Reliability and performance of optimised Schmitt trigger gates

Mihai Tache1,2, Walid Ibrahim3,2, Fekri Kharbash3, Valeriu Beiu1
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@uae.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-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 significantly less power than CLS gates.

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 to 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 $V_{IN} = V_{IN} + \sigma_{V_{NOISE}}$. Similarly, variations manifest themselves as $\sigma_{V_{THN}}$ and $\sigma_{V_{THP}}$, so $V_{THN} = V_{IN} + \sigma_{V_{NOISE}} + \sigma_{V_{THN}}$ and $V_{THP} = V_{THP} + \sigma_{V_{THP}}$. Therefore, in case of logic ‘low’ input, if $V_{NOISE}$ and $\sigma_{V_{THN}}$ are large enough such that $V_{NOISE} \geq V_{THN} - \sigma_{V_{THN}}$, $N_0$ 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 $V_{THN}$ is ‘on’ and saturated, while $N_1$ and $N_0$ are both ‘off’. When $V \geq V_{THN}$, $N_1$ 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]

$$V_{FB} = V_{DD} - V_{THN} - (V_{IN} - V_{THN}) \times \frac{W_1 \times L_1}{W_2 \times L_2}$$

(1)

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

$$V_{IN} = V_{THN} + (V_{DD} - V_{THN}) \times \left(1 + \frac{W_1 \times L_1}{W_2 \times L_2}\right)^{-1}$$

(2)

Assuming that $N_1$ and $N_2$ are sized identically, the switching threshold becomes ($V_{DD} + V_{THN}$)/2.
Similarly, when \( V_{DD} \geq V_N > V_{DD} - 1V_{THp} \), \( N_0, N_1, P_2, P_3 \) are all 'on' and saturated while \( P_0, P_1, N_2 \) are 'off'. For these conditions \( V_{OUT} = 0 \) and \( V_{FB} \) (the feedback voltage at the source terminal of \( P_2 \)) = \( V_{THp} \). As \( V_N \) decreases, \( P_1 \) starts switching 'on' from \( V_N = V_{DD} - 1V_{THp} \) and the feedback voltage \( V_{FB} \) drops to:

\[
V_{FB} = (V_{DD} - V_{THP} - V_N) + \frac{w_P \times L_P}{w_P \times L_N} - V_{THP}
\]  

(3)

At this point \( V_{OUT} \) is still 0 and it will maintain a logic 'low' value until \( P_2 \) switches 'on' at \( V_N = V_{DD} - 1V_{THP} - V_{FB} \). Hence, the ST-IN' 'low' switching threshold is:

\[
V_N = V_{DD} - 1V_{THP} - (V_{DD} - 1V_{THP})\left(1 + \frac{w_P \times L_P}{w_P \times L_N}\right).
\]  

(4)

Assuming that \( P_1 \) and \( P_2 \) are sized identically, the ST-IN' 'low' switching threshold is \( (V_{DD} - 1V_{THP})/2 \), as shown in Fig. 1d.

Equations (1) and (3) show that \( V_{THP} \), and hence the SNM of ST-IN', depends not only on \( V_{TH} \) 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 \( W_N \) and/or \( L_N \).

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 different environments where they tolerate different noise levels. This was suggested in 2003 by using body biasing [53].

A differential ST with tunable SNMs in 0.18 \( \mu m \) was proposed in [55]. By using the even further, from 0.5 to 1.8 \( V \) in [56].

Recently, independent tuning of the ST switching thresholds within different environments where they tolerate different noise levels. This was suggested in 2003 by using body biasing [53].

Independent tuning of the ST switching thresholds within different environments where they tolerate different noise levels. This was suggested in 2003 by using body biasing [53].

