

# Automating the sizing of transistors in CMOS gates for low-power and high-noise margin operation

Azam Beg<sup>\*,†</sup>

*College of Information Technology, UAE University, United Arab Emirates*

## SUMMARY

This paper presents an automatic method for sizing the transistors in CMOS gates. The method utilizes a feedback control system to efficiently optimize the transistor sizes in small and large fan-in gates, with the primary goal of enhancing noise robustness (as characterized by the static noise margin). The gates retain their robustness under threshold-voltage variations over a range of supply voltages. The optimized gates not only expend reduced power and energy, but also take up less area than the conventional ones. These multifaceted gains, however, do incur some performance loss. Copyright © 2014 John Wiley & Sons, Ltd.

Received 24 December 2013; Revised 4 September 2014; Accepted 5 September 2014

**KEY WORDS:** CMOS; logic gates; transistor sizing; static noise margin; power dissipation; energy consumption; PID feedback control

## 1. INTRODUCTION

The semiconductor industry has been able to sustain the Moore's law [1] for the past several decades. Along the way, the industry had to overcome many obstacles. The recent International Technology Roadmap for Semiconductors (ITRS) recognizes *power* as one of the major challenges for current and forthcoming VLSI designs [2–4]. An additional challenge is the *reliability*, which includes tolerance to variations and endurance to intrinsic and extrinsic noises.

Power consumption in CMOS circuits consists of two components: *static* and *dynamic*. Static power consumption occurs due to static conducting paths between the power supply ( $V_{DD}$ ) and the ground (GND), and due to leakage currents. The dynamic power consumption happens during switching due to temporary current paths between  $V_{DD}$  and GND, and because of charging of capacitors [3].

In the past, numerous techniques have been used for reducing different components of power. Some of the well-known techniques are device optimization [5], use of multiple threshold voltages ( $V_{TH}$ ) for the devices [6], multiple  $V_{DD}$ 's [7], dynamically scaled  $V_{DD}$  [8], selective power supply shutdown, adaptive biasing of substrate [9–12], near or sub-threshold voltage ( $V_{TH}$ ) operation [13]–[15], and the reduction in clock frequency [3].

The dynamic power has a quadratic relationship to the  $V_{DD}$ , while the leakage power is linearly related to  $V_{DD}$ . So the most obvious and the simplest way of power reduction would be to reduce  $V_{DD}$ . But the downside is that the lower  $V_{DD}$  degrades performance and makes the circuit highly prone to manufacturing variations and to noise [16–18].

*Static noise margin* (SNM) [16], [19], [20] is one of the metrics for assessing the tolerance of a circuit to noise and variations. Using SNM as a metric has been more common in case of static

\*Correspondence to: Azam Beg, College of Information Technology, UAE University, United Arab Emirates.

†E-mail: abeg@uaeu.ac.ae

RAM (SRAM) as compared to basic logic gates [21]. (Refer to Appendix A for a brief explanation of SNM; further details can be found in most books on digital circuit design).

This paper builds up on the idea of using transistor sizing as a method of increasing SNM, proposed by us in [22], where the feasibility of the technique was demonstrated for simple gates: INV, NAND2, and NOR2. For large fan-in gates, the need for systemizing the process was pointed out because of the tedious and manual nature of the sizing methodology. Here, we will introduce a novel, automatic system for sizing the multi-input gates with the purpose of enhancing noise robustness while reducing the power and energy consumption.

This paper is organized as follows: Previous work on MOS transistor length extension and related analyses are presented in Section 2. This is followed by a discussion of  $V_{TH}$  variations in the transistors in Section 3. Issues related to the SNM are looked into in Section 4. Our scheme for automatic sizing of transistors in different gates is explained in Section 5. The conclusions and the directions for future research are given in Section 6.

## 2. EXTENDING THE TRANSISTOR CHANNEL LENGTH

The techniques for sizing the transistors in CMOS gates have been well established; however, there are instances when the techniques have been revisited for the sake of reducing power or for boosting reliability. Upsizing the transistor channel from its traditional minimum has also been proposed in [21–26].

With the primary aim of performance maximization, VLSI designers conventionally set the channel lengths ( $L$ 's) of the nMOS and pMOS transistors to the minimum (i.e.  $L_{nMOS} = L_{pMOS} = L_{min}$ ) and then increase the channel widths ( $W_{nMOS}$  and  $W_{pMOS}$ ) to equalize the rising and falling transition times. However, the sustained scaling of the transistor dimensions has resulted in higher leakage power and device variations; these factors have prompted the researchers to break away from the tradition of setting the channel lengths to  $L_{min}$ .

Gupta *et al.* [27] proposed that the  $L$ 's be selectively increased by nearly 10% to minimize the leakage power. Counter-intuitively, in [28], it was shown that some  $L > L_{min}$  can minimize the delay. The idea of upsized- $L$  along with an optimized doping profile was brought forth for sub- $V_{TH}$  operation of SRAMs [29]. In general, incremental changes in  $L$  proved to be a simple cure for the static-energy problems caused by the scaling trends [30]. Actually, the extension of the  $L$  by just a few nanometers resulted in two orders of magnitude reduction in energy. Tajalli and Leblebici [31] reached similar conclusions but for the system level performance. They emphasized the need for careful selection of  $V_{DD}$  and transistor upsizing for minimizing energy consumption. Gupta and Ghosh [32] used ant colony for optimizing transistor sizes of (a rather dated) technology node of 180 nm.

Alioto [33] studied the effect of different types of variations on SNMs. His analysis of voltage, temperature, and process showed that their variations have a very adverse effect on the gates operating in sub- $V_{TH}$  regime. Similar findings were reported in [34], wherein a six-transistor SRAM cell failed to operate at sub- $V_{TH}$  due to curtailed SNM and other variations. In [31], the relationship of process parameters (i.e. drain-induced barrier lowering/DIBL,  $V_{DD}$ , and sub- $V_{TH}$  and slope factor) to power consumption and reliability was highlighted.

At low operating voltages, there can be considerable imbalances of nMOS and pMOS transistors, resulting in reduced noise margin [8]. One way of offsetting the effect of the imbalance is to increase the channel widths, either conventionally, or by using multiple minimum-sized transistors. The latter technique would result in sufficient drive while lowering capacitance and the area.

Calhoun *et al.* [8] and Blesken *et al.* [35] pointed out that for optimal operation, the gates operating at nominal  $V_{DD}$  need to be sized differently for the sub- $V_{TH}$  operation. However, in [21], it was shown that SNM could be nearly maximized by fine-tuning the crossover points to the proximity of  $V_{DD}/2$ . Carefully selected (to be explained shortly) channel lengths ( $L_{nMOS}$  and  $L_{pMOS}$ ) make it possible to attain sufficient SNM even in sub- $V_{TH}$  region.

In order to have a closer look at the impact of scaling on delay and SNM, we utilize an INV, and sweep the  $L$ 's of nMOS and pMOS transistors:  $L_{min} < L < 1.5 \times L_{min}$ , and  $\Delta L = 1$  nm. For each

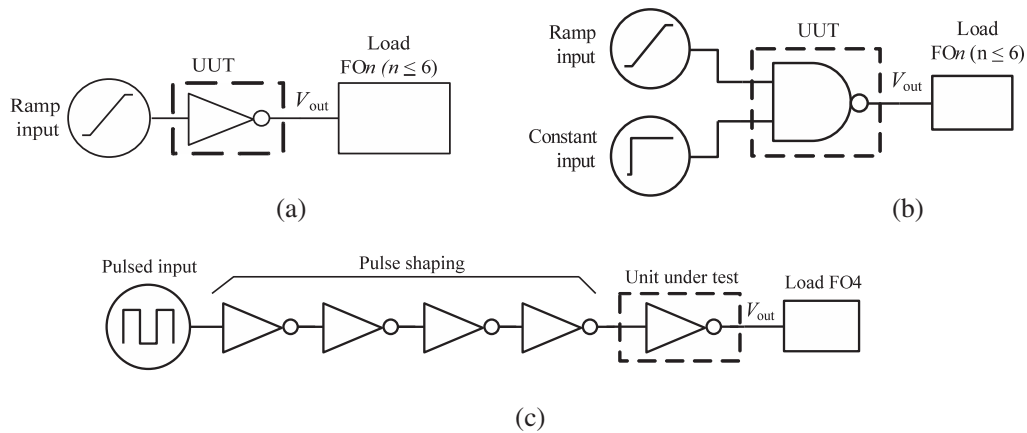


Figure 1. Test circuits for measuring (a) the SNM of an INV, (b) the SNM of a NAND2, and (c) the power and the delay of an INV.

combination of  $L_{nMOS}$  and  $L_{pMOS}$ , we balance the rise and the fall times [36] by adjusting the  $W_{pMOS}$ ;  $W_{nMOS}$  is fixed at 44 nm [37].

The test circuit for measuring the SNMs of an INV is shown in Figure 1(a). (UUT stands for *unit-under-test*). A simple ramp input is needed for an INV. Larger fan-in gates, such as NAND2 of Figure 1(b), need different but relevant combinations of ramp and constant inputs [38–41]. All gates in this paper are based on 22-nm PTM HP v2.1 (high- $k$ /metal gate and stress effect) MOS transistor models [42], [43] and BSIM4v4.7 level 54 [44]. We have chosen the *fan-out* of 4 (FO-4) for all UUTs used for SNM measurement.

