

Aging-Aware Reliable Multiplier Design With Adaptive Hold Logic

Ing-Chao Lin, *Member, IEEE*, Yu-Hung Cho, and Yi-Ming Yang

Abstract—Digital multipliers are among the most critical arithmetic functional units. The overall performance of these systems depends on the throughput of the multiplier. Meanwhile, the negative bias temperature instability effect occurs when a pMOS transistor is under negative bias ($V_{gs} = -V_{dd}$), increasing the threshold voltage of the pMOS transistor, and reducing multiplier speed. A similar phenomenon, positive bias temperature instability, occurs when an nMOS transistor is under positive bias. Both effects degrade transistor speed, and in the long term, the system may fail due to timing violations. Therefore, it is important to design reliable high-performance multipliers. In this paper, we propose an aging-aware multiplier design with a novel adaptive hold logic (AHL) circuit. The multiplier is able to provide higher throughput through the variable latency and can adjust the AHL circuit to mitigate performance degradation that is due to the aging effect. Moreover, the proposed architecture can be applied to a column- or row-bypassing multiplier. The experimental results show that our proposed architecture with 16×16 and 32×32 column-bypassing multipliers can attain up to 62.88% and 76.28% performance improvement, respectively, compared with 16×16 and 32×32 fixed-latency column-bypassing multipliers. Furthermore, our proposed architecture with 16×16 and 32×32 row-bypassing multipliers can achieve up to 80.17% and 69.40% performance improvement as compared with 16×16 and 32×32 fixed-latency row-bypassing multipliers.

Index Terms—Adaptive hold logic (AHL), negative bias temperature instability (NBTI), positive bias temperature instability (PBTI), reliable multiplier, variable latency.

I. INTRODUCTION

DIGITAL multipliers are among the most critical arithmetic functional units in many applications, such as the Fourier transform, discrete cosine transforms, and digital filtering. The throughput of these applications depends on multipliers, and if the multipliers are too slow, the performance of entire circuits will be reduced.

Furthermore, negative bias temperature instability (NBTI) occurs when a pMOS transistor is under negative bias ($V_{gs} = -V_{dd}$). In this situation, the interaction between inversion layer holes and hydrogen-passivated Si atoms breaks

the Si–H bond generated during the oxidation process, generating H or H₂ molecules. When these molecules diffuse away, interface traps are left. The accumulated interface traps between silicon and the gate oxide interface result in increased threshold voltage (V_{th}), reducing the circuit switching speed. When the biased voltage is removed, the reverse reaction occurs, reducing the NBTI effect. However, the reverse reaction does not eliminate all the interface traps generated during the stress phase, and V_{th} is increased in the long term. Hence, it is important to design a reliable high-performance multiplier.

The corresponding effect on an nMOS transistor is positive bias temperature instability (PBTI), which occurs when an nMOS transistor is under positive bias. Compared with the NBTI effect, the PBTI effect is much smaller on oxide/polygate transistors, and therefore is usually ignored. However, for high- k /metal-gate nMOS transistors with significant charge trapping, the PBTI effect can no longer be ignored. In fact, it has been shown that the PBTI effect is more significant than the NBTI effect on 32-nm high- k /metal-gate processes [2]–[4].

A traditional method to mitigate the aging effect is overdesign [5], [6], including such things as guard-banding and gate oversizing; however, this approach can be very pessimistic and area and power inefficient. To avoid this problem, many NBTI-aware methodologies have been proposed. An NBTI-aware technology mapping technique was proposed in [7] to guarantee the performance of the circuit during its lifetime. In [8], an NBTI-aware sleep transistor was designed to reduce the aging effects on pMOS sleep-transistors, and the lifetime stability of the power-gated circuits under consideration was improved. Wu and Marculescu [9] proposed a joint logic restructuring and pin reordering method, which is based on detecting functional symmetries and transistor stacking effects. They also proposed an NBTI optimization method that considered path sensitization [12]. In [10] and [11], dynamic voltage scaling and body-biasing techniques were proposed to reduce power or extend circuit life. These techniques, however, require circuit modification or do not provide optimization of specific circuits.

Traditional circuits use critical path delay as the overall circuit clock cycle in order to perform correctly. However, the probability that the critical paths are activated is low. In most cases, the path delay is shorter than the critical path. For these noncritical paths, using the critical path delay as the overall cycle period will result in significant timing waste. Hence, the variable-latency design was proposed to reduce the timing waste of traditional circuits.

The variable-latency design divides the circuit into two parts: 1) shorter paths and 2) longer paths. Shorter paths

Manuscript received February 26, 2013; revised September 26, 2013 and December 14, 2013; accepted February 11, 2014. Date of publication April 29, 2014; date of current version February 20, 2015. This work was supported by the National Science Council of Taiwan under Grant NSC 100-2221-E-006-177 and Grant NSC 102-2221-E-006-281.

I.-C. Lin is with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan (e-mail: iclin@csie.ncku.edu.tw).

Y.-H. Cho is with Welltrend Semiconductor, Inc., Kaohsiung 80147, Taiwan (e-mail: steven751008@hotmail.com).

Y.-M. Yang is with VIA Technology, Inc., New Taipei City 23644, Taiwan (e-mail: leo771031@gmail.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVLSI.2014.2311300

can execute correctly in one cycle, whereas longer paths need two cycles to execute. When shorter paths are activated frequently, the average latency of variable-latency designs is better than that of traditional designs. For example, several variable-latency adders were proposed using the speculation technique with error detection and recovery [13]–[15]. A short path activation function algorithm was proposed in [16] to improve the accuracy of the hold logic and to optimize the performance of the variable-latency circuit. An instruction scheduling algorithm was proposed in [17] to schedule the operations on nonuniform latency functional units and improve the performance of Very Long Instruction Word processors. In [18], a variable-latency pipelined multiplier architecture with a Booth algorithm was proposed. In [19], process-variation tolerant architecture for arithmetic units was proposed, where the effect of process-variation is considered to increase the circuit yield. In addition, the critical paths are divided into two shorter paths that could be unequal and the clock cycle is set to the delay of the longer one. These research designs were able to reduce the timing waste of traditional circuits to improve performance, but they did not consider the aging effect and could not adjust themselves during the runtime. A variable-latency adder design that considers the aging effect was proposed in [20] and [21]. However, no variable-latency multiplier design that considers the aging effect and can adjust dynamically has been done.