To maximise performance, VLSI designers have routinely set the channel lengths for the nMOS and pMOS transistors to the standard 6 and 8 \( \mu m \), 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 \( V_{TH} = v \) for a transistor can be calculated as [66]

\[
P(V_{TH} = v) = \frac{1}{2 \pi \sigma_{VTH}^2} \exp \left[-\left(\frac{v - \mu_{VTH}}{2 \sigma_{VTH}}\right)^2\right]
\]  

(5)

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

\[
\sigma_{VTH} \approx 3.19 \times 10^{-4} \times \frac{T_{OX} \times N_{DEP}^{eff}}{\sqrt{L_{INT} \times W_{eff}}}
\]  

(6)

\[
L_{eff} = L_{DIBL} + X_{L} + 2 \times L_{INT}
\]  

(7)

\[
W_{eff} = W_{DIBL} + X_{W} + 2 \times W_{INT}
\]  

(8)

Here, \( L_{eff} \) and \( W_{eff} \) are the effective length and width of the transistor's channel, \( T_{OX} \) is the oxide thickness, \( N_{DEP} \) is the channel doping concentration at depletion edge for zero body bias, \( X_{L} \) is the channel length offset (due to mask/etch effect), \( L_{INT} \) is the channel length offset parameter, \( X_{W} \) is the channel width offset (due to mask/etch effect), \( W_{INT} \) 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 \( V_{th} = 0 \), we could estimate \( \mu_{VTH} \), as

\[
\mu_{VTH} = V_{TH} + V_{TH, DBL}
\]  

(9)

where

\[
V_{TH, DBL} = \frac{\eta_{BS} \times V_{DS}}{2 \times (\cosh(D_{SUB} \times L_{eff}/L_0) - 1)}
\]  

(10)

\[
L_0 = \frac{\left(\eta_{BS} \times T_{OX} \times X_{DEP}\right)}{E_{OX}}
\]  

(11)

\[
X_{DEP} = \frac{2 \times \eta_{BS} \times \phi}{q \times N_{DEP}}
\]  

(12)

\[
\phi = \frac{2kT}{q} \times \ln\left(N_{DEP}/n_i\right)
\]  

(13)

Here \( V_{TH, DBL} \) is the long channel threshold voltage at \( V_{th} = 0 \), \( \eta_{BS} \) is the DBL (drain induced barrier lowering) coefficient in sub-threshold, \( D_{SUB} \) is the DBL coefficient exponent in sub-threshold, \( L_0 \) is the characteristic length, \( \eta_{BS} \) is permittivity of silicon, \( T_{OX} \) is the oxide thickness, \( X_{DEP} \) is the depletion width, \( E_{OX} \) is the permittivity of the oxide, \( \phi \) is the surface potential, \( q \) is the electron charge, \( T \) is the temperature, \( N_{DEP} \) 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, \( V_{TH} \) can be calculated much more accurately as

\[
\mu_{VTH} = V_{TH} + V_{TH, BE} - V_{TH, CS} - V_{TH, DBL} + V_{TH, RSC}
\]

+ \(V_{TH, NW} - V_{TH, SS} - V_{TH, PH} - V_{TH, PH, em}
\]  

(14)

For full details of (14) see BSIM4 v4.8 level 54. Equation (14) is used in this paper for calculating \( V_{TH} \) as precisely as possible.

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. \( L_{MIN} = L_{pMOS} = L_{nMOS} \)), 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 \( V_{TH} = v \) for a transistor can be calculated as [66]...
Fig. 3 shows that a slight increase of $L_{\text{drawn}}$ above minimum ($L_{\text{min}} = 22 \, \text{nm}$) causes a steep modification of $\mu_{\text{TH}}$. It also shows a reduction of $\sigma_{\text{TH}}$ as $L_{\text{drawn}}$ increases ($\pm \sigma_{\text{TH}}$, given by (6) are shown as bars). On the other hand, increasing $W_{\text{drawn}}$ above minimum ($W_{\text{min}} = 22 \, \text{nm}$) reduces $\sigma_{\text{TH}}$, but it does not affect $\mu_{\text{TH}}$.

