Aging-Aware Standard Cell Library Design

Saman Kiamehr Farshad Firouzi Mojtaba Ebrahimi Mehdi B. Tahoori
Karlsruhe Institute of Technology, Karlsruhe, Germany
e-mails: {kiamehr, farshad.firouzi, mojtaba.ebrahimi, mehdi.tahoori}@kit.edu

***Abstract*—Transistor aging, mostly due to Bias Temperature Instability (BTI), is one of the major unreliability sources at nano-scale technology nodes. BTI causes the circuit delay to increase and eventually leads to a decrease in the circuit lifetime. Typically, standard cells in the library are optimized according to the design time delay, however, due to the asymmetric effect of BTI, the rise and fall delays might become significantly imbalanced over the lifetime. In this paper, the BTI effect is mitigated by balancing the rise and fall delays of the standard cells at the excepted lifetime. We find an optimal tradeoff between the increase in the size of the library and the lifetime improvement (timing margin reduction) by non-uniform extension of the library cells for various ranges of the input signal probabilities. The simulation results reveal that our technique can prolong the circuit lifetime by around 150% with a negligible area overhead.**

I. Introduction

Among various reliability issues in nano-scale technology nodes, transistor aging is a serious concern which leads to a gradual increase of the threshold voltage and the circuit delay over time. It is shown that *Bias Temperature Instability* (BTI) is the dominant aging factor in digital circuits designed with the recent technology nodes [2]. The BTI-induced threshold voltage shift is a function of different factors such as temperature, supply voltage and the duty cycle (the ratio between the stress to total time). The duty cycle of each transistor in a gate, is a function of the signal probabilities (i.e. the occurrence probability of a node to be logic "1") of the gates inputs.

In the standard cell library design, the transistors in each gate are sized in a way that the rise and fall delays become equal for a typical load capacitance and transition time [6]. However, BTI asymmetrically affects the rise and fall delays of a gate according to its input signal probabilities [3]. As a result, the rise and fall delays of the gate become significantly imbalanced during the operational time.

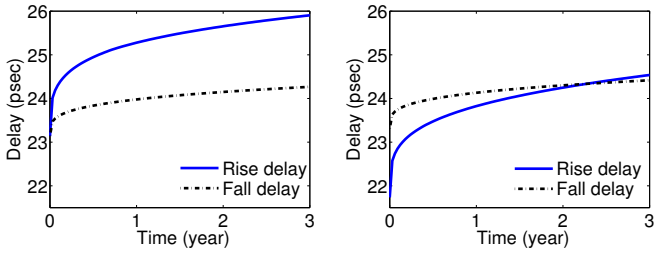
To address this issue, we propose a technique to optimize the standard cells considering BTI effect. The main idea is to optimize the library cells in order to balance their rise and fall delays at the expected lifetime rather than the design time. In our technique, we consider uneven BTI-induced degradation by taking the duty cycle of each transistor into account. This is achieved by replicating and redesigning the library cells for different ranges of input signal probabilities. In order to keep a reasonable tradeoff between the library size increase and the lifetime improvement (i.e., BTI-induced timing margin reduction), we investigate the optimal sampling ranges of input signal probabilities. The simulation results on ISCAS89 benchmark circuits show that our technique improves the lifetime of the circuits by about 150% in average while it has a negligible effect on the area.

II. Aging-aware cell sizing

In the typical library cell design, the optimal ratio of Wp (the width of the PMOS transistor) to Wn (the width of the NMOS transistor) is adjusted to balance the rise and fall delays of the gate [6]. However, different transistors inside the circuit which have different *signal probabilities* (SPs), have different BTI-induced threshold voltage shifts. Therefore, due to the BTI effect, the threshold voltage of transistors degrades unevenly leading to unequal rise and fall delays of the gates at the end of the expected lifetime. Figure 1(a) shows the rise and fall delays of a simple inverter with an input $SP_{in} = 0.1$ over the time. As shown in this figure, although the rise and fall delays are equal at $time = 0$, they become different after 3 years. Since the *duty cycle* (DC) of the PMOS transistor ($DC_{NBTI} = 1 - SP_{in} = 0.9$) is higher than that of the NMOS transistor ($DC_{PBTI} = SP_{in} = 0.1$), the rise delay degradation is higher than the fall delay degradation.