A. Paper Contribution

In this paper, we propose an aging-aware reliable multiplier design with a novel adaptive hold logic (AHL) circuit. The multiplier is based on the variable-latency technique and can adjust the AHL circuit to achieve reliable operation under the influence of NBTI and PBTI effects. To be specific, the contributions of this paper are summarized as follows:

- 1) novel variable-latency multiplier architecture with an AHL circuit. The AHL circuit can decide whether the input patterns require one or two cycles and can adjust the judging criteria to ensure that there is minimum performance degradation after considerable aging occurs;
- 2) comprehensive analysis and comparison of the multiplier’s performance under different cycle periods to show the effectiveness of our proposed architecture;
- 3) an aging-aware reliable multiplier design method that is suitable for large multipliers. Although the experiment is performed in 16- and 32-bit multipliers, our proposed architecture can be easily extended to large designs;
- 4) the experimental results show that our proposed architecture with the 16×16 and 32×32 column-bypassing multipliers can attain up to 62.88% and 76.28% performance improvement compared with the 16×16 and 32×32 fixed-latency column-bypassing (FLCB) multipliers. In addition, our proposed architecture with 16×16 and 32×32 row-bypassing multipliers can achieve up to 80.17% and 69.40% performance improvement as compared with 16×16 and 32×32 fixed-latency row-bypassing multipliers.

The paper is organized as follows. Section II introduces the background of the column-bypassing multiplier,

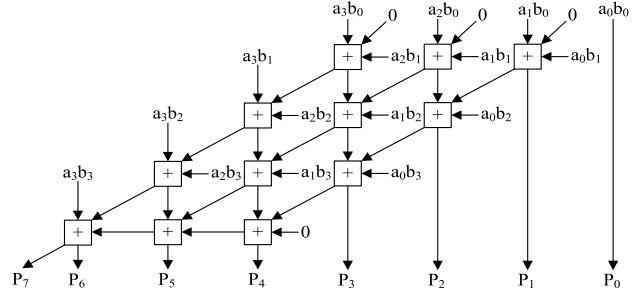


Fig. 1. 4×4 normal AM.

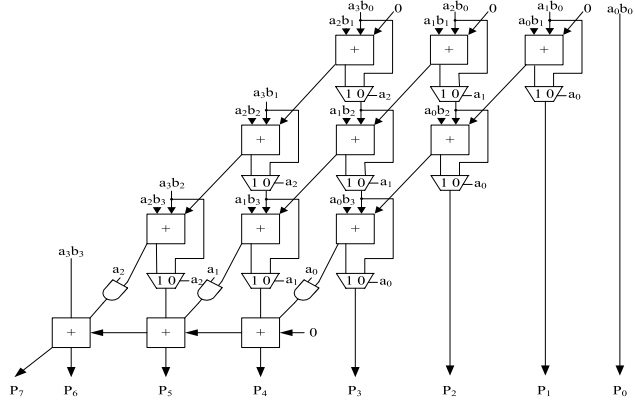


Fig. 2. 4×4 column-bypassing multiplier.

row-bypassing multiplier, variable-latency design, and NBTI/PBTI models. Section III details the aging-aware variable-latency multiplier based on the column- or row-bypassing multiplier. The experimental setup and results are presented in Section IV. Section V concludes this paper.

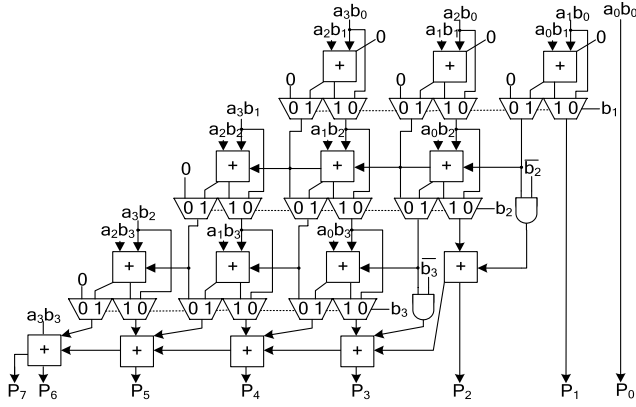
II. PRELIMINARIES

A. Column-Bypassing Multiplier

A column-bypassing multiplier is an improvement on the normal array multiplier (AM). The AM is a fast parallel AM and is shown in Fig. 1. The multiplier array consists of $(n - 1)$ rows of carry save adder (CSA), in which each row contains $(n - 1)$ full adder (FA) cells. Each FA in the CSA array has two outputs: 1) the sum bit goes down and 2) the carry bit goes to the lower left FA. The last row is a ripple adder for carry propagation.

The FAs in the AM are always active regardless of input states. In [22], a low-power column-bypassing multiplier design is proposed in which the FA operations are disabled if the corresponding bit in the multiplicand is 0. Fig. 2 shows a 4×4 column-bypassing multiplier. Supposing the inputs are $1010_2 * 1111_2$, it can be seen that for the FAs in the first and third diagonals, two of the three input bits are 0: the carry bit from its upper right FA and the partial product $a_i b_i$. Therefore, the output of the adders in both diagonals is 0, and the output sum bit is simply equal to the third bit, which is the sum output of its upper FA.

Hence, the FA is modified to add two tristate gates and one multiplexer. The multiplicand bit a_i can be used as the

Fig. 3. 4×4 row-bypassing multiplier.

selector of the multiplexer to decide the output of the FA, and a_i can also be used as the selector of the tristate gate to turn off the input path of the FA. If a_i is 0, the inputs of FA are disabled, and the sum bit of the current FA is equal to the sum bit from its upper FA, thus reducing the power consumption of the multiplier. If a_i is 1, the normal sum result is selected. More details for the column-bypassing multiplier can be found in [22].

B. Row-Bypassing Multiplier

A low-power row-bypassing multiplier [23] is also proposed to reduce the activity power of the AM. The operation of the low-power row-bypassing multiplier is similar to that of the low-power column-bypassing multiplier, but the selector of the multiplexers and the tristate gates use the multiplier.

Fig. 3 is a 4×4 row-bypassing multiplier. Each input is connected to an FA through a tristate gate. When the inputs are $1111_2 * 1001_2$, the two inputs in the first and second rows are 0 for FAs. Because b_1 is 0, the multiplexers in the first row select $a_i b_0$ as the sum bit and select 0 as the carry bit. The inputs are bypassed to FAs in the second rows, and the tristate gates turn off the input paths to the FAs. Therefore, no switching activities occur in the first-row FAs; in return, power consumption is reduced. Similarly, because b_2 is 0, no switching activities will occur in the second-row FAs. However, the FAs must be active in the third row because the b_3 is not zero. More details for the row-bypassing multiplier can also be found in [23].

C. Variable-Latency Design

Section I mentioned that the variable-latency design was proposed to reduce the timing waste occurring in traditional circuits that use the critical path cycle as an execution cycle period. The basic concept is to execute a shorter path using a shorter cycle and longer path using two cycles. Since most paths execute in a cycle period that is much smaller than the critical path delay, the variable-latency design has smaller average latency.