We used Spice (specifically, Ngspice [45]) to simulate the circuits. We employed a Matlab script [21] to derive the SNMs from the text-based simulation log files.

The surface plot of Figure 2(a) shows an INV's SNM relationship to different sets of  $L_{nMOS}$  and  $L_{pMOS}$ . It is evident that higher values of SNM can be achieved by elongating either one or both  $L$ 's ( $L_{nMOS}$ ,  $L_{pMOS}$ ). For the INV, an increase of 8–9 nm in  $L$  can deliver approximately 50% increase in SNM (over a design that uses  $L_{min}$ ).

Keeping in mind the fact that SNM is not the sole design criterion, and that power and performance (delay) also need to be taken into account. So we used the test setup of Figure 1(c) for power and delay measurement. In Figure 2(b), we observe that 2–3 nm extension of  $L$  keeps the performance penalty in check but provides limited gain in SNM. Therefore, if one is willing to pay price in terms of performance, it is better to increase the  $L$  by 6–7 nm. (Intel uses  $L=30$  nm for its '22-nm node' 3-D gates! [46]).

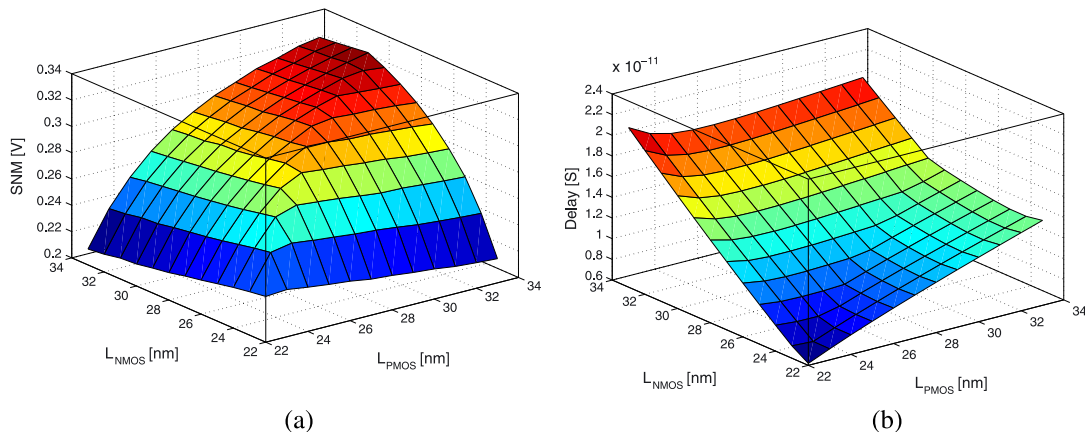


Figure 2. The effect of an INV's  $L$  ( $L_{nMOS}$  and  $L_{pMOS}$ ) on its (a) SNM and (b) delay [27].

It is worth noting that mere sweeping of  $L$  hardly constitutes a systematic approach for determining the optimum  $L_{\text{nMOS}}$  and  $L_{\text{pMOS}}$ ; therefore, we need a formal methodology for doing so (more on this in the next section).

### 3. $V_{\text{TH}}$ VARIATIONS IN NANOMETRIC TRANSISTORS

We utilize BSIM4v4.7 level 54 [44] to describe the  $V_{\text{TH}}$  of a MOS transistor:

$$V_{\text{TH}} = V_{\text{TH0}} - \frac{\eta_0 \times V_{\text{DS}}}{2 \times [\cosh(D_{\text{SUB}} \times L_{\text{eff}} / L_t) - 1]} \quad (1)$$

where:

$$L_{\text{eff}} = L_{\text{drawn}} + X_L - 2 \times L_{\text{INT}} \quad (2)$$

$$W_{\text{eff}} = W_{\text{drawn}} + X_W - 2 \times W_{\text{INT}} \quad (3)$$

$$L_t = \sqrt{\frac{\epsilon_{\text{Si}} \times T_{\text{ox}} \times X_{\text{dep}}}{\epsilon_{\text{ox}}}} \quad (4)$$

$$X_{\text{dep}} = \sqrt{\frac{2 \times \epsilon_{\text{Si}} \times \phi}{q \times N_{\text{dep}}}} \quad (5)$$

and

$$\phi = \frac{2kT}{q} \times \ln(N_{\text{dep}} / n_i) \quad (6)$$

Here  $V_{\text{TH0}}$  is the long channel threshold voltage at  $V_{\text{BS}}=0$ . (For the 22-nm node [42], nominal  $V_{\text{TH0\_nMOS}}=0.5031$  V, and nominal  $V_{\text{TH0\_pMOS}}=-0.4606$  V).  $\eta_0$  is the DIBL (drain induced barrier lowering) coefficient in sub-threshold,  $D_{\text{SUB}}$  is the DIBL coefficient exponent in sub-threshold,  $L_{\text{eff}}$  is the effective channel length,  $L_t$  is the characteristic length,  $L_{\text{drawn}}$  is the drawn length,  $X_L$  is the channel length offset,  $L_{\text{INT}}$  is the channel length offset parameter,  $W_{\text{eff}}$  is the effective channel width,  $W_{\text{drawn}}$  is the drawn width,  $X_W$  is the channel width offset,  $W_{\text{INT}}$  is the channel width offset parameter,  $\epsilon_{\text{Si}}$  is the 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.

It is known that basic MOS transistors' probabilistic differences depend on the transistor type (nMOS or pMOS), the transistor dimensions, and the input voltage levels [21]. The variation in  $V_{\text{TH}}$  due to random doping fluctuations can be estimated using the following equation [47]:

$$\sigma_{V_{\text{THRDF}}} \approx 3.19 \times 10^{-8} \times \frac{T_{\text{ox}} \times N_{\text{dep}}^{0.4}}{\sqrt{L_{\text{eff}} \times W}} \quad (7)$$

For an INV built from minimum-sized nMOS and pMOS transistors ( $W \times L = 22 \text{ nm} \times 22 \text{ nm}$ ), the  $V_{\text{TH}}$  of the transistors can be calculated using eq. (1), and the probability density function (PDF) of  $V_{\text{TH}}$  by [21]:

$$PDF_{V_{TH}}(V_{GS}) = \frac{\exp\left[-(V_{GS} - V_{TH})^2 / 2\sigma^2 V_{THRDF}\right]}{\sigma V_{THRDF} \sqrt{2\pi}} \quad (8)$$

The INV has non-zero probabilities for the transistors' switching errors at either GND or  $V_{DD}$  [21] (see Figure 3(a)). The four probabilities (for logic low and logic high inputs for each transistor) of switching errors can be reduced by: (i) equating the probabilities (balancing them at GND and  $V_{DD}$  for each transistor; ideally  $V_{TH} = V_{DD}/2$ ), and (ii) reducing  $\sigma_{V_{THRDF}}$  (at least theoretically). Eq. (1) shows that  $V_{TH}$  depends on  $L$ , so we need  $L_{eff}$  for which  $V_{TH} = V_{DD}/2$ . The resulting  $L$  would be called optimal- $L$  ( $L_{opt}$ ) and the transistor with  $L_{opt}$  would be called an *optimal- $L$*  transistor (or just *optimal transistor*). A *normal- $L$*  transistor (or simply a *normal transistor*) would have  $L = L_{min} = 22$  nm; for this node [42], the nMOS' optimal- $L = 24.9$  nm, and the pMOS' optimal- $L = 29.4$  nm (see Figure 3(b)). Such precise sub-nm tuning of transistors may be attainable with *optical proximity correction* (OPC) [48]. The next step would be to adjust transistors'  $W$ 's for reduction of  $\sigma_{V_{THRDF}}$  and for balancing rise and fall-times of the INV [21]. In  $V_{TH}$ -PDF plots of Figure 3(c), we can see the resultant shifts of  $\mu_{V_{TH\_nMOS}}$  and  $[\mu_{V_{TH\_pMOS}} + V_{DD}]$  from 0.322 V and 0.541 V, respectively, to  $V_{DD}/2 = 0.4$  V.

Even with OPC, it may not be feasible to manufacture MOS transistors with the exact  $L_{opt}$ 's identified above, so the experiments in the rest of this paper will use optimal lengths (rounded to nearest nm) of 25 nm (instead of 24.9 nm) and 29 nm (instead of 29.4 nm) for nMOS and pMOS transistors, respectively. Additionally, these *shifted-from-ideal*  $L$ 's can help avoid higher switching currents when input voltage crosses  $V_{DD}/2$ . This is to be reiterated here is that once selected, the same value of optimal- $L$  is to be used for all gates in a particular circuit.

#### 4. STATIC NOISE MARGIN OF LOGIC GATES

Use of scaling for making SRAMs noise- and variation-tolerant has been in vogue for quite some but mostly remained an untapped scheme for digital logic gates. A low SNM in SRAM can cause a bit flip, while a logic gate's glitched output may end up as a sampled error.