Our objective is to design the cell (by changing the Wp/Wn ratio), in a way that its rise and fall delays become equal at the end of the expected lifetime (see Figure 1(b)). As a result, the Wp/Wn ratio has to increase compared to the typical mode in order to have the same rise and fall delays at the expected lifetime. As shown in Figure 1, by optimizing Wp/Wn ratio to balance the rise and fall delays at the expected lifetime, at the expense of upsizing only the PMOS transistors in the gate, a better post-aging delay is achieved. Since the BTI effect is a function of the DC (see Figure 2(a)), the optimized BTI-aware Wp/Wn ratio for each cell is a function of the input DC (and hence SP).

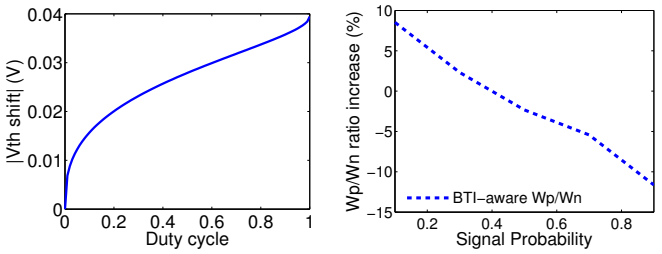
Figure 2(b) shows the optimized Wp/Wn ratio for different input SPs normalized to the case that the BTI effect is not considered. As shown in this figure, for smaller SPs the Wp/Wn ratio is higher. This is due to the fact that for the smaller SPs, the NBTI effect is larger and, as a result, the PMOS transistor in the pull-up network has to be designed larger to compensate the NBTI effect. Another observation is that if $SP = 0.5$, the Wp/Wn ratio is almost equal to the case that the aging effect is neglected. In fact, when $SP = 0.5$, both NMOS and PMOS transistors are almost under the same stress and as a result both transistors degrade at almost the same pace. Therefore, their ratio in this case is almost equal to the typical case. Moreover, for the larger SPs, the Wp/Wn ratio is less than the typical case. Since we consider a constant Wn and for the larger SPs ($SP > 0.5$) the NMOS transistor degrades



(a) Time-zero-balanced (b) Lifetime-balanced

Fig. 1. The effect of Wp/Wn optimization on BTI-induced delay degradation of an inverter with input SP=0.1

978-3-9815370-2-4/DATE14/©2014 EDAA



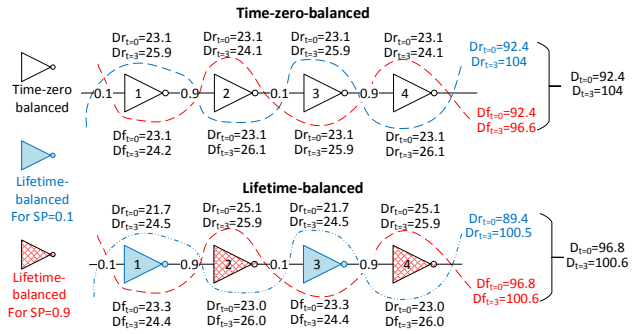
(a) (b)

Fig. 2. a) The effect of different duty cycle on the BTI-induced Vth shift, b) The optimized Wp/Wn ratio increase for different SPs normalized to the case that it is optimized for time=0

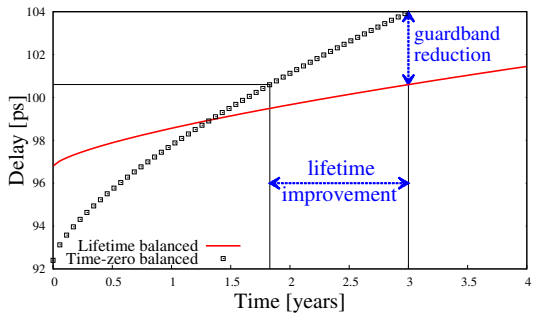
more than the PMOS transistor, the fall delay becomes larger than the rise delay and hence we can decrease Wp to make the rise and the fall delays equal in order to save area and power (with no impact on the delay).

The efficiency of this approach is demonstrated in the following by the example circuit given in Figure 3. The circuit is an inverter chain with the primary input SP of 0.1. Figure 3(a) shows the rise and the fall delays when the time-zero-balanced (typical) library cells are used. As shown in this figure, although the rise and the fall delays of the path are the same at the design time, they diverge significantly throughout the lifetime. Therefore, the total rise and the fall delays of the path become significantly imbalanced. Figure 3(a) also shows the rise and fall delays when the lifetime-balanced library cells are used. As shown for this case, the rise and fall delays of the path become similar at the end of the operational lifetime of the circuit and the overall circuit delay for this case ($100.6ps$) is less than the case in which the time-zero-balanced library cells are used ($104ps$).