For example, Fig. 4 is an 8-bit variable-latency ripple carry adder (RCA). A_8-A_1 , B_8-B_1 are 8-bit inputs, and S_8-S_1 are

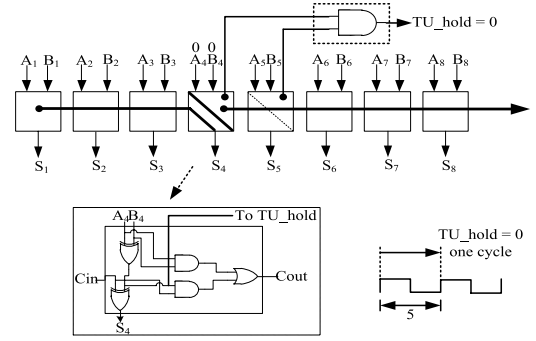


Fig. 4. 8-bit RCA with a hold logic circuit.

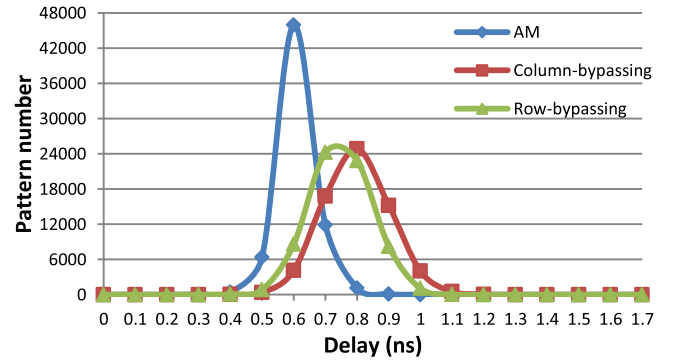


Fig. 5. Path delay distribution of AM, column-, and row-bypassing multipliers for 65536 input patterns.

the outputs. Supposing the delay for each FA is one, and the maximum delay for the adder is 8.

Through simulation, it can be determined that the possibility of the carry propagation delay being longer than 5 is low. Hence, the cycle period is set to 5, and hold logic is added to notify the system whether the adder can complete the operation within a cycle period.

Fig. 4 also shows the hold logic that is used in this circuit. The function of the hold logic is $(A_4 \text{ XOR } B_4)(A_5 \text{ XOR } B_5)$. If the output of the hold logic is 0, i.e., $A_4 = B_4$ or $A_5 = B_5$, either the fourth or the fifth adder will not produce a carryout. Hence, the maximum delay will be less than one cycle period. When the hold logic output is 1, this means that the input can activate paths longer than 5, so the hold logic notifies the system that the current operation requires two cycles to complete. Two cycles are sufficient for the longest path to complete ($5 * 2$ is larger than 8).

The performance improvement of the variable-latency design can be calculated as follows: if the possibility of each input being 1 is 0.5, the possibility of $(A_4 \text{ XOR } B_4)(A_5 \text{ XOR } B_5)$ being 1 is 0.25. The average latency for the variable-latency design is $0.75 * 5 + 0.25 * 10 = 6.25$. Compared with the simple fixed-latency RCA, which has an average latency of 8, the variable-latency design can achieve a 28% performance improvement.

Fig. 5 shows the path delay distribution of a 16×16 AM and for both a traditional column-bypassing and traditional row-bypassing multiplier with 65536 randomly chosen input patterns. All multipliers execute operations on a fixed cycle

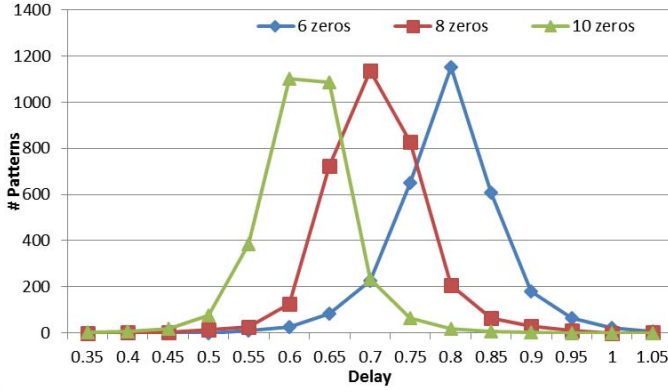


Fig. 6. Delay distribution of a 16 column-bypassing multiplier under three different numbers of zeros in the multiplicands.

period. The maximum path delay is 1.32 ns for the AM, 1.88 ns for the column-bypassing multiplier, and 1.82 ns for the row-bypassing multiplier. It can be seen that for the AM, more than 98% of the paths have a delay of <0.7 ns. Moreover, more than 93% and 98% of the paths in the FLCB and row-bypassing multipliers present a delay of <0.9 ns, respectively. Hence, using the maximum path delay for all paths will cause significant timing waste for shorter paths, and redesigning the multiplier with variable latency can improve their performance.

Another key observation is that the path delay for an operation is strongly tied to the number of zeros in the multiplicands in the column-bypassing multiplier. Fig. 6 shows the delay distribution of the 16×16 column-bypassing multiplier under three different numbers of zeros in the multiplicands: 1) 6; 2) 8; and 3) 10. Three thousand randomly selected patterns are used in each experiment. It can be seen as the number of zeros in the multiplicands increases, delay distribution is left shifted, and average delay is reduced. The reason for this is the multiplicand is used as the select line for column-bypassing multipliers, and if more zeros exist in the multiplicand, more FAs will be skipped, and the sum bit from the upper FA is passed to the lower FA, reducing the path delay. Note that similar experiments are also done for row-bypassing multipliers. However, because the results are similar, they are not shown to avoid duplications.

For a row-bypassing multiplier, the multipliers are used to determine whether a pattern needs one cycle or two cycles to complete an operation because the multiplier is used as the select line.

This makes the column-bypassing multiplicand and row-bypassing multiplier excellent candidates for the variable-latency design since we can simply examine the number of zeros in the multiplicand or multiplier to predict whether the operation requires one cycle or two cycles to complete.

D. Aging Model

As mentioned in Section I, the NBTI (PBTI) effect occurs when a pMOS (nMOS) transistor is under negative (positive) bias voltage, resulting in V_{th} drift. When the bias voltage is removed, the recovery process occurs, reducing the V_{th} drift.

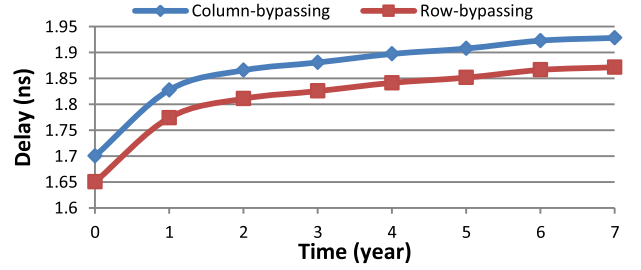


Fig. 7. Trend for circuit aging for a 16×16 column-bypassing multiplier.

