}. At nominal V_{DD}, the CLS design method becomes the favourable method, while ST-OPT can be used for application running in near threshold.

Table 5: Effect of design method on the performance of INV, and NOR-2 gates

| V_{DD} | Normalised SNM | Normalised power | Normalised delay | Normalised PDP |
|--------|----------------|-----------------|-----------------|----------------|
|        | CLS | OPT | ST | ST-OPT | CLS | OPT | ST | ST-OPT | CLS | OPT | ST | ST-OPT |
| 200    | 0.379 | 0.490 | 0.647 | 0.717 | 0.332 | 1.000 | 0.170 | 0.547 | 5.6 × 10^{-4} | 1.4 × 10^{-4} | 9.2 × 10^{-5} | 2.7 × 10^{-5} | 0.013 | 0.009 | 1.1 × 10^{-3} | 0.001 |
| 300    | 0.451 | 0.568 | 0.788 | 0.847 | 0.112 | 0.487 | 0.060 | 0.271 | 5.6 × 10^{-3} | 0.001 | 8.8 × 10^{-4} | 2.5 × 10^{-4} | 0.043 | 0.042 | 0.004 | 0.005 |
| 400    | 0.485 | 0.606 | 0.863 | 0.915 | 0.042 | 0.266 | 0.024 | 0.151 | 0.045 | 0.010 | 0.007 | 0.002 | 0.128 | 0.182 | 0.011 | 0.020 |
| 500    | 0.503 | 0.627 | 0.905 | 0.954 | 0.017 | 0.154 | 0.010 | 0.089 | 0.205 | 0.055 | 0.032 | 0.111 | 0.234 | 0.057 | 0.022 | 0.064 |
| 600    | 0.509 | 0.640 | 0.929 | 0.976 | 0.007 | 0.089 | 0.004 | 0.053 | 0.481 | 0.158 | 0.080 | 0.031 | 0.225 | 0.959 | 0.023 | 0.112 |
| 700    | 0.502 | 0.644 | 0.945 | 0.991 | 0.003 | 0.052 | 0.002 | 0.031 | 0.759 | 0.282 | 0.131 | 0.058 | 0.144 | 1.000 | 0.016 | 0.123 |
| 800    | 0.481 | 0.640 | 0.955 | 1.000 | 1.1 × 10^{-3} | 0.030 | 7.0 × 10^{-4} | 0.019 | 1.000 | 0.395 | 0.175 | 0.083 | 0.073 | 0.816 | 0.008 | 0.105 |

where PWR_j is the minimum measured power of a gate j across the four design methods and the whole voltage range (200–800 mV). Using the normalised values, it is much easier to identify the design and operating voltage that has the best performance (i.e. normalised value = 1).
The cells with bold text delay and PDP sensitive applications running in subthreshold. The positive feedback, caused by performance for delay sensitive applications, while the OPT gates with optimal sizing (ST-OPT) should be used as it performs significantly better than the other methods when applied for power sensitive application (FoM2) running in subthreshold, sensitive applications operating at near and above threshold.