Fig. 4 shows the effect of increasing $W_{\text{drawn}}$ and $L_{\text{drawn}}$ on the $V_{\text{TH}}$ distribution of the nMOS transistors. The figure reveals several important issues. First, the $V_{\text{TH}}$ distribution is not balanced. The probability of having $V_{\text{TH}} = 0$ is much higher than the probability of having $V_{\text{TH}} = 0.8$ (nominal voltage). Increasing $W_{\text{drawn}}$ from 22 to 44 nm have significantly reduces $P(V_{\text{TH}} = 0.8)$ further. However, it has a very modest effect on $P(V_{\text{TH}} = 0)$. On the other hand, increasing $L_{\text{drawn}}$ significantly reduces $P(V_{\text{TH}} = 0)$ while have almost no effect on $P(V_{\text{TH}} = 0.8)$. 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 $V_{\text{TH}}$ 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

$$P_{\text{fN}} = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-t^2/2} dt$$

(15)

While $P_{\text{fN}}$ (high) can be calculated as

$$P_{\text{fN}}(\text{high}) = 0.5 \times \frac{V_{\text{IN}} - V_{\text{NOISE}} - \mu_{\text{TH}}}{\sigma_{\text{TH}}} \times \frac{1 + \text{erf}\left(\frac{V_{\text{IN}} + V_{\text{NOISE}} - \mu_{\text{TH}}}{\sigma_{\text{TH}} \times \sqrt{2}}\right)}$$

(16)

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

$$P_{\text{fP}}(\text{low}) = 0.5 \times \left(1 + \text{erf}\left(\frac{-V_{\text{DD}} + (V_{\text{IN}} + V_{\text{NOISE}} + \mu_{\text{TH}})}{\sigma_{\text{TH}} \times \sqrt{2}}\right)\right)$$

(18)

$$P_{\text{fP}}(\text{high}) = 0.5 \times \text{erfc}\left(\frac{-V_{\text{DD}} + (V_{\text{IN}} - V_{\text{NOISE}} + \mu_{\text{TH}})}{\sigma_{\text{TH}} \times \sqrt{2}}\right)$$

(19)

Using (5) in conjunction with (6) and (14–19), the probability that a transistor exhibits a switching error (PF) was estimated for the $V_{\text{IN}}$ 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 $V_{\text{NOISE}} = 50\, \text{mV}$. 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 × 20 nm], when logic low is applied at the gate terminal ($V_{\text{IN}} = 0 \pm 50\, \text{mV}$), 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 ($V_{\text{IN}} = V_{\text{DD}} = -50\, \text{mV}$), 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 $V_{\text{TH}}$ 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 $V_{\text{TH}}(\mu_{\text{VTH}})$. It shows that increasing $W_{\text{drawn}}$ has no effect on $\mu_{\text{VTH}}$ (the line drawn at the bottom of the figure). Table 1 shows that increasing $L_{\text{drawn}}$ from 22 to 44 nm, and then to 66 significantly reduces $V_{\text{TH}}$ variations ($\sigma_{\text{TH}}$) and hence the switching PF. Increasing $W_{\text{drawn}}$ from 22 to 44 nm reduces $\sigma_{\text{TH}}$ from 0.101 to 0.060 V, while $\mu_{\text{VTH}}$ remains unchanged at 0.303 V. This consequently reduces $P_{\text{fN}}$ to $1.28 \times 10^{-5}$, while $P_{\text{fP}}$ 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

Fig. 4 $P(V_{\text{TH}})$ versus sizing:

(a) Varying $W$ ($L = \text{Lmin}$), (b) Varying $L$ ($W = \text{Wmin}$), (c) Varying $W$ ($L = \text{Lopt}$)
Obviously, shows that slightly increasing while PF orders of magnitude). PF switching PF as a function of PTM HP v2.1 have identified i. Making the switching probabilities of failure equal at logic low and logic high for each and every transistor.

ii. Reducing \( \sigma_{vth} \) while also aiming to make \( \text{PF}_{\text{nMOS}} \approx \text{PF}_{\text{pMOS}} \), i.e. balancing the nMOS and pMOS transistors as trying to make \( \sigma_{vth,\text{nMOS}} \approx \sigma_{vth,\text{pMOS}} \).