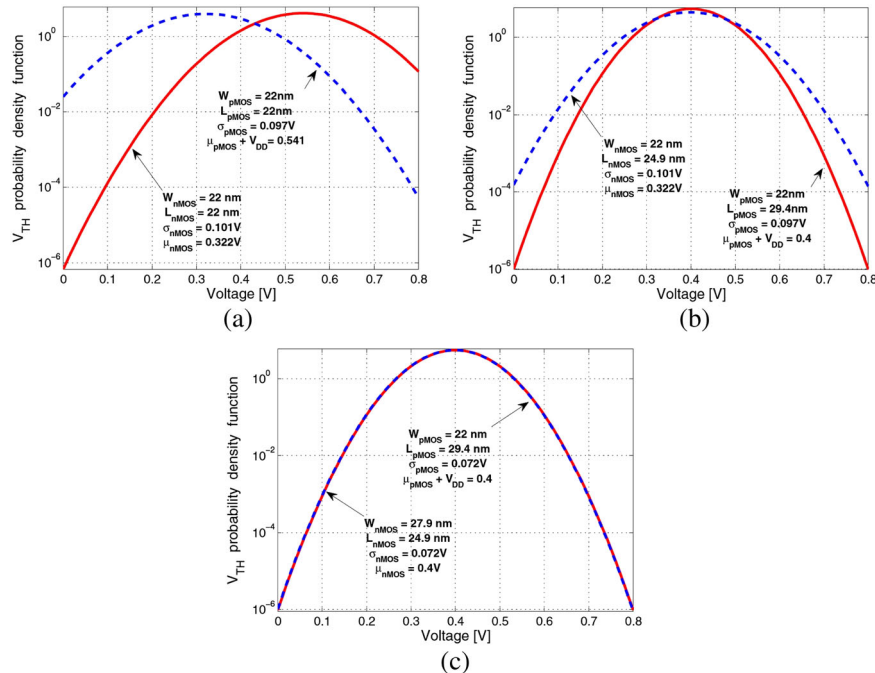


Figure 3. Exploiting sizing to balance the  $V_{TH}$ 's of nMOS and pMOS transistors: (a) minimum  $L$ 's and  $W$ 's; (b) optimal- $L$ 's and minimum  $W$ 's; and (c) optimal- $L$ 's and adjusted  $W_{pMOS}$  [14].

We built circuits for NAND2 and NOR2 using the transistor dimensions of Figure 4 and the test setup of Figure 1(b). Figures 5(a) and 5(b) show the *butterfly* curves ( $V_{in}$  vs.  $V_{out}$ , and  $V_{out}$  vs.  $V_{in}$ ) for a NAND2 gate with normal- and optimal- $L$ , respectively. For this gate, there are two instances in which a single input change causes an output transition:  $10 \rightarrow 11$ , and  $01 \rightarrow 11$ ; similarly, a NOR2 switches with these inputs:  $00 \rightarrow 01$ , and  $00 \rightarrow 10$ . Each set of input and output results in a different butterfly curve and hence a different SNM.

As a two-input gate results in four different output curves, the *worst* of the four SNMs needs to be found using:

- The largest allowed input voltage for logic LOW ( $V_{IL}$ )
- The smallest allowed input voltage for logic HIGH ( $V_{IH}$ )
- The largest output voltage for logic LOW ( $V_{OL}$ ), and
- The smallest output voltage for logic HIGH ( $V_{OH}$ ).

For fan-in of more than two, we must consider a larger set of input combinations. The combinations of constant and ramp inputs that cause an output to switch for a 5-input NAND gate (NAND5) are shown in Figure 6(a). Ideally, the output ( $V_{out}$ ) must crossover the ramp input (for example,  $V_{in0}$ ) at  $V_{DD}/2$ . But, in reality, some  $V_{out}$ 's end up above  $V_{DD}/2$  and others below; two such examples are shown in Figure 6(b).

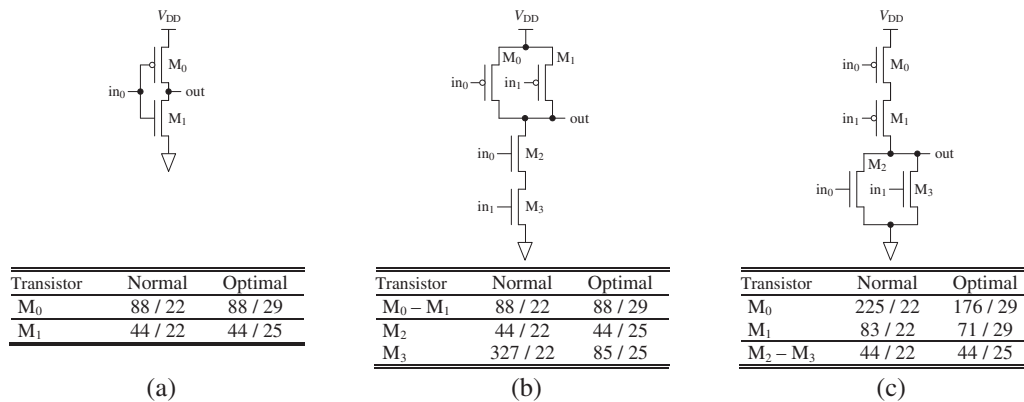


Figure 4. The schematics and the transistor dimensions of gates with normal- $L$  and optimal- $L$  (a) INV, (b) NAND2, and (c) NOR2.

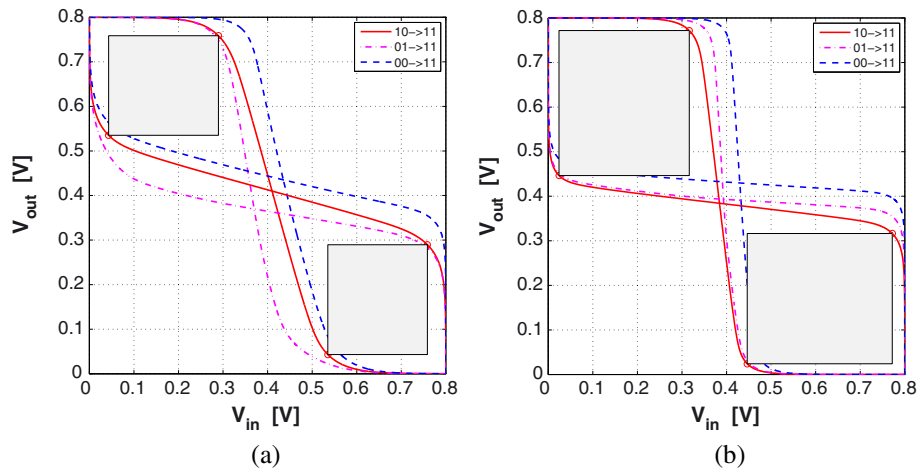


Figure 5. The SNMs of NAND2 gates with (a) normal- $L$ , and (b) optimal- $L$ .



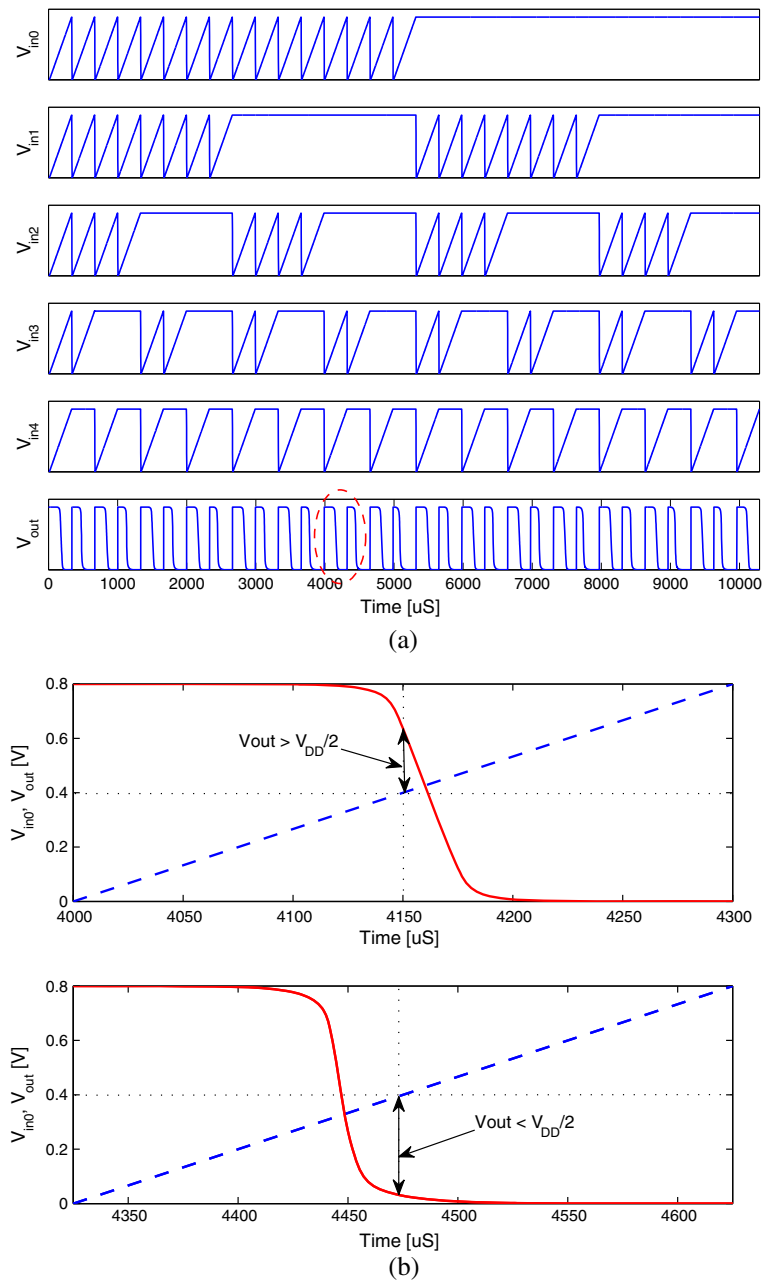


Figure 6. (a) Thirty-one different input vectors used for measuring the SNM of a NAND5 gate; (b) two sets of  $V_{in0}$  (blue/linear) and  $V_{out}$  (red/curve) (red-circle din (a)) that have significantly different rise and fall times, and hence the SNMs.