If a pMOS (nMOS) transistor is under constant stress, this is referred to as static NBTI (PBTI). If both stress and recovery phases exist, it is referred to as dynamic NBTI (PBTI). The V_{th} drift of pMOS (nMOS) transistor due to the static NBTI (PBTI) effect can be described by dc reaction-diffusion (RD) framework. If transistors are under alternative stress and recovery phases, the dc RD model should be modified to an ac RD model [24], [25]

$$\Delta V_{th}(t) \cong K_{AC} \times t^n \cong \alpha(S, f) \times K_{DC} \times t^n \quad (1)$$

where α is a function of stress frequency (f) and signal probability (S). Since the impact of frequency is relatively insignificant, the effect of signal frequency is ignored. K_{DC} is a technology-dependent constant

$$K_{DC} = A \times T_{OX} \times \sqrt{C_{OX}(V_{GS} - V_{th})} \times [1 - V_{DS}/\alpha(V_{GS} - V_{th})] \times \exp(E_{OX}/E_0) \times \exp\left(-\frac{E_a}{kT}\right) \quad (2)$$

where A is a constant, and T_{OX} is the oxide thickness. E_{OX} is the gate electric field, which is $(V_{GS} - V_{th})/T_{OX}$; k is the Boltzmann constant, and T is the temperature. E_0 and E_a are technology-independent characteristics of the reaction that are equal to 1.9–2.0 MV/cm and 0.12 eV, respectively. More details about this model can be found in [26].

In this paper, we use 32-nm high- k metal gate models. We set the temperature at 125 °C in our simulation and use the above BTI model to predict the BTI effect on the circuits. Fig. 7 shows the simulated delays of the 16×16 column- and row-bypassing multipliers under a seven-year NBTI/PBTI effect. From this figure, it can be seen that the BTI effect increased the critical path circuit delay by $\sim 13\%$. Hence, if the BTI effect is not considered during circuit design, the increased delay may cause system failure in the long term.

III. PROPOSED AGING-AWARE MULTIPLIER

This section details the proposed aging-aware reliable multiplier design. It introduces the overall architecture and the functions of each component and also describes how to design AHL that adjusts the circuit when significant aging occurs.

A. Proposed Architecture

Fig. 8 shows our proposed aging-aware multiplier architecture, which includes two m -bit inputs (m is a positive number), one $2m$ -bit output, one column- or row-bypassing multiplier, $2m$ 1-bit Razor flip-flops [27], and an AHL circuit.

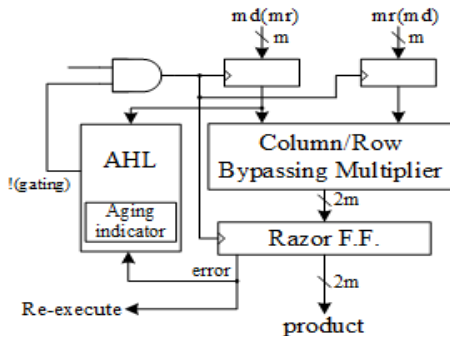


Fig. 8. Proposed architecture (md means multiplicand; mr means multiplier).

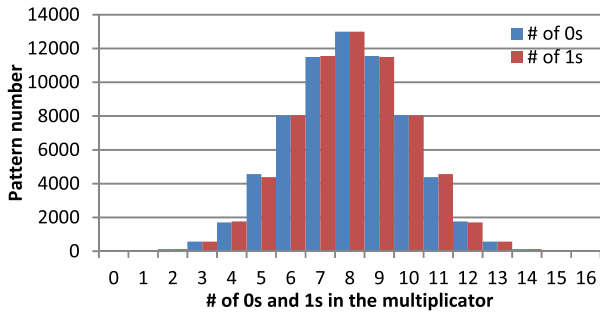


Fig. 9. Pattern number distribution based on the number of 0 s and 1 s in the multiplier.

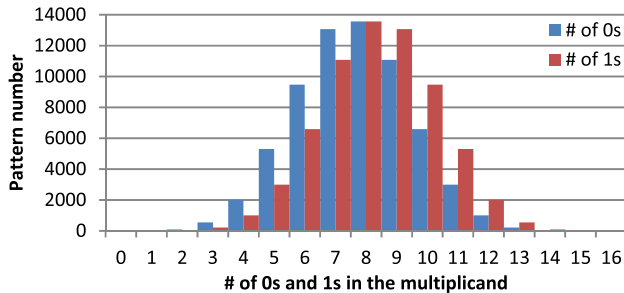


Fig. 10. Pattern number distribution based on the number of 0 s and 1 s in the multiplicand.

The inputs of the row-bypassing multiplier are the symbols in the parentheses.

In the proposed architecture, the column- and row-bypassing multipliers can be examined by the number of zeros in either the multiplicand or multiplier to predict whether the operation requires one cycle or two cycles to complete. When input patterns are random, the number of zeros and ones in the multiplier and multiplicand follows a normal distribution, as shown in Figs. 9 and 10. Therefore, using the number of zeros or ones as the judging criteria results in similar outcomes.

Hence, the two aging-aware multipliers can be implemented using similar architecture, and the difference between the two bypassing multipliers lies in the input signals of the AHL. According to the bypassing selection in the column- or row-bypassing multiplier, the input signal of the AHL in the architecture with the column-bypassing multiplier is the multiplicand, whereas that of the row-bypassing multiplier is the multiplier. Razor flip-flops can be used to detect

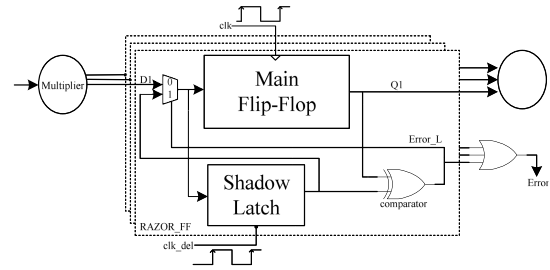


Fig. 11. Razor flip flops.

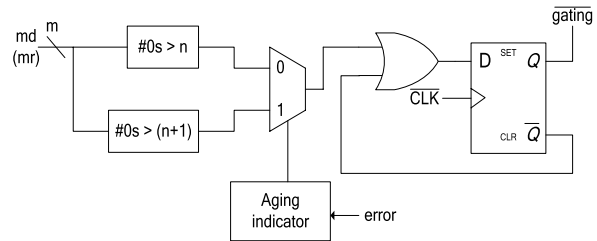


Fig. 12. Diagram of AHL (md means multiplicand; mr means multiplier).

whether timing violations occur before the next input pattern arrives.