This is obtained by the upsizing of PMOS transistors for the inverters under NBTI stress (higher Wp/Wn ratio for the



(a)



(b)

Fig. 3. A simple circuit to show the efficiency of aging-aware standard cell sizing: a) time-zero-balanced vs lifetime-balanced mapping b) delay of lifetime-balanced vs time-zero balanced

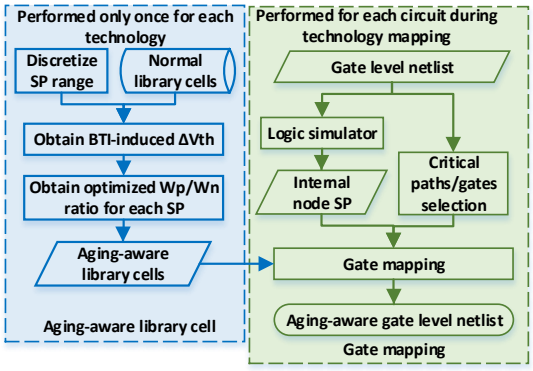


Fig. 4. The overall flow of proposed methodology

inverters 1 and 3 with smaller SPs according to Figure 2(b)). This upsizing is compensated (in terms of power and area) by downsizing the PMOS transistors of the other inverters (inverters 2 and 4) which are more under PBTI stress (smaller Wp/Wn ratio for larger SPs according to Figure 2(b)).

It should be noted that, as shown in Figure 3(b), the delay of the circuit at time zero may even become larger, however, it becomes smaller at the expected lifetime. Therefore, with the reduction in the amount of aging-induced timing margin, the clock period is reduced in overall.

III. Cell Library Redesign and Mapping

Figure 4 shows the overall flow of the proposed methodology. It consists of two phases: i) aging-aware standard cell library redesign, and ii) circuit library mapping using the new library cells. The details of each phase are explained next.

A. Aging-aware cell library

We propose aging-aware standard cell library redesign, in which the library cells are optimized for different SPs considering the BTI effect. According to Figure 2(b), the optimized Wp/Wn ratio is a function of the SPs of cell inputs. However, it is not possible to extend the library for all combinations of SPs. For this purpose, the SP range ([0.0, 1.0]) is discretized and for each combination of these SP values a new library cell is added and optimized by finding a suitable Wp/Wn ratio for that range using SPICE simulations.

In order to obtain the optimized Wp/Wn ratio, first the BTI-induced ΔVth for all the internal transistors of the cell is calculated according to the particular SP value. Then, the Wp/Wn ratio of the cell is swept using a binary search to obtain the best ratio leading to equal rise and fall delays. For example, if we discretize the SP range to the $\{0, 0.25, 0.5, 0.75, 1)\}$, then for a simple inverter (INV_X1) we need to extend the library with five cells: $\{INV\_X1\_0$, $INV\_X1\_0.25$, $INV\_X1\_0.5$, $INV\_X1\_0.75$, $INV\_X1\_1\}$ and for each cell the optimized Wp/Wn ratio is obtained to have equal rise and fall delays at the expected lifetime. In order to build the library, each cell is characterized to obtain the delay and leakage *look-up tables* (LUTs). By increasing the library size, the characterization time/effort increases accordingly, however, it should be noted that the aging-aware library cell design and characterization are done only once for each technology.

Library size increase and non-uniform SP sampling: If the SP range ([0.0, 1.0]) is discretized to m ranges, for each cell with n inputs, m^n cells are added to the library. In other words, the number of new library cells increases exponentially using the proposed aging-aware library cell. More sampling

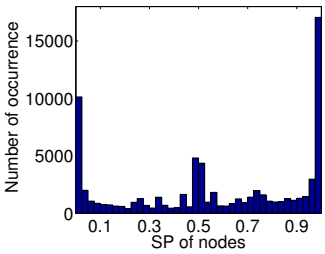


Fig. 5. The histogram of internal node SP distribution for ISCAS89 benchmark circuits (over all benchmarks)

points may increase the efficiency of this approach in terms of delay balancing, however, it leads to a very large size of the library. This makes the approach infeasible for industrial-scale libraries which contain 1000-1500 cells. This implies that a suitable discretization resolution has to be considered for a reasonable trade-off between the efficiency of the method and the library size. For this purpose, two important parameters have to be considered:

i) The sensitivity of the BTI-induced Vth shift to SPs: As shown in Figure 2(a), the BTI-induced Vth shift has different sensitivities to the SP (DC) in different ranges of SPs. Therefore, more samples (at least one sample) has to be considered for more sensitive ranges (e.g. [0.0, 0.1] range in Figure 2(a)).