We compared the SNMs of optimal and normal gates operating at different  $V_{DD}$ 's. Decreasing both the  $V_{DD}$  and the  $V_{TH}$  can reduce significantly increase the leakage power and the  $V_{TH}$  variations. So we limit the  $V_{DD}$  range from 0.4 V to the nominal  $V_{DD}$  of 0.8 V. (For 22-nm PTM models, [42] specifies the nominal  $V_{DD}$  as 0.8 V). The optimal- $L$  gates consistently yield higher SNMs over the aforementioned range of  $V_{DD}$ . In Figure 7(a), we have plotted the *normalized* SNM (measured SNM/ $V_{DD}$ ) as a function of  $V_{DD}$ , for the three basic gates. When compared with normal- $L$ , the use of optimal- $L$  exhibits SNM enhancements of 11%–29% for the INV, 59%–64% for the NAND2, and 9%–19% for the NOR2.

Figure 7(b) provides another perspective into the merits of optimal- $L$  by using SNM vs. power plots. (Figure 1(c) shows the setup for measuring the power and delay of an INV-UUT; for a multi-input UUT, each input has its own series of waveform-shaping gates). For a given value of power, each

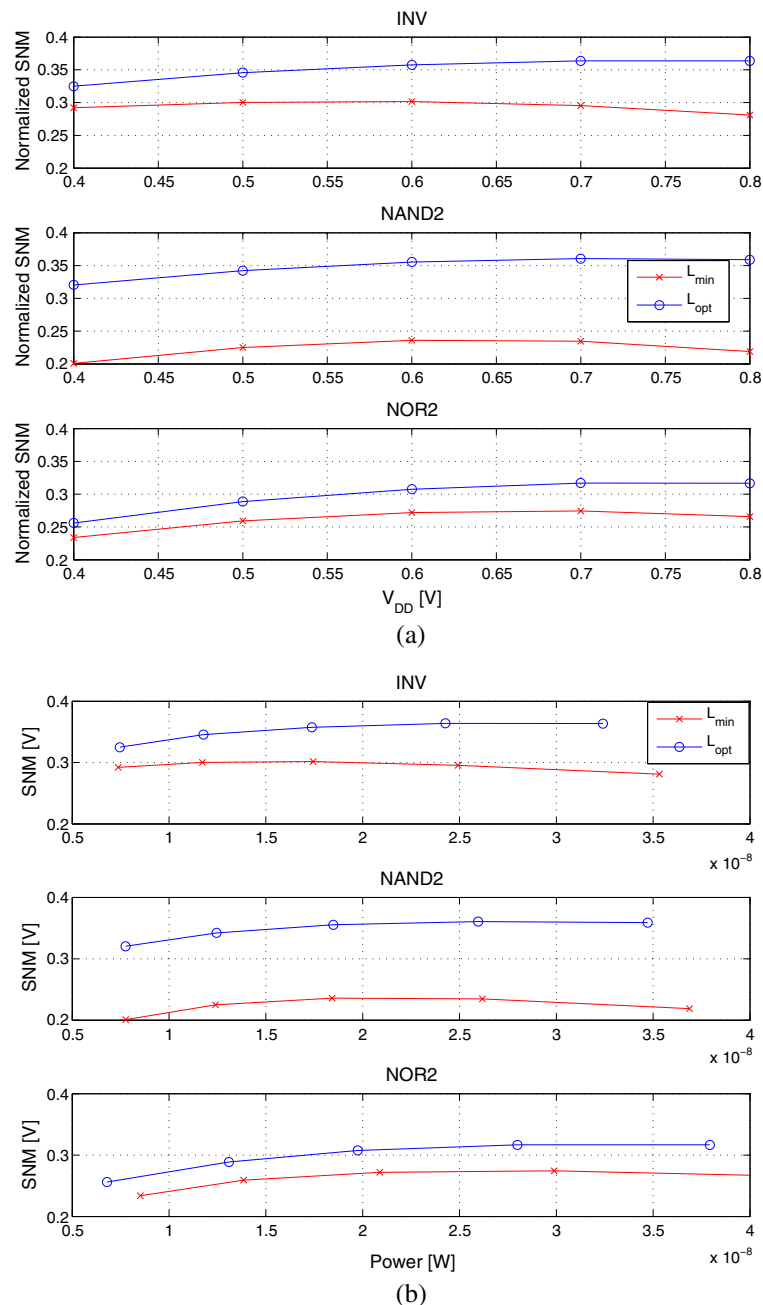


Figure 7. The comparison of normal- $L$  and optimal- $L$  basic gates: (a) normalized SNM (actual SNM/ $V_{DD}$ ) vs.  $V_{DD}$ , and (b) SNM vs. power.

gate with optimal- $L$  consistently provides better SNM than its normal- $L$  counterpart. The optimal- $L$  INV provides 11%–27% higher SNM than the normal one; NAND2's and NOR2's SNM gains are 60%–64%, and 15%–18%, respectively.

The noise robustness (again, in terms of SNM) of optimal gates was also compared with the normal- $L$  gates when the gates were subject to  $V_{TH}$  variations. The graphical results (histograms) of 1000 Monte Carlo (MC) simulations of the three gates operating at two different  $V_{DD}$ 's, 0.5 V and 0.8 V, are shown in Figure 8. The robustness of the optimal- $L$  gates (blue curves on the right) is quite evident. Numerically speaking, the average improvement in SNM for the optimal INV was 33.5% and 38.5% for  $V_{DD}$ 's of 0.5 V and 0.8 V, respectively. NAND2 showed a change of 49.3% and 47.9%, while NOR2 exhibited values of 32.5% and 37.3% (Table I). Furthermore, we defined a *failure* as a case when SNM < 20% of  $V_{DD}$ . The failure-counts



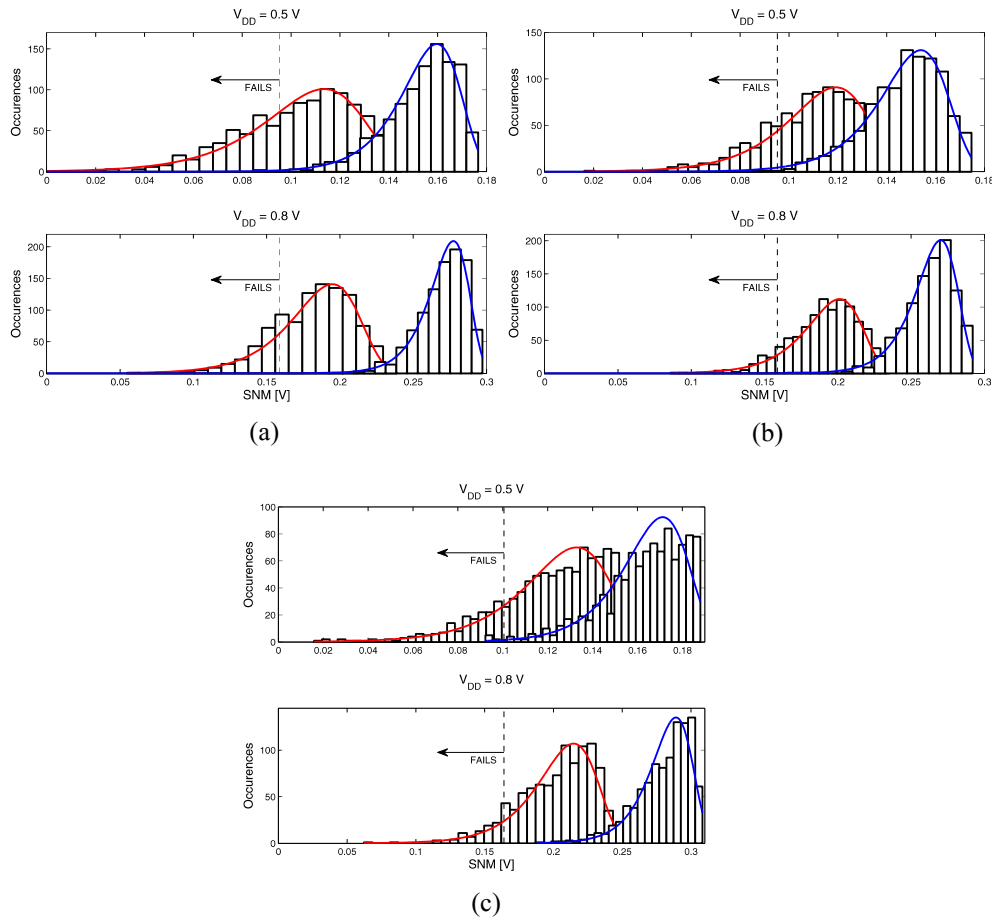


Figure 8. The effect of random variations of  $V_{TH}$  on SNM of normal (red curves on the left) and optimal gates (blue curves on the right): (a) INV, (b) NAND2, and (c) NOR2.

of optimal- $L$  gates were significantly lower than the normal- $L$  gates, as shown in Table II. For  $V_{DD}$ 's of 0.5 V and 0.8 V, INV's failures dropped from 179 to 7, and 74 to zero, respectively; NAND2's failures dropped from 405 to 8, and 211 to just 1; and NOR2 had its failures decrease from 279 to 12, and from 106 to only 2. Understandably, due to their random nature, the preceding statistics may vary from simulation-set to simulation-set, but they still give us a general comparison of the failure rates for the normal and the optimal gates.