Fig. 11 shows the details of Razor flip-flops. A 1-bit Razor flip-flop contains a main flip-flop, shadow latch, XOR gate, and mux. The main flip-flop catches the execution result for the combination circuit using a normal clock signal, and the shadow latch catches the execution result using a delayed clock signal, which is slower than the normal clock signal. If the latched bit of the shadow latch is different from that of the main flip-flop, this means the path delay of the current operation exceeds the cycle period, and the main flip-flop catches an incorrect result. If errors occur, the Razor flip-flop will set the error signal to 1 to notify the system to reexecute the operation and notify the AHL circuit that an error has occurred. We use Razor flip-flops to detect whether an operation that is considered to be a one-cycle pattern can really finish in a cycle. If not, the operation is reexecuted with two cycles. Although the reexecution may seem costly, the overall cost is low because the reexecution frequency is low. More details for the Razor flip-flop can be found in [27].

The AHL circuit is the key component in the aging-aware variable-latency multiplier. Fig. 12 shows the details of the AHL circuit. The AHL circuit contains an aging indicator, two judging blocks, one mux, and one D flip-flop. The aging indicator indicates whether the circuit has suffered significant performance degradation due to the aging effect. The aging indicator is implemented in a simple counter that counts the number of errors over a certain amount of operations and is reset to zero at the end of those operations. If the cycle period is too short, the column- or row-bypassing multiplier is not able to complete these operations successfully, causing timing violations. These timing violations will be caught by the Razor flip-flops, which generate error signals. If errors happen frequently and exceed a predefined threshold, it means the circuit has suffered significant timing degradation due to the aging effect, and the aging indicator will output signal 1;

otherwise, it will output 0 to indicate the aging effect is still not significant, and no actions are needed.

The first judging block in the AHL circuit will output 1 if the number of zeros in the multiplicand (multiplier for the row-bypassing multiplier) is larger than n (n is a positive number, which will be discussed in Section IV), and the second judging block in the AHL circuit will output 1 if the number of zeros in the multiplicand (multiplier) is larger than $n + 1$. They are both employed to decide whether an input pattern requires one or two cycles, but only one of them will be chosen at a time. In the beginning, the aging effect is not significant, and the aging indicator produces 0, so the first judging block is used. After a period of time when the aging effect becomes significant, the second judging block is chosen. Compared with the first judging block, the second judging block allows a smaller number of patterns to become one-cycle patterns because it requires more zeros in the multiplicand (multiplier).

The details of the operation of the AHL circuit are as follows: when an input pattern arrives, both judging blocks will decide whether the pattern requires one cycle or two cycles to complete and pass both results to the multiplexer. The multiplexer selects one of either result based on the output of the aging indicator. Then an OR operation is performed between the result of the multiplexer, and the \bar{Q} signal is used to determine the input of the D flip-flop. When the pattern requires one cycle, the output of the multiplexer is 1. The $!(\text{gating})$ signal will become 1, and the input flip flops will latch new data in the next cycle. On the other hand, when the output of the multiplexer is 0, which means the input pattern requires two cycles to complete, the OR gate will output 0 to the D flip-flop. Therefore, the $!(\text{gating})$ signal will be 0 to disable the clock signal of the input flip-flops in the next cycle. Note that only a cycle of the input flip-flop will be disabled because the D flip-flop will latch 1 in the next cycle.

The overall flow of our proposed architecture is as follows: when input patterns arrive, the column- or row-bypassing multiplier, and the AHL circuit execute simultaneously. According to the number of zeros in the multiplicand (multiplier), the AHL circuit decides if the input patterns require one or two cycles. If the input pattern requires two cycles to complete, the AHL will output 0 to disable the clock signal of the flip-flops. Otherwise, the AHL will output 1 for normal operations. When the column- or row-bypassing multiplier finishes the operation, the result will be passed to the Razor flip-flops. The Razor flip-flops check whether there is the path delay timing violation. If timing violations occur, it means the cycle period is not long enough for the current operation to complete and that the execution result of the multiplier is incorrect. Thus, the Razor flip-flops will output an error to inform the system that the current operation needs to be reexecuted using two cycles to ensure the operation is correct. In this situation, the extra reexecution cycles caused by timing violation incurs a penalty to overall average latency. However, our proposed AHL circuit can accurately predict whether the input patterns require one or two cycles in most cases. Only a few input patterns may cause a timing variation when the AHL

circuit judges incorrectly. In this case, the extra reexecution cycles did not produce significant timing degradation.

In summary, our proposed multiplier design has three key features. First, it is a variable-latency design that minimizes the timing waste of the noncritical paths. Second, it can provide reliable operations even after the aging effect occurs. The Razor flip-flops detect the timing violations and reexecute the operations using two cycles. Finally, our architecture can adjust the percentage of one-cycle patterns to minimize performance degradation due to the aging effect. When the circuit is aged, and many errors occur, the AHL circuit uses the second judging block to decide if an input is one cycle or two cycles.

IV. EXPERIMENTAL RESULT

Our experiments are conducted in a Linux operating system. We adopt a 32-nm high- k predictive technology model [1] to estimate the BTI degradation for seven years. The proposed multiplier is designed in Verilog and converted to SPICE files using SpringSoft Laker. Then Synposys Nanosim is used to analyze the delay and power of the circuit. The V_{th} drift caused by BTI is estimated using the BTI model proposed in Section II-D and is added into the SPICE files during simulation.

In the variable-latency design, the average latency is affected by both the percentage of one-cycle patterns and the cycle period. If more patterns only require one cycle, the average latency is reduced. Similarly, if the cycle period is reduced, the average latency is also reduced. However, the cycle period cannot be too small. If the cycle period is too small, large amounts of timing violations will be detected by the Razor flip-flops, and the average latency will increase. Hence, it is important to analyze the tradeoff between the percentage of one-cycle patterns and the cycle period. To achieve this, we analyze three scenarios for both 16×16 and 32×32 variable-latency column-bypassing (VLCB) and variable-latency row-bypassing (VLRB) multipliers. We also compare the results with the AM, a FLCB multiplier, and a fixed-latency row-bypassing (FLRB) multiplier.

A. Average Latency Comparison

We first compare the average latency of the AM, FLCB, FLRB, VLCB, and VLRB multipliers in both 16×16 and 32×32 multipliers. In the case of the 16×16 variable-latency bypassing multiplier, there are three scenarios: 1) Skip-7; 2) Skip-8; and 3) Skip-9. For VLCB multipliers, Skip-7 means the patterns that have seven or more zeros in the multiplicand are regarded as one-cycle patterns, and Skip-8 means the patterns that have eight or more zeros in the multiplicand are regarded as one-cycle patterns, and so on. Similarly, for VLRB multipliers, Skip-7 means the patterns that have seven or more zeros in the multiplier are regarded as one-cycle patterns, and Skip-8 means the patterns that have eight or more zeros in the multiplier are regarded as one-cycle patterns, and so on.