ii) Distribution of the SPs of internal nodes in typical circuits: The SPs of the internal nodes in typical circuits are not uniformly distributed over the entire range. Figure 5 shows the histogram of the SP distribution for the internal nodes of ISCAS89 benchmark circuits. As shown in this figure, SP values around 0.1, 0.5, and 0.9 are more frequent. Therefore, in a non-uniform sampling, more samples have to be chosen in such ranges.

Considering these two factors, a non-uniform discretization and sampling can be used in order to keep the sampling points as few as possible, while maintaining a high efficiency of this technique in terms of aging mitigation.

B. *Technology mapping with aging-aware cell library*

Once the aging-aware cell library is constructed, it can be used for the technology mapping phase for different circuits. For this purpose, we start from a netlist mapped into the original aging-unaware library. Then, the gate level netlist is given to a logic simulator to obtain the internal node SPs. According to the input SPs of each gate, a new cell with the closest set of input SPs from the new library is chosen to replace the initial cell. For example, if we have a two-input NAND gate with the SPs of 0.15 and 0.70 for its inputs, according to the discretization example of previous subsection, this NAND gate will be replaced with the $NAND\_0.25\_0.75$ aging-aware cell. In order to minimize the area/power overhead, this remapping is done only for the critical gates (gates which are in the critical/near-critical paths) since the others have no contribution to the delay of the circuit.

IV. Simulation Results

In this section, we show the efficiency of our proposed method by comparing it to the time-zero-balanced library cell design as well as scenarios in which a representative SP is considered for all gates. We also investigate the trade-off between the library size and the aging mitigation by considering various sample sizes and strategies. The impact of different workloads is also discussed. Figure 6 shows the details of our flow to obtain the simulation results. The gate-level netlist, is obtained by synthesizing the ISCAS89 benchmark circuits using Nangate 45 nm library [1] containing 42 cells (INVERTER, BUFFER and two inputs AND, OR, NAND, NOR, , XOR, and XNOR gates). The worst case BTI-induced delay degradation is assumed to be 10% in 3 years and the parameters of the BTI model are set accordingly.

The first step is to conduct the aging-aware library cell design as proposed in Section III-A. Here, we consider four different scenarios for the discretization of the SP range:

1) Uniform sampling with 5 points (U5): The SP range is discretized uniformly to 5 points: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The number of logical cells is increased by ≈50X (from 42 to 2010 cells). Obviously this library size increase is unacceptable.

2) Non-uniform sampling with 3 points (NU3): Here we only consider 3 samples for SP ($\{0.1, 0.5, 0.9\}$) to decrease the library size. In this case, the library size consists of 522 cells which is almost 4 times smaller than that of U5.

3) Non-uniform sampling with 2 points (NU2): Here we consider only 2 samples for SP ($\{0.1, 0.9\}$) to further decrease the library size, which now consists of 192 cells ($>$ 10x reduction compared to U5).

4) Non-uniform worst case sampling with 2 points (NU2W): Here also 2 samples for SP ($\{0.1, 0.5\}$) are considered. The library size is equal to the previous case, however, in this case all the gates with an input SP larger than 0.5 are mapped to a cell with SP=0.5, to upsize the PMOS transistors (according to Figure 2(b)) to further reduce BTI-degradation compared to NU2.

To obtain the results for the case in which the effect of SP is neglected in the library cell design (e.g. the method in [4]), we considered two cases. In the first case, we assumed that all the cells are optimized with the input SP of 0.5. For the second case, we considered a worst-case approach in which, according to Figure 2(b), a very small SP of 0.1 is considered to upsize the PMOS transistors to reduce BTI-degradation. After obtaining the new aging-aware cells, each cell is characterized to obtain the delay LUTs for different load capacitances, transition times, and ΔVth.

In the mapping phase, only the critical gates in the gate-level netlist are replaced with the new cells, as described in Section III-B. Next, the netlist mapped into the original library and the one mapped into the aging-aware library are given to a logic simulator to obtain the SPs of internal nodes. Then the BTI-induced ΔVth of all the transistors are obtained according to the model proposed in [3]. Finally, the gate level netlists, ΔVth values, and the delay LUTs are given to an aging-aware Static Timing Analyzer (STA), similar to the one proposed in [5], to obtain the fresh and aged circuit delays.