The first step means that \( V_{\text{TH}} \) of both transistors should be as close as possible to \( V_{\text{DD}}/2 \), i.e. symmetrical and as far as possible from both \( V_{\text{DD}} \) and GND (i.e. \( V_{\text{TH}} = |V_{\text{TH}}| = V_{\text{DD}}/2 \)). This equalises (balances) the PF of a given transistor at GND and \( V_{\text{DD}} \). Obviously, \( V_{\text{TH}} \) can be modified by tuning \( L \) as shown in (14) and Table 1. Equation (14) is used to determine the \( L_{\text{eff}} \) for which \( V_{\text{TH}} = V_{\text{DD}}/2 \). \( L_{\text{drawn}} \) for these ‘tuned’ transistors are then calculated using (7). We call \( L_{\text{drawn}} \), for which \( V_{\text{TH}} = V_{\text{DD}}/2 \), the optimum channel length \( (L_{\text{opt}}) \). Extensive simulations for 22 nm PTM HP v2.1 have identified \( L_{\text{opt, } \text{nMOS}} = 25.3 \text{nm} \), and \( L_{\text{opt, } \text{pMOS}} = 29.8 \text{nm} \) [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 \( L_{\text{drawn}} \) and \( W_{\text{drawn}} \) on the switching PF as a function of \( V_{\text{NOISE}} \). Fig. 5a shows the PF when the minimum transistor size is used (i.e. \( 22 \times 22 \) nm). Fig. 5b shows the PF of the transistors when using \( L_{\text{opt}} \) and \( W_{\text{MIN}} \). It shows that slightly increasing \( L \) from \( L_{\text{MIN}} \) to \( L_{\text{opt}} \) reduces \( \text{PF}_{\text{pMOS}} \) (low) by more than two orders of magnitude to \( 2.65 \times 10^{-5} \), while \( \text{PF}_{\text{pMOS}} \) (high) is reduced to \( 4.0 \times 10^{-7} \) (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 \( \sigma_{vth} \) and at the same time balancing the nMOS and pMOS transistors (i.e. \( \text{PF}_{\text{nMOS}} \approx \text{PF}_{\text{pMOS}} \)). Reducing \( \sigma_{vth} \) 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 \( L_{\text{opt}} \) 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 \( W \). Increasing the \( W_{\text{MIN}} \) from 22 to 27.9 nm reduces \( \sigma_{vth} \) from 0.0864 to 0.07075, which \( \approx \sigma_{vth} \) (0.0708). Consequently reduces \( \text{PF}_{\text{nMOS}} \) further to \( 4.0 \times 10^{-7} \) (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.

### Table 1 Effect of sizing on switching PF of bulk CMOS transistors, assuming \( V_{\text{NOISE}} = 50 \text{mV} \)

<table>
<thead>
<tr>
<th>Size (W \times L), nm</th>
<th>nMOS</th>
<th>pMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_{\text{TH}} )</td>
<td>( \sigma_{\text{vth}} )</td>
</tr>
<tr>
<td>22 \times 22</td>
<td>0.303</td>
<td>0.101</td>
</tr>
<tr>
<td>44 \times 2</td>
<td>0.303</td>
<td>0.06</td>
</tr>
<tr>
<td>66 \times 22</td>
<td>0.303</td>
<td>0.047</td>
</tr>
<tr>
<td>22 \times 44</td>
<td>0.488</td>
<td>0.054</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{PF} ) (low)</th>
<th>( \text{PF} ) (high)</th>
<th>( \text{PF} ) (low)</th>
<th>( \text{PF} ) (high)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 6.20 \times 10^{-3} )</td>
<td>( 1.28 \times 10^{-5} )</td>
<td>( 3.30 \times 10^{-6} )</td>
<td>( 4.44 \times 10^{-16} )</td>
</tr>
<tr>
<td>( 4.74 \times 10^{-6} )</td>
<td>( 4.51 \times 10^{-14} )</td>
<td>( 5.50 \times 10^{-22} )</td>
<td>( 7.37 \times 10^{-7} )</td>
</tr>
</tbody>
</table>