Table I lists the one-cycle pattern ratio of a one-cycle pattern ratio in the 16×16 variable-latency bypassing multiplier, and

TABLE I
ONE-CYCLE PATTERN RATIO IN 16×16 MULTIPLIER

	16x16 VLCB	16x16 VLRB
Skip-7	73.58%	77.39%
Skip-8	53.78%	59.89%
Skip-9	33.22%	40.20%

TABLE II
ONE-CYCLE PATTERN RATIO IN 32×32 MULTIPLIER

	32x32 VLCB	32x32 VLRB
Skip-15	66.46%	66.99%
Skip-16	52.68%	52.74%
Skip-17	38.18%	38.42%

Table II lists the one-cycle pattern ratio in the 32×32 variable-latency bypassing multiplier. It can be seen that the one-cycle pattern ratio decreases as the skip number increases.

Fig. 13 compares the average latency between the AM, FLCB, FLRB, adaptive variable-latency column bypassing (A-VLCB), and adaptive variable-latency row bypassing (A-VLRB) multipliers without aging in the 16×16 multiplier. To simplify the comparison, the results are divided into three parts: 1) Skip-7; 2) Skip-8; and 3) Skip-9. The latency of the AM, FLRB, and FLCB is shown to be 1.32, 1.82, and 1.88, respectively.

In Fig. 13(a), when the skip number is 7, the latency of the A-VLCB is 37.3% less than the FLCB and 10.7% less than the AM when the cycle period is 0.9. The latency of the A-VLRB is 39.9% less than the FLRB and 17.2% less than the AM when the cycle period is 0.85.

In Fig. 13(b), when the skip number is 8, the latency of the A-VLCB is 32.2% less than the FLCB and 3.4% less than the AM when the cycle period is 0.8. The latency of the A-VLRB is 35.5% less than the FLRB and 11.1% less than the AM when the cycle period is 0.8.

In Fig. 13(c), when the skip number is 9, the latency of the A-VLCB is 28.8% less than the FLCB, but 1.3% higher than the AM when the cycle period is 0.8. The latency of the A-VLRB is 32.0% less than the FLRB and 6.3% less than the AM when the cycle period is 0.7.

It can be seen the average latency of the variable-latency bypassing multiplier can be lower than the fixed-latency bypassing multiplier and the AM when appropriate cycle periods and skip numbers are used. In addition, the average latency of the A-VLRB is smaller than that of A-VLCB, and the average latency of the FLRB is smaller than that of the FLCB in the 16×16 multipliers.

Note that both the A-VLCB and A-VLRB have a preferred cycle period range that can achieve lower average latency than the AM, and the preferred cycle period range changes when the skip number of the variable-latency multiplier changes. For example, the preferred cycle period range for the A-VLCB when the skip number is 7 is between 0.85 and 1 ns, and when the skip number is 8, the preferred cycle period range is between 0.8 and 0.9 ns.

When the system clock period is smaller than the preferred range, the average latency is increased because more timing violations occur, and more operations are reexecuted in smaller cycle periods. Similarly, when the system clock period is larger than the preferred range, the average latency is increased because more timing waste occurs. To achieve an average latency lower than that of fixed-latency multipliers, it is important to match the system cycle period with the multiplier's preferred cycle period. If both do not match, methods, such as transistor sizing or using another skip number, can be used to adjust the multiplier's cycle period.

Fig. 14 compares the average latency of AM, FLCB, FLRB, A-VLCB, and A-VLRB multipliers without aging in a 32×32 multiplier. To simplify the comparison, the results are divided into three parts: 1) Skip-15; 2) Skip-16; and 3) Skip-17. The latency of 32×32 AM, FLRB, and FLCB is shown to be 2.74, 3.95, and 3.88 ns, respectively.

In Fig. 14(a), when the skip number is 15, the latency of the A-VLCB is 46.6% less than the FLCB and 24.2% less than the AM when the cycle period is 1.5. The latency of the A-VLRB is 42.5% less than the FLRB and 17.1% less than the AM when the cycle period is 1.65.

In Fig. 14(b), when the skip number is 16, the latency of the A-VLCB is 43.1% less than the FLCB and 19.5% less than the AM when the cycle period is 1.45. The latency of the A-VLRB is 38.3% less than the FLRB and 11.1% less than the AM when the cycle period is 1.6.

In Fig. 14(c), when the skip number is 17, the latency of the A-VLCB is 40% less than the FLCB and 14.8% less than the AM when the cycle period is 1.4. The latency of the A-VLRB is 35.0% less than the FLRB and 6.3% less than the AM when the cycle period is 1.55.

It can be seen that more latency reduction can be achieved in a larger multiplier. For example, in Fig. 13(c), the latency of the A-VLCB is slightly larger than that of the AM when the cycle period is 0.75, but in Fig. 14(c), the latency of the A-VLCB is better than that of the AM. This is because in larger multipliers, the delay difference between long and short paths is increased, and thus, a variable-latency design can reduce more timing waste. In addition, the preferred cycle period that has lower latency than the AM is increased. This is because variable-latency in a larger multiplier has more latency reduction. Therefore, the preferred region is also increased.

Note that in Fig. 13, the latency of the VLCB is larger than that of the VLRB. However, in the 32×32 multiplier, the latency of the VLCB is less than that of the VLRB, on average. This is because the critical path of the VLRB has more multiplexers, and in a larger multiplier, the critical path of the VLRB will increase faster than that of the VLCB.

B. Latency Comparison Among Different Skip Numbers

We also compare the average latency of a variable-latency bypassing multiplier under three different skip numbers. Fig. 15(a) and (b) compares the average latency in the 16×16 VLCB and VLRB multipliers under three different skip numbers without the aging effect. Fig. 16(a) and (b) displays the error count for three different skip numbers over different

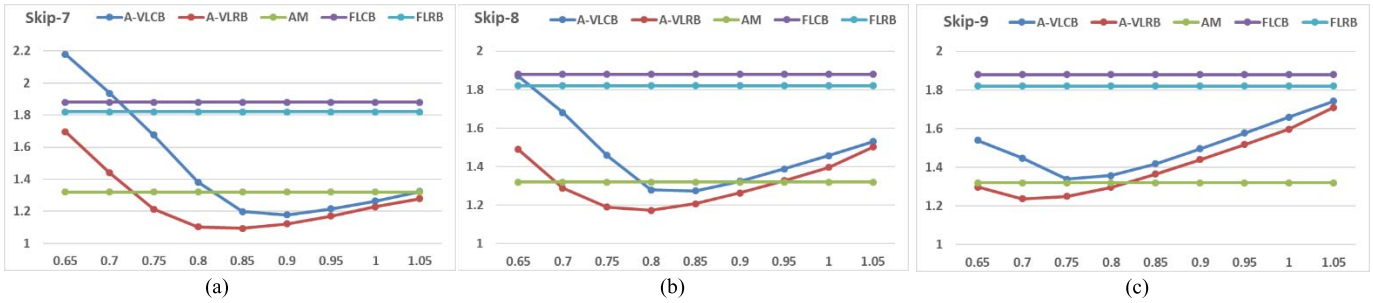


Fig. 13. Average latency comparison of AM, FLCB, FLRB, A-VLCB, and A-VLRB in 16×16 multiplier without aging. The x -axis is the cycle period, and the y -axis is the average latency.