## 5. THE PROPOSED SCHEME FOR TRANSISTOR SIZING

As mentioned in the last section, an ideal gate should have its ramp-input and output crossover at  $V_{DD}/2$ . Unfortunately, transistor sizing to achieve this balance is a non-trivial task for the gates with multiple inputs; the larger the fan-in, the harder it is to do the matching. We had identified the need for an automatic *sizing-for-balancing* mechanism in [21]. To fill this lacuna, we present a systematic approach for sizing the gate transistors.

Our proposed scheme for gate-transistor sizing utilizes a *progressive-sizing* approach [3]; the scheme iteratively finds the *progressive-sizing factor*  $K_{\text{prog}}$  (see Figure 9) using a fast PID (*proportional-integral-derivative*) controller [49] for sizing the pMOS and nMOS transistors in the gates. (Refer to Appendix B for a brief introduction to the PID control systems). The block diagram of our PID-controller for transistor sizing is shown in Figure 10.

Table I. The average SNM [V] for 1000 MC simulations with varying  $V_{TH}$ .

$V_{DD}$ [V]	INV			NAND2			NOR2		
	$L_{min}$	$L_{opt}$	Change	$L_{min}$	$L_{opt}$	Change	$L_{min}$	$L_{opt}$	Change
0.5	0.122	0.163	33.5%	0.102	0.152	49.3%	0.110	0.146	32.5%
0.8	0.202	0.280	38.5%	0.182	0.270	47.9%	0.191	0.262	17.3%

Table II. The failures for 1000 MC simulations with varying  $V_{TH}$ .

$V_{DD}$ [V]	INV		NAND2		NOR2	
	$L_{min}$	$L_{opt}$	$L_{min}$	$L_{opt}$	$L_{min}$	$L_{opt}$
0.5	179	7	405	8	279	12
0.8	74	0	211	1	106	2

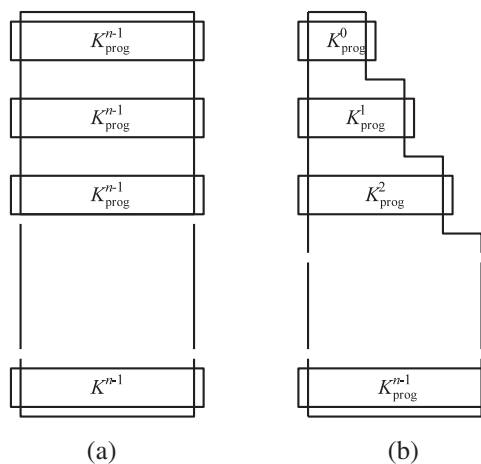


Figure 9. Transistor stacks: (a) with uniform sizing, and (b) with progressive sizing.

As it would be impossible to perfectly match the crossover points for all input combinations, we define a metric called *average output error* and aim to reduce it to a pre-defined value. If a gate has  $N$  input-sets that cause the output to change, we define the average output error as:

$$\varepsilon_{avg} = \sum_{i=1}^N \left( \frac{V_{DD}}{2} - V_{out_i} \right) / N \quad (9)$$

We used an arbitrary value of  $\varepsilon_{avg} < 0.1 \times V_{DD}$  for the examples that follow. Designers of a gate library may choose higher or lower values of  $\varepsilon_{avg}$ , based on other criteria, such as the gate area. For example, trying to reach  $\varepsilon_{avg} < 0.01 \times V_{DD}$  may increase the gate area beyond practical limits.

Initially, the transistors are uniformly sized as shown in Figure 9(a). As the PID controller iterates through the design process, the value of  $K_{prog}$  is altered (to become *progressive*) according to the preset PID parameters ( $K_P$ ,  $K_I$ , and  $K_D$ ). Each value of  $K_{prog}$  determines the dimensions of the top and the bottom transistors in the pMOS and the nMOS stacks. These values are used to determine the *progressive* dimensions of the remaining transistors in the two stacks. The Spice-netlist generation engine uses these dimensions to fully specify the UUT. Two different netlists are generated in every iteration of the feedback design loop, one for finding SNM (see Figure 1(a)), and the other for power and delay (see Figure 1(c)). For the SNM-netlist, the engine also defines the needed stimuli to ensure that all possible values of input combinations of ramps and constant inputs (for multi-input UUTs/gates) are included, because different input vectors result in different SNMs (as discussed earlier).

A post-simulation processor parses through the Spice simulation log files to find  $\varepsilon_{avg}$ . A Matlab script (mentioned earlier) is used to calculate the worst-case SNM.

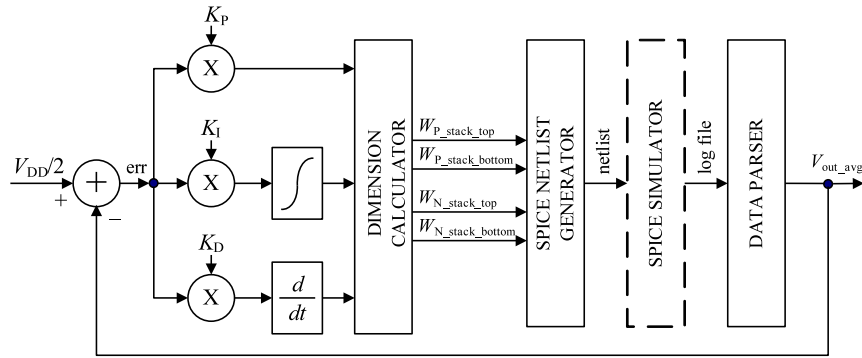


Figure 10. The PID-controller-based transistor sizing scheme for gates.

Due to the absence of a *known* model for the system, the three gains/parameters ( $K_P$ ,  $K_I$ , and  $K_D$ ) for the PID controller have to be found experimentally. A set of parameters that can be used for different types of gates in a gate library needs to ensure that the system reaches a steady state within a reasonable number of iterations (let's say <30), and that the steady-state error is less than the desired limit (for example, <10%).

In order to find the three PID parameters for our design system, we used the Zeigler–Nichols closed-loop [49] method because it takes the dynamics of the full system into account. The technique involves these steps: (1) Set  $K_I$  and  $K_D$  to zero; (2) set the desired value of output, and adjust  $K_P$  so that oscillation amplitude is constant; and (3) finally adjust  $K_I$  and  $K_D$  to remove oscillations while keeping the iteration count within limits. Using these criteria and with extensive experimentation, we arrived at these parameter values that are valid (and fixed, and do not need to be adjusted for different simulations) for automatic sizing of all multi-fan-in gates, i.e. NORs, NANDs, etc:  $K_P=0.1$ ,  $K_I=0.025$ , and  $K_D=0.05$ . (Note that 'automatic' refers to the sizing of gates and not to the PID parameters). To demonstrate, how the PID-parameters affect the design process, we show in Figure 11, the outcomes of using three different sets of PID parameters for sizing a NAND5 gate. We notice that a smaller value of  $K_I$  leads to a constant error, while a large value of  $K_D$  results in oscillations.

We used our PID-controller-based system to design gates with different fan-ins, for example, NAND with 3, 4, or 5 inputs. The system approached the desired  $\varepsilon_{avg}$  (<10%) in less than 10 design iterations in all cases. Figure 12 shows the schematics and transistor sizes for NAND3–NAND5 gates of the normal and optimal gates, using our system; the gates produce *balanced* rise- and fall-times.

The optimal- $L$  gates result in lower values of  $K_{prog}$ , and hence the smaller gate areas, as shown in the four plots of Figure 13. In Figure 13(a), we see that the optimal gates have low  $\varepsilon_{avg}$  with similarly sized (uniform  $K_{prog}$ ) normal gates. The SNMs are also higher for the optimal gates as shown Figure 13(b).

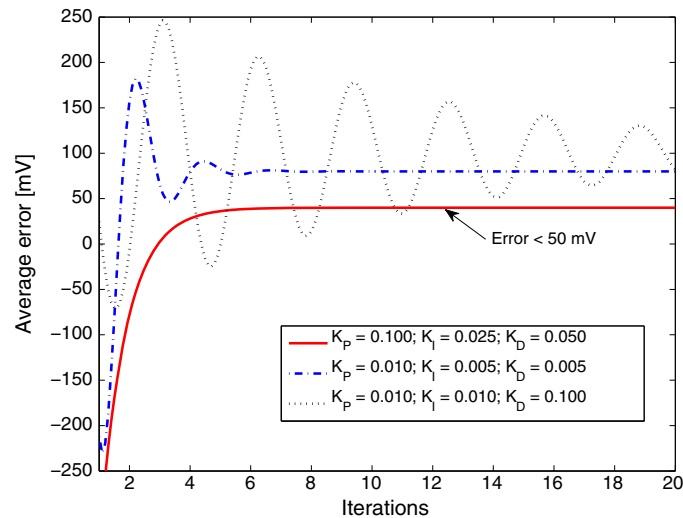


Figure 11. The PID parameters and the average output error of a NAND5 gate.