Fig. 5 Switching PF for CMOS transistors \((W \times L)\):

(a) Minimum (22 \times 22 nm), (b) Optimal (nMOS: 22 \times 25.3 nm, pMOS: 22 \times 29.8 nm), (c) Balanced (nMOS: 27.9 \times 25.3, pMOS: 22 \times 29.8 nm)

nm will significantly reduce down to \( 4.44 \times 10^{-16} \), hence making 7.37 \times 10^{-7} the worst case condition. This means that the worst case condition when \( L_{\text{drawn}} \) is increased is more than two orders of magnitude better than when \( L_{\text{drawn}} \) 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 \( \sigma_{vth} \) while also aiming to make \( \text{PF}_{\text{nMOS}} \approx \text{PF}_{\text{pMOS}} \), i.e. balancing the nMOS and pMOS transistors as trying to make \( \sigma_{vth,\text{nMOS}} \approx \sigma_{vth,\text{pMOS}} \).

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 \( V_{\text{DD}}/2 \). 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 \( V_{\text{DD}}/2 \) 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 \( V_{\text{DD}}/2 \) 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 \( V_{\text{DD}}/2 \). Table 2
Table 2 CLS and OPT sizing (W×L) in nm, and their effects on $V_{TH}$ and $\sigma_{VTH}$ of CMOS gates