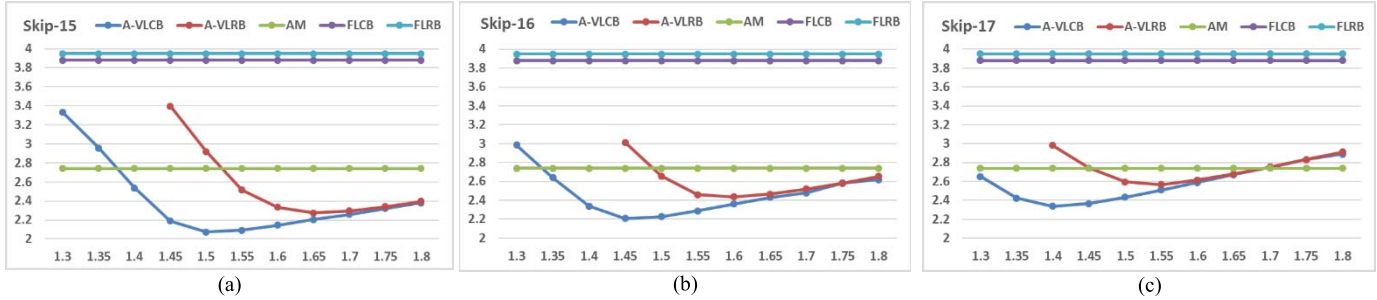


Fig. 14. Average latency comparison of AM, FLCB, FLRB, A-VLCB, and A-VLRB in 32×32 multiplier without aging. The x -axis is the cycle period, and the y -axis is the average latency.

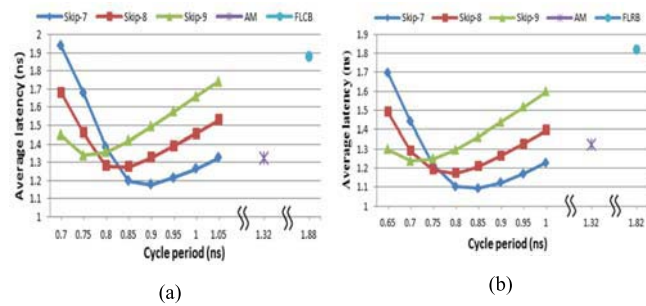


Fig. 15. Average latency of a 16×16 variable-latency bypassing multiplier under three different skip numbers. (a) 16×16 A-VLCB. (b) 16×16 A-VLRB.

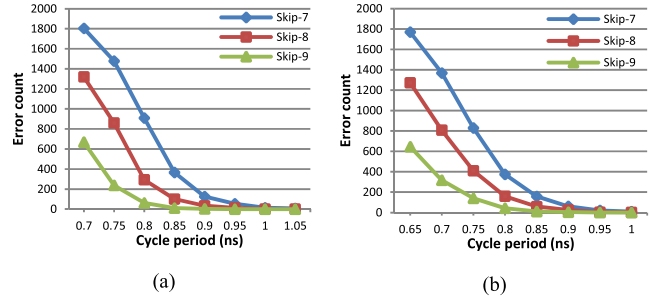


Fig. 16. Error count in 10000 cycles. (a) 16×16 A-VLCB. (b) 16×16 A-VLRB.

cycle periods. The average latency is calculated using 10000 input simulations. Because AM, FLCB, and FLRB use only fixed cycle periods, their average latency in 16×16 multipliers is a single number, 1.32, 1.88, and 1.82 ns, respectively.

In Fig. 15(a), the average latency of the three different VLCBs in the 16×16 multiplier is smaller than that of the FLCBs. Compared with the AM, the average latency can be either larger or smaller. For example, the average latency is smaller than the AM when the cycle period is larger than 0.85 ns in the 16×16 multiplier. When the cycle period in 16×16 VLCBs is larger than 0.85 ns, the average latency of Skip-7 is the smallest of the three scenarios. However, the average latency of Skip-7 is the largest of the three scenarios when the cycle period is < 0.8 ns.

Similarly, in Fig. 15(b), the average latency of these VLRB multipliers is smaller than that of the FLRBs. Compared with the AM, the average latency can be either larger or smaller. For example, the average latency is smaller than that of the AM when the cycle period is larger than 0.74 ns in the

16×16 VLRBs. The Skip-7 in the 16×16 VLRBs has the smallest average latency when the cycle periods are > 0.75 ns. However, Skip-7 has the largest average latency when the cycle periods are < 0.7 ns. This situation is similar to that of the VLCBs. The VLRB with a smaller skip number has more errors in smaller cycle periods, so the average latency is increased due to the penalties incurred. In contrast, it has fewer errors in larger cycle periods, so the VLRB with a smaller skip number, i.e., Skip-7, has the smallest average latency in large cycle periods.

This is because the path delay of one-cycle patterns in the 16×16 variable-latency bypassing multipliers with Skip-7 is larger than that of one-cycle patterns in the 16×16 variable-latency bypassing multipliers with Skip-8 and Skip-9. When the cycle period is < 0.8 ns, the 16×16 variable-latency bypassing multiplier with Skip-7 has more errors, as shown in Fig. 16(a) and (b). When an error occurs, the operation needs to be executed again using three extra cycles (one cycle for Razor flip-flops and two cycles for reexecution). More errors

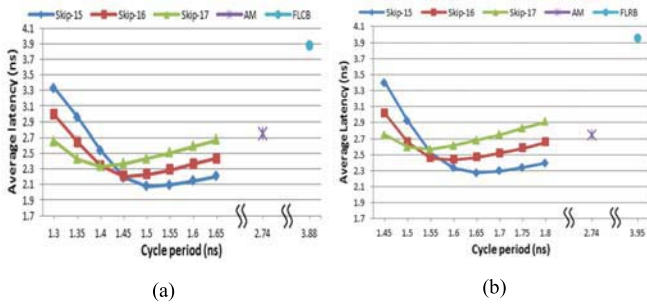


Fig. 17. Average latency of a 32×32 variable-latency bypassing multiplier under three different skip numbers. (a) 32×32 VLCB. (b) 32×32 VLRB.