So far, we have looked at normal- $L$  and optimal- $L$  gates with progressive  $W$  sizing. In Table III, we have included an additional set of NAND2–NAND5 gates that are sized with uniform  $W$  (as in Figure 9 (a)). So the three sizing configurations in the table are: (1) *Type-I* that has uniformly-sized  $W$  and normal- $L$ 's; (2) *Type-II* that has progressively-sized  $W$  and optimal- $L$ 's; and (3) *Type-III* that has progressively-sized  $W$  and optimal- $L$ 's. Using the table data, we make the following observations about the NAND2–NAND5 of Types I–III:

Type-III has consistently higher SNMs than Type-I and Type-II. Just progressively sizing of  $W$ 's (but retaining normal- $L$ ) does not help with the SNM; rather we see that Type-II ends up having ~5% lower SNM than Type-I. Type-III gates have 33–51% higher SNMs for different fan-ins; the higher the fan-in, the bigger the difference is between the SNM of Type-III and Type-I gates.

Progressive sizing is expected to reduce the delay of the gates, but Type-II shows 11–27% performance penalty for different fan-ins, due to larger capacitances. Type-III carries the heaviest delay cost of 56–112% versus Type-I. Higher fan-in gates tend to have lesser delay cost as

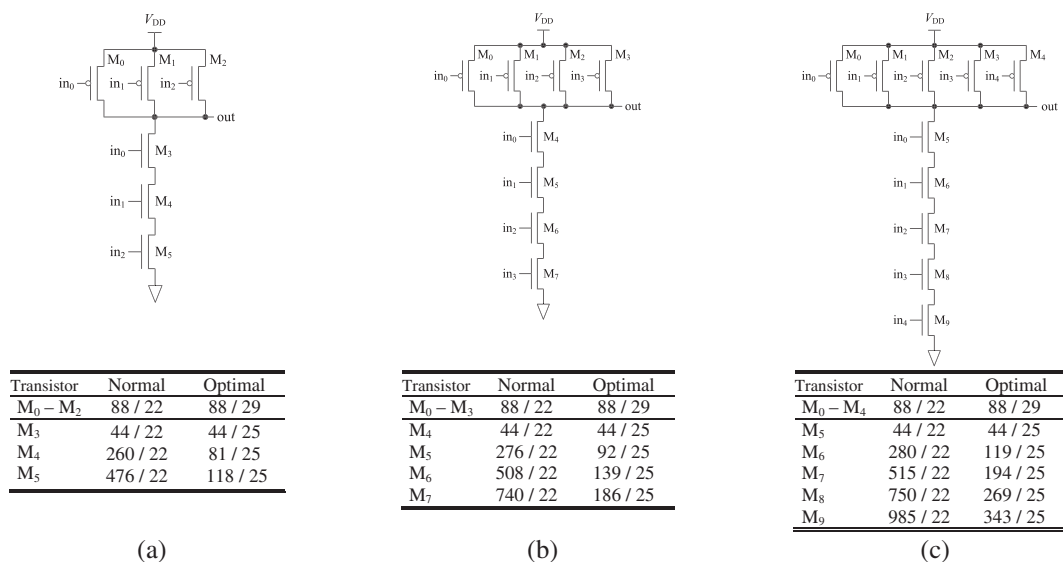


Figure 12. The schematics and the transistor dimensions of normal- $L$  and optimal- $L$  gates: (a) NAND3, (b) NAND4, and (c) NAND5.

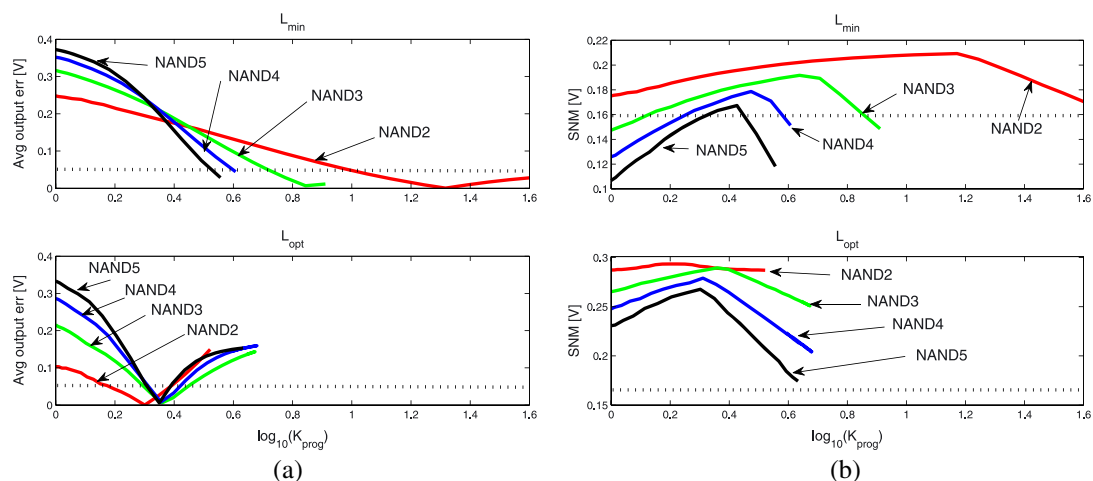


Figure 13. The comparison of normal- $L$  and optimal- $L$  NAND gates: (a) average output error vs. progressive-sizing factor; (b) SNM vs. progressive-sizing factor.

Table III. The comparison of NAND gates with different sizing schemes [F = 1E + 9].

Gate	Sizing scheme	Type	pMOS W/L [nm]	nMOS ( $W_{top}-W_{bottom}$ )/L [nm]	Gate area $\sum_i W_i \times L_i$ [nm <sup>2</sup> ]	SNM [V]	Delay [s]	Power [W]	PDP [W·s]
NAND2	Uniform-W-and-normal-L	I	88/22	(97-97)/22	8140	0.221	9.01E-12	1.01E-07	8.90E-19
	Progressive-W-and-normal-L	II	88/22	(44-327)/22	12034	0.209	1.14E-11	3.43E-08	3.91E-19
	Progressive-W-and-optimal-L	III	88/29	(44-85)/25	8329	0.294	1.91E-11	2.32E-08	4.43E-19
NAND3	Uniform-W-and-normal-L	I	88/22	(168-168)/22	16896	0.205	1.22E-11	8.88E-08	1.04E-18
	Progressive-W-and-normal-L	II	88/22	(44-476)/22	22968	0.193	1.38E-11	3.16E-08	4.37E-19
	Progressive-W-and-optimal-L	III	88/29	(44-118)/25	13731	0.290	2.33E-11	1.65E-08	3.84E-19
NAND4	Uniform-W-and-normal-L	I	88/22	(304-304)/22	34496	0.187	1.60E-11	9.17E-08	1.38E-18
	Progressive-W-and-normal-L	II	88/22	(44-740)/22	42240	0.179	1.84E-11	3.02E-08	5.55E-19
	Progressive-W-and-optimal-L	III	88/29	(44-186)/25	21733	0.278	2.88E-11	1.11E-08	3.21E-19
NAND5	Uniform-W-and-normal-L	I	88/22	(502-502)/22	64900	0.177	2.32E-11	9.84E-08	2.05E-18
	Progressive-W-and-normal-L	II	88/22	(44-985)/22	66308	0.168	2.59E-11	2.08E-08	5.36E-19
	Progressive-W-and-optimal-L	III	88/29	(44-343)/25	36985	0.268	3.62E-11	1.17E-08	4.24E-19

Note: Power, delay, and PDP differ from [14] due to (1) difference in  $L_{opt}$ 's, (2) increase in clock frequency, and (3) automated (vs. manual) sizing.

compared to low fan-in ones, because progressive sizing helps out more with *taller* nMOS stacks. Although not included here, raising the  $V_{DD}$  ( $>0.8\text{ V}$ ) and the use of low- $V_{TH}$  devices are two possible techniques for delay reduction.

Power consumption of Type-III gates is the lowest for all given fan-ins. Type-II gates reduce the power dissipation which is up to 33% of Type-I gates, whereas Type-III consume even less power, i.e. 23% of Type-I and less. Generally, higher fan-in gate Type-III gates exhibit better power advantages than lower ones.

Type-II gates result in significant PDP savings over their Type-I counterparts. Despite higher delay, Type-III gates have even lower PDP than Type-I gates, i.e. between 50 and 79%.

The power and PDP savings of Type-II gates carry an added area cost of up to 48% versus Type-I. Type-III NAND2's area is comparable to Type-I, whereas higher fan-in gates have up to 43% less area than Type-I's.

## 6. CONCLUSIONS

This paper has presented an automatic method for building gates that enhance SNMs, and are power and energy-efficient. The application of this technique has first been demonstrated for the basic INV, NAND2, and NOR2 gates. And then the scalability of the technique has been exhibited for higher fan-in gates (i.e. NAND3, NAND4, and NAND5). According to the standard design practice, the optimal gates are to be incorporated into the cell libraries, which in turn would be utilized for building larger designs or systems.