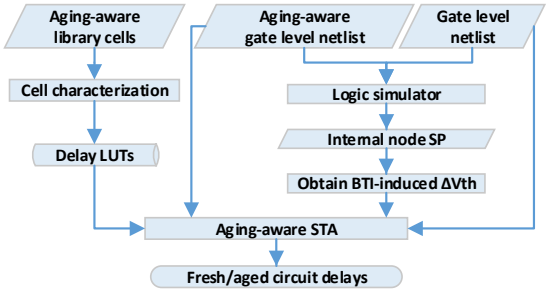


Fig. 6. The overall flow to obtain simulation results

Benchmark Circuit	# of gates	Timing margin reduction				Lifetime improvement						Area overhead					
		U5	NU3	NU2	NU2W	U5	NU3	NU2	NU2W	SP=0.5	SP=0.1	U5	NU3	NU2	NU2W	SP=0.5	SP=0.1
s953	683	37.3%	39.1%	31.7%	20.3%	255.0%	274.3%	200.6%	109.7%	-11.7%	78.2%	0.0%	0.0%	0.0%	0.1%	0.0%	0.3%
s1196	797	23.4%	22.1%	12.0%	19.6%	132.0%	121.9%	57.5%	104.8%	-13.3%	61.2%	0.0%	0.0%	0.0%	0.4%	-0.1%	1.0%
s1423	824	31.8%	33.9%	25.5%	22.4%	201.8%	221.5%	147.9%	124.8%	-15.9%	93.8%	0.0%	-0.1%	-0.1%	0.3%	-0.1%	0.7%
s1238	881	52.3%	36.2%	25.8%	26.5%	438.5%	244.3%	150.1%	155.9%	-13.8%	127.4%	0.1%	0.0%	0.0%	0.1%	0.0%	0.3%
s1488	902	24.0%	21.0%	17.3%	14.8%	136.1%	114.3%	89.4%	73.4%	-2.8%	31.8%	0.0%	0.0%	0.0%	0.2%	0.0%	0.3%
s9234	1725	32.4%	29.5%	22.4%	20.5%	207.5%	181.2%	124.2%	110.8%	-11.3%	71.2%	0.0%	-0.1%	0.0%	0.3%	-0.1%	0.6%
s5378	2926	36.2%	38.3%	28.7%	28.7%	244.1%	265.5%	174.2%	174.4%	-13.4%	135.6%	0.0%	0.0%	0.0%	0.1%	0.0%	0.2%
s13207	4074	9.8%	29.4%	19.9%	19.9%	45.4%	180.7%	106.9%	106.7%	-10.6%	56.9%	0.3%	0.1%	0.0%	0.2%	0.0%	0.5%
s15850	5244	21.9%	23.7%	8.1%	13.6%	120.7%	133.8%	36.5%	66.8%	-8.4%	36.6%	0.0%	0.0%	0.0%	0.3%	-0.1%	0.6%
s38584	18142	26.2%	19.8%	14.4%	18.1%	153.8%	106.0%	71.5%	94.5%	-12.5%	52.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
average		26.7%	25.4%	14.2%	18.0%	167.2%	155.3%	85.5%	97.1%	-11.8%	56.8%	0.1%	0.0%	0.0%	0.3%	-0.1%	0.6%

TABLE I. THE EFFICIENCY OF OUR TECHNIQUE COMPARED TO THE NORMAL STANDARD CELL LIBRARY DESIGN IN TERMS OF LIFETIME IMPROVEMENT AND AREA OVERHEAD. **U5**: UNIFORM SAMPLING WITH 5 POINTS ($\{0.1,0.3,0.5,0.7,0.9\}$), **NU3**: NON-UNIFORM SAMPLING WITH 3 POINTS($\{0.1,0.5,0.9\}$), **NU2**: NON-UNIFORM SAMPLING WITH 3 POINTS($\{0.1,0.9\}$), **NU2W**: NON-UNIFORM WORST CASE SAMPLING WITH 2 POINTS($\{0.1,0.5\}$),.