<table>
<thead>
<tr>
<th></th>
<th>W×L</th>
<th>$V_{TH}$</th>
<th>$\sigma_{VTH}$</th>
<th>W×L</th>
<th>$V_{TH}$</th>
<th>$\sigma_{VTH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>P1</td>
<td>66×22</td>
<td>−235</td>
<td>44.8</td>
<td>196×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>N1</td>
<td>44×22</td>
<td>303</td>
<td>60</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td>NAND-2</td>
<td>P1</td>
<td>55×22</td>
<td>−235</td>
<td>50</td>
<td>183×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>N1</td>
<td>44×22</td>
<td>303</td>
<td>60</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td>NOR-2</td>
<td>P1</td>
<td>96×22</td>
<td>−235</td>
<td>36.2</td>
<td>220×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>N1</td>
<td>44×22</td>
<td>303</td>
<td>60</td>
<td>44×25.3</td>
<td>399</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>W×L</th>
<th>$V_{TH}$</th>
<th>$\sigma_{VTH}$</th>
<th>W×L</th>
<th>$V_{TH}$</th>
<th>$\sigma_{VTH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-INV</td>
<td>P1</td>
<td>67×22</td>
<td>−235</td>
<td>37.5</td>
<td>188×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>66×22</td>
<td>−235</td>
<td>44.8</td>
<td>196×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>66×22</td>
<td>−235</td>
<td>44.8</td>
<td>196×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>55×22</td>
<td>−235</td>
<td>55</td>
<td>183×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>P5</td>
<td>55×22</td>
<td>−235</td>
<td>55</td>
<td>183×29.8</td>
<td>−400</td>
</tr>
<tr>
<td></td>
<td>N1</td>
<td>44×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>44×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td>ST-NAND2</td>
<td>P1</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>P5</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>N1</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>303×22</td>
<td>399</td>
<td>51.3</td>
<td>44×25.3</td>
<td>399</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$V_{DD}$</th>
<th>CLS</th>
<th>OPT</th>
<th>ST</th>
<th>OPT</th>
<th>CLS</th>
<th>OPT</th>
<th>ST</th>
<th>OPT</th>
<th>CLS</th>
<th>OPT</th>
<th>ST</th>
<th>OPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>23.0</td>
<td>30.0</td>
<td>39.5</td>
<td>44.0</td>
<td>22.5</td>
<td>29.0</td>
<td>39.5</td>
<td>44.0</td>
<td>22.5</td>
<td>29.0</td>
<td>39.5</td>
<td>44.0</td>
</tr>
<tr>
<td>300</td>
<td>27.7</td>
<td>34.7</td>
<td>48.3</td>
<td>52.0</td>
<td>27.3</td>
<td>34.3</td>
<td>47.7</td>
<td>51.7</td>
<td>27.7</td>
<td>34.3</td>
<td>47.7</td>
<td>34.3</td>
</tr>
<tr>
<td>400</td>
<td>29.8</td>
<td>37.3</td>
<td>53.0</td>
<td>56.0</td>
<td>29.8</td>
<td>37.0</td>
<td>51.8</td>
<td>55.3</td>
<td>30.0</td>
<td>36.8</td>
<td>51.0</td>
<td>55.3</td>
</tr>
<tr>
<td>500</td>
<td>30.8</td>
<td>38.4</td>
<td>55.6</td>
<td>58.4</td>
<td>30.8</td>
<td>38.2</td>
<td>54.2</td>
<td>57.4</td>
<td>31.0</td>
<td>38.2</td>
<td>54.0</td>
<td>56.8</td>
</tr>
<tr>
<td>600</td>
<td>31.2</td>
<td>39.2</td>
<td>57.0</td>
<td>59.8</td>
<td>31.2</td>
<td>39.0</td>
<td>55.7</td>
<td>58.6</td>
<td>31.3</td>
<td>38.8</td>
<td>56.2</td>
<td>59.0</td>
</tr>
<tr>
<td>700</td>
<td>30.9</td>
<td>39.4</td>
<td>58.0</td>
<td>60.7</td>
<td>30.9</td>
<td>39.1</td>
<td>56.7</td>
<td>59.7</td>
<td>31.0</td>
<td>39.1</td>
<td>57.3</td>
<td>60.4</td>
</tr>
<tr>
<td>800</td>
<td>29.5</td>
<td>39.3</td>
<td>58.5</td>
<td>61.3</td>
<td>29.4</td>
<td>38.9</td>
<td>57.3</td>
<td>60.5</td>
<td>29.8</td>
<td>38.8</td>
<td>56.0</td>
<td>61.4</td>
</tr>
</tbody>
</table>

shows the CLS and OPT sizing and their effect on $V_{TH}$ and $V_{TH}$ 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, which 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 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 measured 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 ($nPWR_{i,j}$) is calculated as

$$nPWR_{i,j} = PWR_{j}/PWR_{j}^{*}$$

where $PWR_{j}^{*}$ is the base power of gate j.
shows that the CLS design method exhibits the best normalised design and operating voltage that has the best performance (i.e. using the normalised values, it is much easier to identify the delay (nDLY) compared to the other designs, whereas the shortest and 800 mV in the case of SNM). On the other hand, the bold 

Fig. 6 illustrates the normalised values for the NAND-2 gate. It shows that the CLS design method exhibits the best normalised delay (nDLY) compared to the other designs, whereas the shortest delay occurs at nominal V_DD. Using either the ST or the OPT method increases the delay significantly. On the other hand, in the case of power dissipation, Fig. 6a shows that the OPT method has the lowest power dissipation (i.e. best nPWR) followed by the ST-OPT regardless of the operating voltage.

Fig. 6 also shows that the best design method with respect to PDP (energy) depends on the operating voltage. While the CLS design has the best PDP performance in subthreshold (200 mV), the OPT design significantly outperformed the other three methods for V_DD > 400 mV. Fig. 6 also shows that the best PDP performance occurs at different V_DD for different design methods. In the case of CLS method, the best PDP occurs at V_DD = 540mV, while it occurs at 550, 680, and 700 mV for the ST-CLS, OPT, and ST-OPT methods, respectively.