In the future, we plan to integrate other power reduction techniques (such as multiple- $V_{TH}$  and substrate biasing) into our sizing system. Application of this sizing technique for the FinFET gates is also under consideration; in this case, sizing steps would be quite *discrete* in nature due to the fixed widths of the fins.

We also plan to make the sizing system available on the Internet. This would entail online implementation of the overall PID-control system including generation and simulation of different netlists (for simple gates and for other standard cells), and data parsing.

## APPENDIX A: STATIC NOISE MARGIN—A BRIEF EXPLANATION

SNM (sometimes just referred to as *noise margin/NM*) of a gate is related to its DC/voltage transfer characteristics [3]. The SNM is a metric for the acceptable noise voltage on a gate's input so that the logical value of the output stays intact. Two components that make up the SNM are:

low noise margin  $NM_L = V_{IL} - V_{OL}$ , and

high-noise margin  $NM_H = V_{OH} - V_{IH}$

where

$V_{IL}$  = maximum *low* input voltage

$V_{OL}$  = maximum *low* output voltage

$V_{OH}$  = minimum *high* output voltage, and

$V_{IH}$  = minimum *high* input voltage

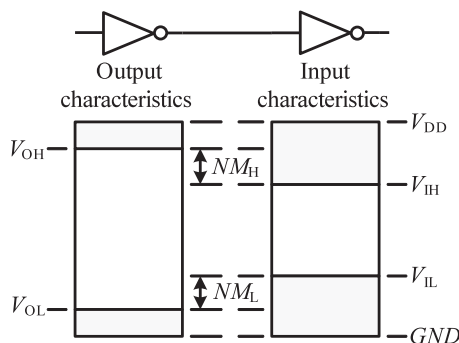


Figure A1. The high- and the low-noise margins of an INV.



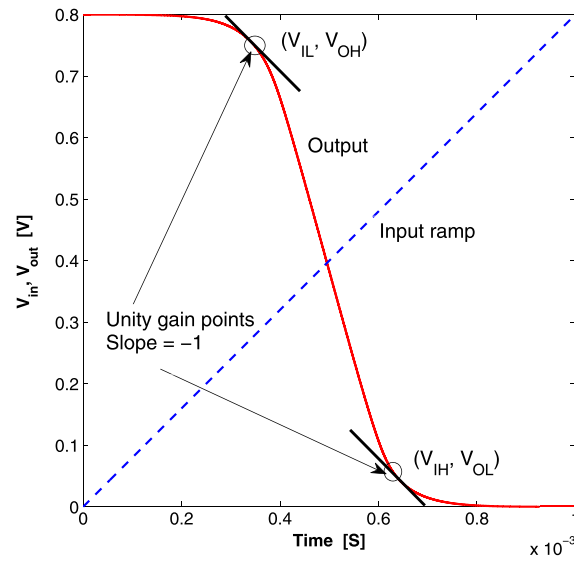


Figure A2. The voltage (DC) transfer characteristic of an INV.

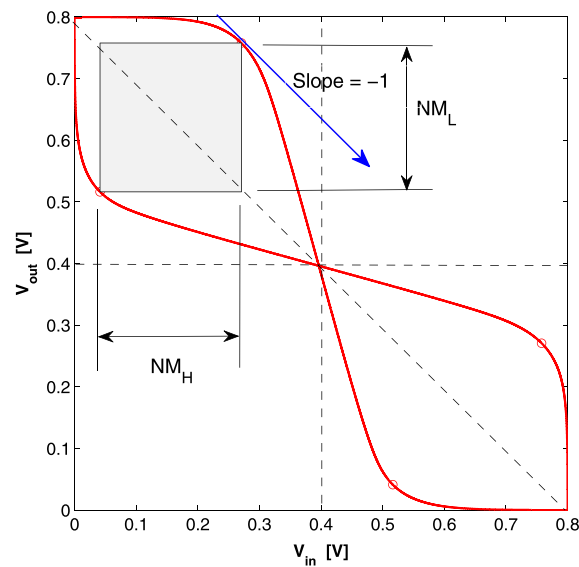
Figure A3. The *butterfly curve* for finding the NM/SNM of an INV.

Figure A1 is a pictorial representation of the definitions of the two noise margins.

The NMs can be derived from the voltage transfer characteristic curves. For this purpose, a ramp is input to an INV and the output is measured (as shown in Figure A2). The *butterfly* curves comprise an overlapping of  $V_{out}$ -vs- $V_{in}$  and  $V_{in}$ -vs- $V_{out}$  curves (see Figure A3).

In order to measure the NMs in an automated fashion a Matlab script has been developed. The script accurately finds the unity gain points by differentiating the  $V_{out}$  and  $V_{in}$  curves:  $dv_{out} = \text{diff}(v_{out}) ./ \text{diff}(v_{in})$ . The inflection points in two halves ('left' and 'right') of  $dv_{out}$  curve coincide with the *unity gain points* (slope = -1; refer to Figure A2) of  $V_{out}$ . Left inflection point is used to find  $V_{IL}$  and  $V_{OL}$ , and the right point corresponds to  $V_{OH}$  and  $V_{IH}$ . The four voltages determine  $NM_L$  and  $NM_H$  (as given in the formulae above). SNM is the smaller of the two NMs. Note that the loading (fanout) of the UUT has little effect on the SNM, in our experiments, as shown in Table A1.

Table AI. The effect of loading (fan-out) on SNMs of NAND gates (optimal-*L* and progressive sizing).

NAND2 as the UUT			NAND5 as the UUT		
Fan-out	SNM with INV-load [V]	SNM with NAND2-load [V]	Fan-out	SNM with INV-load [V]	SNM with NAND5-load [V]
0	0.294298	0.294298	0	0.268466	0.268466
1	0.294296	0.294295	1	0.268466	0.268443
2	0.294294	0.294292	2	0.268465	0.268419
3	0.294293	0.294289	3	0.268464	0.268395
4	0.294291	0.294286	4	0.268463	0.268371
5	0.294289	0.294283	5	0.268462	0.268347
6	0.294287	0.294280	6	0.268461	0.268323

## APPENDIX B: PID CONTROL SYSTEMS—A SHORT INTRODUCTION

Feedback control systems are found in both natural and engineered systems, from cars to aircrafts, and from air-conditioners to lighting systems. Most basic components of a feedback control include sensing, computation, and actuation. A feedback controller (Figure A4(a)) dynamically adjusts the behavior of one or components of systems in order to achieve the desired system output. PID (*proportional-integral-derivative*) controllers are useful even when the underlying process is undefined (see Figure A4 (b)). For the output  $y(t)$ , the PID algorithm is defined as:

$$y(t) = K_P e(t) + K_I \int e(t) dt + K_D \frac{de(t)}{dt}$$

where  $K_P$  is the *proportional gain*,  $K_I$  is the *integral gain*, and  $K_D$  is the *derivative gain*. The  $K_P$  helps attain stability but a large value of  $K_P$  results in a very fast response. The  $K_D$  limits the overshoot but large values of  $K_D$  can slow down the transient response and even cause instability. The  $K_I$  is important for shrinking steady-state errors. Understandably the three gains/parameters need to be properly *tuned* for the given system. Ziegler and Nichols [49] method for PID *tuning* (i.e. finding the three gains,  $K_P$ ,  $K_I$ , and  $K_D$ ) has been in practice for a long time. The method makes a priori assumptions about the system's model but does not necessitate that the model be fully known.

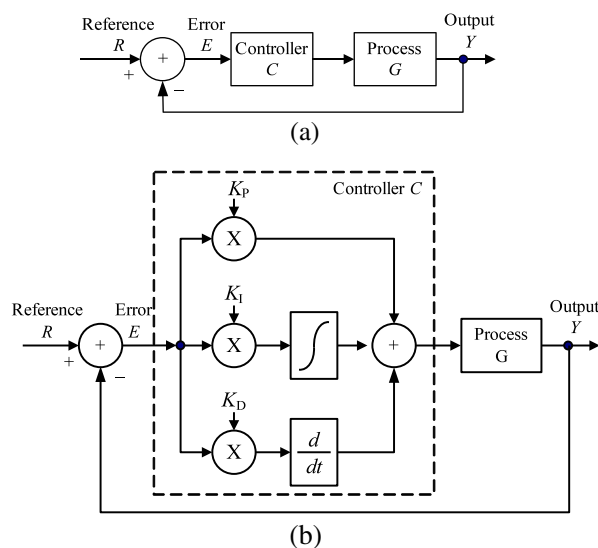


Figure A4. (a) A general feedback control system, and (b) a feedback system with a PID controller.

The applications of general feedback control systems in computing and networks include routing, data caching, and power management. Temperature and power management [50], and clock frequency scaling [51] are two of the practical examples in general-purpose microprocessors (single and multi-core) and systems-on-chip.

#### ACKNOWLEDGEMENTS

This research was partially funded by the SRC grant 2012-TJ-2332.