The simulation results are shown in Table I. The area is approximated by the summation of all the transistors widths inside the circuit. Therefore, it is not only a representative for the area, but also shows the trend of the power consumption. As shown in this table, the uniform sampling with 5 points (U5) and the non-uniform sampling with 3 points (NU3) scenarios lead to 26.7% and 25.4% timing margin reduction (167% and 155% lifetime improvement), respectively. This implies that, while NU3 needs much smaller library size compared to U5, their efficiencies in terms of the lifetime improvement and the timing margin reduction are comparable. This shows that the lifetime improvement saturates when the number of SP samples exceeds a particular limit. Both scenarios have negligible area overhead. There are two reasons for that: i) the aging-aware technology mapping is only performed for the critical gates, and ii) for the gates with large input SPs, the PMOS transistors are down-sized to save area/power. The results for NU2 scenario shows that with a much smaller library size, 14% timing margin reduction (85% lifetime improvement) is obtained. However, NU2W gives better results compared to NU2 in terms of timing margin (lifetime improvement) with the same library size at the expense of a small area/power overhead (0.3%). This shows the importance of the proper SP sampling for the aging-aware cell library design.

Compared to the other alternative (in which the SP distribution of the internal nodes is neglected [4]), considering a fixed SP of 0.5 leads to even a worse lifetime compared to the original library cell design, although it might be beneficial only when the NBTI effect is considered. For the other scenario (fixed SP of 0.1), the lifetime improvement is much less than all the four scenarios above. Moreover, for this worst case scenario, the area/power overhead is higher (0.6%).

To account for wearout mechanisms, the clock frequency has to be set according to the delay of the circuit at the expected lifetime (not at zero-time), by adding aging-induced timing margins. This means that the circuit performance is determined by the post-aging delay. Although our proposed method may even lead to a higher circuit delay at time-zero (up to 2%), it provides an overall performance improvement by reducing the post-aging delay and its associated timing margin, as shown in Table I. For the case where the performance is fixed, the proposed technique results in an improvement in the lifetime of the circuit.

Effect of the workload. For the simulations results presented in Table I, we assumed that the primary input SPs are 0.5 for both the gate mapping phase and the SP calculation of internal nodes (with which BTI-induced ΔVth values are obtained accordingly). However, different workloads result in different primary input SPs, and internal SPs accordingly, observed by the circuit during its operational lifetime. To account for this on the efficiency of our methodology, we performed two sets of experiments. In the first experiment, the gates are mapped (optimized) according to the primary input SP of 0.5 but the internal nodes SPs (and BTI-induced Vth shifts) are calculated for the primary input SP of 0.2. In the second experiment, the primary input SPs used for the mapping phase and the internal SP (and degradation) calculation were chosen as 0.2 and 0.5, respectively, i.e. the reverse situation as in the first experiment. The results show that the efficiency of U5 decreases by around 25% (from 167% lifetime improvement to 127%) when different primary input SPs are used for the mapping phase and the delay degradation calculation. However, this has negligible effect on the efficiency of the methods with the fewer SP samples (less than 1% decrease in the lifetime improvement of NU2W). Therefore, while the optimization is done at "design time" using a particular assumption over the primary input SPs, the method still remains effective as the workload changes during runtime.

V. CONCLUSION

In nano-scale technology nodes, BTI has become a crucial aging effect leading to reduced circuits lifetime. In this paper, we proposed a BTI-aware library cell design to mitigate the BTI effect. The main idea is to balance the rise and fall delays of a cell by considering the target lifetime delay degradations instead of time-zero delays. We also presented a technology mapping technique in which the critical gates in the circuit are mapped to suitable cells within this aging-aware library based on their input signal probabilities. The simulation results show that our technique can improve the lifetime by approximately 150% with negligible area/power overheads. Our experiments also show that the proposed approach remains effective even with workload changes during runtime.

REFERENCES

[1] "Nangate," http://www.nangate.com/.

[2] K. Bernstein *et al.*, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development - Advanced silicon technology*, vol. 50, pp. 433–449, 2006.

[3] S. Bhardwaj *et al.*, "Predictive modeling of the NBTI effect for reliable design," in *CICC,*, 2006.

[4] M. B. da Silva *et al.*, "Nbti-aware technique for transistor sizing of high-performance cmos gates," in *LATW'09*, pp. 1–5.

[5] F. Firouzi *et al.*, "Incorporating the impacts of workload-dependent runtime variations into timing analysis," in *DATE*, 2013.

[6] J. M. Rabaey *et al.*, *Digital integrated circuits*. Prentice hall Englewood Cliffs, 2002, vol. 2.