Table 6: Effect of design method on SNM for the INV and NOR-2

| VDD | FoM1 (normal) | FoM2 (power sensitive) | FoM3 (delay sensitive) | FoM4 (PDP sensitive) |
|-----|---------------|------------------------|------------------------|-----------------------|
|     | CLS | OPT | ST | STOPT | CLS | OPT | ST | STOPT | CLS | OPT | ST | STOPT | CLS | OPT | ST | STOPT |
| 200 | 0.251 | 0.496 | 0.310 | 0.451 | 0.318 | 0.696 | 0.344 | 0.560 | 0.185 | 0.296 | 0.276 | 0.342 | 0.159 | 0.202 | 0.259 | 0.287 |
| 300 | 0.215 | 0.374 | 0.333 | 0.420 | 0.237 | 0.471 | 0.345 | 0.474 | 0.194 | 0.276 | 0.322 | 0.366 | 0.206 | 0.252 | 0.317 | 0.342 |
| 400 | 0.220 | 0.325 | 0.354 | 0.412 | 0.220 | 0.376 | 0.358 | 0.442 | 0.221 | 0.274 | 0.351 | 0.382 | 0.271 | 0.351 | 0.352 | 0.378 |
| 500 | 0.268 | 0.313 | 0.375 | 0.411 | 0.230 | 0.333 | 0.370 | 0.427 | 0.305 | 0.294 | 0.379 | 0.396 | 0.342 | 0.594 | 0.375 | 0.420 |
| 600 | 0.350 | 0.330 | 0.397 | 0.415 | 0.265 | 0.316 | 0.382 | 0.420 | 0.445 | 0.344 | 0.412 | 0.411 | 0.411 | 0.339 | 0.631 | 0.385 | 0.457 |
| 700 | 0.429 | 0.358 | 0.418 | 0.423 | 0.278 | 0.312 | 0.392 | 0.418 | 0.581 | 0.404 | 0.444 | 0.428 | 0.427 | 0.858 | 0.387 | 0.470 |
| 800 | 0.493 | 0.384 | 0.435 | 0.430 | 0.293 | 0.311 | 0.400 | 0.418 | 0.692 | 0.456 | 0.470 | 0.443 | 0.236 | 0.745 | 0.387 | 0.463 |

Fig. 7b shows that the OPT application is also the best method for power sensitive application (FoM2) running in subthreshold, while the ST-OPT is the preferred method for applications operating at nominal VDD. ST-OPT is also the preferred method for delay and PDP sensitive applications running in subthreshold. When operating at nominal VDD, the CLS method gives the best performance for delay sensitive applications, while the OPT method significantly outperforms the other three methods for PDP sensitive applications operating at near and above threshold.

It is noticeable that the CLS method has always the worst performance in subthreshold operation. In these two cases, the CLS or the OPT method could be used instead.

7 Acknowledgments

This research was funded in part by the United Arab Emirates University grant #31/T057, the Semiconductor Research Corporation grant 2012-TJ-2332, and the European Union through the European Regional Development Fund (ERDF) under the Competitiveness Operational Program (BioCell-NanoART = Novel Bio-inspired Cellular Nano-architectures, POC-A1.1.4-E-2015 nr. 30/01.09. 2016).

8 References

[1] Hagiwara, T., Yamaguchi, K., Shojiro, A.: ‘Threshold voltage deviation in very small MOS transistors due to local impurity fluctuations’. Proc. Int. Symp. VLSI Technology (VLSIT’82), Oiso, Japan, pp. 46–47
[2] Asenov, A.: ‘Random dopant induced threshold voltage lowering and fluctuations in sub-0.1µm MOSFET’s: a 3-D ‘atmospheric’ simulation study’, IEEE Trans. Electron Dev., 1998, 45, (12), pp. 2560–2563
[3] Asenov, A., Brown, A.R., Davies, J.H., et al.: ‘Simulation of intrinsic parameter fluctuations in decanometer and nanometer-scale MOSFETs’, IEEE Trans. Electron Dev., 2003, 50, (5), pp. 1837–1852
[4] Asenov, A.: ‘Statistical device variability and its impact on design’. IEEE Int. Symp. Asynchronous Circuits and Systems, Newcastle, UK, Apr. 2008, pp. xv–xvi
[5] Intl. Tech. Roadmap for Semiconductors (ITRS) 2.0, SEMATECH, Albany, NY, USA, 2015. Available at http://www.itrs2.net/
[6] Soeleman, H., Roy, K.: ‘Ultra-low power digital subthreshold logic circuits’. Proc. Int. Symp. Low Power Electronics and Design, San Diego, CA, USA, 1999, pp. 94–96
[7] Giustolisi, G., Palumbo, G., Criscione, M., et al.: ‘A low-voltage low-power voltage reference based on subthreshold MOSFETs’. IEEE J. Solid-State Circ., 2003, 38, (1), pp. 151–154
Lohstroh, J.: 'Static and dynamic noise margins of logic circuits', *IEEE J. Solid-State Circ.*, 1979, 14, (3), pp. 591–598.