#### REFERENCES

- Moore GE. Cramming more components onto integrated circuits. *Electronics* 1965; **38**(8):114–117. [Reprinted in *Proc. IEEE*, vol. 86, no. 1, pp. 82–85, Jan. 1998].
- Intl. Tech. Roadmap for Semiconductors (ITRS). SEMATECH. Albany: NY, USA, 2011 [Online]. Available: <http://public.itrs.net/>.
- Rabaey JM, Chandrakasan A. Digital Integrated Circuits – A Design Perspective 2<sup>nd</sup> ed. Prentice Hall: NJ, USA, 2003.
- Roska T. Circuits, computers, and beyond Boolean logic. *Int J Circuit Theory Appl* 2007; **35**(5–6):485–496.
- Bansal A, Mukhopadhyay S, Roy K. Device-optimization technique for robust and low-power FinFET SRAM design in nanoscale era. *IEEE Trans Electr Dev* 2007; **54**(6):1409–1419.
- Tawfik SA, Kursun V. Low power and high speed multi threshold voltage interface circuits. *IEEE Trans VLSI Syst* 2009; **17**(5):638–645.
- Chang J-M, Pedram M. Energy minimization using multiple supply voltages. *IEEE Trans VLSI Syst* 1997; **5**(4):436–443.
- Calhoun BH, Chandrakasan A. Ultra-dynamic voltage scaling using sub-threshold operation and local voltage dithering in 90 nm CMOS. *Proc. ISSCC*, San Francisco, CA, USA, 300–301 and 599, 2005.
- Liu X, Mourad S. Performance of submicron CMOS devices and gates with substrate biasing. *Proc. ISCAS*, Geneva, Switzerland, 2000, **4**, 9–12.
- Corsonello P, Lanuzza M, Perri S. Gate-level body biasing technique for high-speed sub-threshold CMOS logic gates. *Int J Circuit Theory Appl* 2014; **42**(1):65–70.
- Albano D, Lanuzza M, Taco R, Crupi F. Gate-level body biasing for subthreshold logic circuits: analytical modeling and design guidelines. *Int J Circuit Theory Appl* 2014. DOI: 10.1002/cta.2016
- Faraji R, Naji HR, Rahimi-Nezhad M, Arabnejhad M. New SRAM design using body bias technique for low-power and high-speed applications. *Int J Circuit Theory Appl* 2013. DOI: 10.1002/cta.1914
- Marković D, Wang CC, Alarcón LP, Liu T-T, Rabaey JM. Ultralow-power design in near-threshold region. *Proc IEEE* 2010; **98**(2):237–252.
- Soeleman H, Roy K. Ultra-low power digital subthreshold logic circuits. *Proc. ISLPED*, San Diego, CA, USA, 1999; 94–96.
- Giustolisi G, Palumbo G, Criscione M, Cutri F. A low-voltage low-power voltage reference based on subthreshold MOSFETs. *IEEE J Solid-State Circ* 2003; **38**(1):151–154.
- Hill CF. Noise margin and noise immunity in logic circuits. *Microelectr* 1968; **1**(4):16–21.
- Flak J, Laiho M. Fault-tolerant programmable logic array for nanoelectronics. *Int J Circuit Theory Appl* 2012; **40**(12):1233–1247.
- Frustaci F, Lanuzza M, Perri S, Corsonello P. Analyzing noise robustness of wide fan-in dynamic logic gates under process variations. *Int J Circuit Theory Appl* 2014; **42**(5):452–467.
- Lohstroh J. Static and dynamic noise margins of logic circuits. *IEEE J Solid-State Circ* 1979; **14**(3):591–598.
- Beg A, Beg A. Reliability of Nano-Scaled Logic Gates Based on Binary Decision Diagrams. *2014 Proc. MSV'14*, Las Vegas, NV, USA, 1–5, 2014.
- Merino JL, Bota SA, Picos R, Segura J. Alternate characterization technique for static random-access memory static noise margin determination. *Int J Circuit Theory Appl* 2013; **41**:1085–1096.
- Beiu V, Beg A, Ibrahim W, Kharbash F, Alioto M. Enabling sizing for enhancing the static noise margins. *Proc. ISQED*, Santa Clara, CA, USA, 2013; 278–285.
- Beiu V, Ibrahim W, Beg A, Kharbash F. Towards ultra-low power/voltage using unconventionally sized arrays of transistors. *Proc. IEEE-NANO*, Birmingham: UK, 2012; 1–5.
- Beiu V, Ibrahim W, Beg A, Tache M. On sizing transistors for threshold voltage variations. *Proc. DFR*. Paris, France, 2012.
- Ibrahim W, Beg A, Beiu V. Highly reliable and low-power full adder cell. *Proc. IEEE-NANO*, Portland, OR, USA, 2011; 500–503.
- Beiu V, Beg A, Ibrahim W. Atto joule gates for the whole voltage range. *Proc. IEEE-NANO*, Portland, OR, USA, 2011; 1424–1429.
- Gupta P, Kahng AB, Sharma P, Sylvester D. Selective gate-length biasing for cost-effective runtime leakage control. *Proc. DAC*, San Diego, CA, USA, 2004; 327–330.

28. Venugopal R, Chakravarthi S, Chidambaram PR. Design of CMOS transistors to maximize circuit FOM using a coupled process and mixed-mode simulation methodology. *IEEE Electr Dev Lett* 2006; **27**(10):863–865.
29. Hanson S, Seok M, Sylvester D, Blaauw D. Nanometer device scaling in subthreshold logic and SRAM. *IEEE Trans Electr Dev* 2008; **55**(1):175–185.
30. Bol D, Ambroise R, Flandre D, Legat J-D. Interests and limitations of technology scaling for subthreshold logic. *IEEE Trans VLSI Syst* 2009; **17**(10):1508–1519.
31. Tajalli, Leblebici Y. Design trade-offs in ultra-low-power digital nanoscale CMOS. *IEEE Trans Circ & Syst* 2011; **58**(9):2189–2200.
32. Gupta H, Ghosh B. Transistor size optimization in digital circuits using ant colony optimization for continuous domain. *Int J Circuit Theory Appl* 2014; **42**(6):642–658.
33. Alioto M. Understanding DC behavior of subthreshold CMOS logic through closed-form analysis. *IEEE Trans Circ & Syst* 2010; **57**(7):1597–1607.
34. Calhoun BH, Wang A, Verma N, Chandrakasan A. Sub-threshold design: The challenges of minimizing circuit energy. *Proc. ISLPED*, Tegernsee, Germany, 2006; 366–368.
35. Blesken M, Lütkeemeier S, Rückert U. Multiobjective optimization for transistor sizing of sub-threshold CMOS logic standard cells. *Proc. ISCAS*, Paris, France, 2010; 1480–1483.
36. Alioto M, Palumbo G, Poli M. Simple and accurate modeling of the output transition time in nanometer CMOS gates. *Int J Circuit Theory Appl* 2010; **38**(10):995–1012.
37. Beg A, Elchouemi A. Enhancing static noise margin while reducing power consumption. *Proc. MWSCAS* 2013, Columbus: OH, USA, 2013; 348–351.
38. Beg A. Automatic Generation of Characterization Circuits - An Application In Academia 2013. *Proc. FIE 2013*, Oklahoma City, OK, USA, 2013; 661–664.
39. Beg A. A Web-Based Method For Building And Simulating Standard Cell Circuits - A Classroom Application. *Computer Applications in Engineering Education* 2014; 1–8.
40. Beg A. Designing Array-Based CMOS Logic Gates By Using A Feedback Control System. 2014 *Proc. SMC 2014*, San Diego, CA, USA, 2014 (in press).
41. SpiceGen (Tools) [Online]. Available: <http://www.azambeg.com>.
42. Predictive Technology Model [Online]. Available: <http://ptm.asu.edu/>.
43. Zhao W, Cao Y. New generation of predictive technology model for sub-45 nm early design exploration. *IEEE Trans Electr Dev* 2006; **53**(11):2816–2823.
44. Berkeley Short-channel IGFET Model, 2011 [Online]. Available: [http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4\\_LR](http://www-device.eecs.berkeley.edu/bsim/?page=BSIM4_LR)
45. Ngspice – Mixed mode – Mixed level circuit simulator [Online]. Available: <http://ngspice.sourceforge.net/>.
46. Courtland R. The status of Moore's Law: It's complicated. *IEEE Spectrum*. [Online]. Available: <http://spectrum.ieee.org/semiconductors/devices/the-status-of-moores-law-its-complicated>.
47. Asenov A, Brown AR, Davies JH, Kaya S, Slavcheva G. Simulation of intrinsic parameter fluctuations in decanometer and nanometer-scale MOSFETs. *IEEE Trans Electr Dev* 2003; **50**(9):1837–1852.
48. Zavyalova L, Lucas K, Zhang Q, Fan Y, Sethi S, Song H, Tyminski J. Analysis of OPC optical model accuracy with detailed scanner information. *Proc. SPIE Optical Microlithography*, San Jose, CA, USA, 2008; 1–12.
49. Leigh JR. *Applied Control Theory*. IEE Press 1988.
50. Meijer M, de Gyvez J, Otten R. On-chip digital power supply control for system-on-chip applications. *Proc. ISLPED*, San Diego, CA, USA, 2005; 311–314.
51. Almeida G, *et al.* Predictive dynamic frequency scaling for multi-processor systems-on-chip. *Proc. ISCAS*, Rio De Janeiro, Brazil, 2011; 1500–1503.