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-CLS 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.
Fig. 7 Figure of merits for NAND-2 gate
(a) FoM1; (b) FoM2; (c) FoM3; (d) FoM4

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

<table>
<thead>
<tr>
<th>V_DD</th>
<th>CLS</th>
<th>OAS</th>
<th>ST</th>
<th>STOPT</th>
<th>CLS</th>
<th>OAS</th>
<th>ST</th>
<th>STOPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.251</td>
<td>0.496</td>
<td>0.311</td>
<td>0.451</td>
<td>0.310</td>
<td>0.496</td>
<td>0.344</td>
<td>0.560</td>
</tr>
<tr>
<td>300</td>
<td>0.215</td>
<td>0.378</td>
<td>0.333</td>
<td>0.420</td>
<td>0.237</td>
<td>0.471</td>
<td>0.345</td>
<td>0.474</td>
</tr>
<tr>
<td>400</td>
<td>0.220</td>
<td>0.328</td>
<td>0.356</td>
<td>0.412</td>
<td>0.220</td>
<td>0.376</td>
<td>0.358</td>
<td>0.442</td>
</tr>
<tr>
<td>500</td>
<td>0.268</td>
<td>0.313</td>
<td>0.375</td>
<td>0.411</td>
<td>0.230</td>
<td>0.333</td>
<td>0.370</td>
<td>0.427</td>
</tr>
<tr>
<td>600</td>
<td>0.350</td>
<td>0.330</td>
<td>0.397</td>
<td>0.415</td>
<td>0.255</td>
<td>0.316</td>
<td>0.382</td>
<td>0.420</td>
</tr>
<tr>
<td>700</td>
<td>0.429</td>
<td>0.358</td>
<td>0.418</td>
<td>0.423</td>
<td>0.278</td>
<td>0.312</td>
<td>0.392</td>
<td>0.418</td>
</tr>
<tr>
<td>800</td>
<td>0.493</td>
<td>0.384</td>
<td>0.435</td>
<td>0.430</td>
<td>0.293</td>
<td>0.311</td>
<td>0.400</td>
<td>0.418</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V_DD</th>
<th>CLS</th>
<th>OAS</th>
<th>ST</th>
<th>STOPT</th>
<th>CLS</th>
<th>OAS</th>
<th>ST</th>
<th>STOPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.240</td>
<td>0.488</td>
<td>0.287</td>
<td>0.429</td>
<td>0.302</td>
<td>0.688</td>
<td>0.325</td>
<td>0.552</td>
</tr>
<tr>
<td>300</td>
<td>0.214</td>
<td>0.369</td>
<td>0.309</td>
<td>0.401</td>
<td>0.235</td>
<td>0.465</td>
<td>0.323</td>
<td>0.461</td>
</tr>
<tr>
<td>400</td>
<td>0.218</td>
<td>0.321</td>
<td>0.342</td>
<td>0.399</td>
<td>0.219</td>
<td>0.371</td>
<td>0.346</td>
<td>0.431</td>
</tr>
<tr>
<td>500</td>
<td>0.255</td>
<td>0.308</td>
<td>0.364</td>
<td>0.402</td>
<td>0.227</td>
<td>0.328</td>
<td>0.360</td>
<td>0.419</td>
</tr>
<tr>
<td>600</td>
<td>0.331</td>
<td>0.323</td>
<td>0.388</td>
<td>0.410</td>
<td>0.249</td>
<td>0.312</td>
<td>0.375</td>
<td>0.415</td>
</tr>
<tr>
<td>700</td>
<td>0.416</td>
<td>0.353</td>
<td>0.410</td>
<td>0.421</td>
<td>0.274</td>
<td>0.308</td>
<td>0.386</td>
<td>0.416</td>
</tr>
<tr>
<td>800</td>
<td>0.494</td>
<td>0.383</td>
<td>0.429</td>
<td>0.429</td>
<td>0.295</td>
<td>0.309</td>
<td>0.396</td>
<td>0.416</td>
</tr>
</tbody>
</table>

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/)