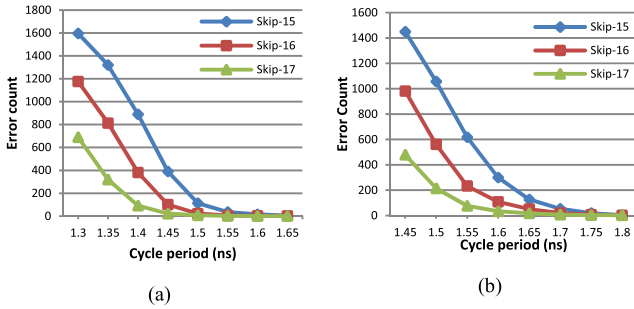


Fig. 18. Error count in 10000 cycles. (a) 32×32 VLCB. (b) 32×32 VLRB.

cause more penalties, and thus the average latency is increased. Although the percentage of one-cycle patterns in the 16×16 variable-latency bypassing multiplier with Skip-7 is the largest, it also has more errors in smaller cycle periods, and, as a result, performance is degraded.

However, when the cycle period is >0.85 ns, the error counts of the Skip-7, Skip-8, and Skip-9 variable-latency bypassing multipliers are similar, as shown in Fig. 16(a) and (b) as well. Since the percentage of one-cycle patterns in the 16×16 variable-latency bypassing multiplier with Skip-7 is higher than that of the other two, the average latency becomes the lowest with fewer error counts. On the contrary, when the cycle period increases, the input patterns with small delays will have more timing waste, leading increased average latency, as shown in Fig. 15(a) and (b).

Fig. 17(a) and (b) compares the average latency of a 32×32 variable-latency bypassing multiplier under three different numbers. Fig. 18(a) and (b) displays the error count for three different skip numbers over different cycle periods.

In 32×32 multipliers, there are also three scenarios: 1) Skip-15; 2) Skip-16; and 3) Skip-17. The percentages of one-cycle patterns are shown in Table II. The latencies of the 32×32 AM, FLCB, and FLRB are 2.74, 3.88, and 3.95 ns, respectively.

Similar to the results for the 16×16 multipliers, the average latency of the 32×32 variable latency bypassing multipliers is lower than that of the AM and much lower than that of the fixed latency bypassing multipliers if proper cycle periods are used.

In addition, the average latency of Skip-15 is the smallest of the three scenarios when the cycle period is large and the largest of the three scenarios when the cycle period is

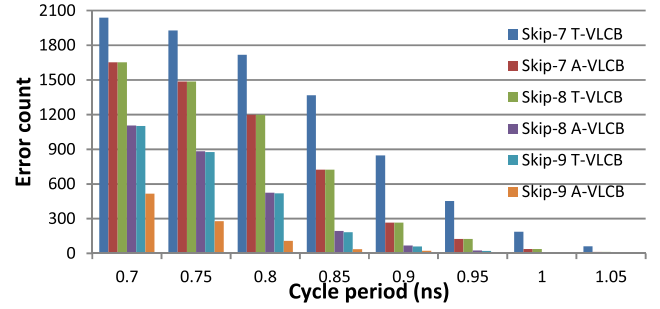


Fig. 19. Comparison of the error count between the T-VLCB and the VLCB with AHL (A-VLCB) in 16×16 multipliers.

small. The reason for this is shown in Fig. 18(a) and (b). The 32×32 variable latency bypassing multipliers with Skip-15 exhibit more errors when the cycle period is short. Therefore, their average latency is the largest of the three scenarios. However, the average latency of the 32×32 variable latency bypassing multipliers with Skip-15 is the smallest when the cycle period is long because they have less timing waste.

C. Average Latency Improved by AHL

Because of the aging effect, the circuit speed is degraded. The path delay of the one-cycle patterns becomes larger than the cycle period and causes timing violations. In this situation, the operations need to be reexecuted. If timing violations occur too frequently, the extra cycles of the reexecution will increase the average latency. In order to mitigate performance degradation, these one-cycle patterns that cause timing violations must be changed to two-cycle patterns.

Our proposed architecture changes one-cycle patterns that cause timing violations to two-cycle patterns by choosing another judging block in the AHL circuit. When the frequency of errors exceeds a threshold (10% in our experiment, that is, 10 errors for each 100 operations), the aging indicator outputs 1 to select another multiplexer input. In this situation, if originally, a multiplicand in the column-bypassing multiplier with n zeros is a one-cycle pattern, a multiplicand must have $n + 1$ zeros to become a one-cycle pattern, reducing the number of one-cycle patterns whose delay is close to the cycle period. For the row-bypassing multiplier, similarly, the criteria of one-cycle patterns must change from n zeros to $n + 1$ zeros in the multiplier.

Figs. 19 and 20 compare the error count between the traditional VLCB (T-VLCB) and our proposed VLCB with AHL (A-VLCB) in 16×16 and 32×32 multipliers, respectively, and Figs. 21 and 22 compare error count between the traditional VLRB (T-VLRB) and our proposed VLRB with AHL (A-VLRB) in 16×16 and 32×32 multipliers, respectively. It can be seen that the error count of the adaptive variable-latency multiplier is smaller than that of the T-VLCB. This is because the T-VLCB has only one judging block in the hold logic circuit, and therefore it cannot change the number of one-cycle patterns to reduce the timing violations.

Fig. 23 compares the average latency between the FLCB, FLRB, T-VLCB, T-VLRB, A-VLCB, and A-VLRB with three scenarios (Skip-7, Skip-8, and Skip-9) in 16×16 multipliers,

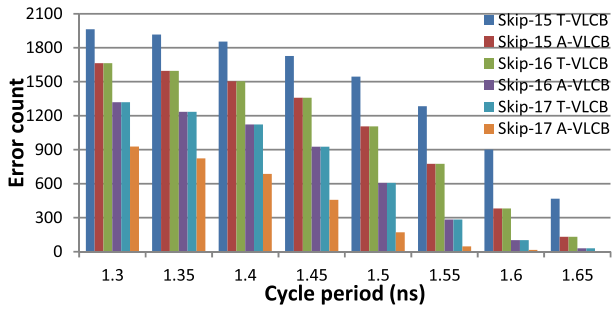


Fig. 20. Comparison of the error count between the T-VLCB and the VLCB with AHL (A-VLCB) in 32×32 multipliers.

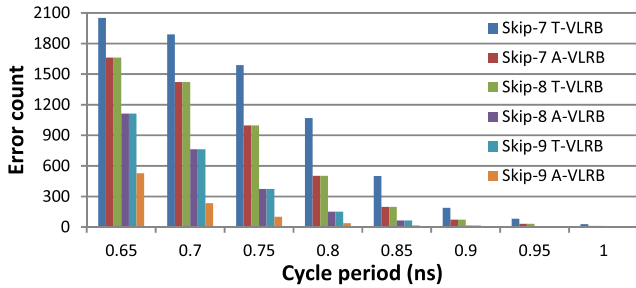


Fig. 21. Comparison of the error count between the T-VLRB and the VLRB with AHL (A-VLRB) in 16×16 multipliers.

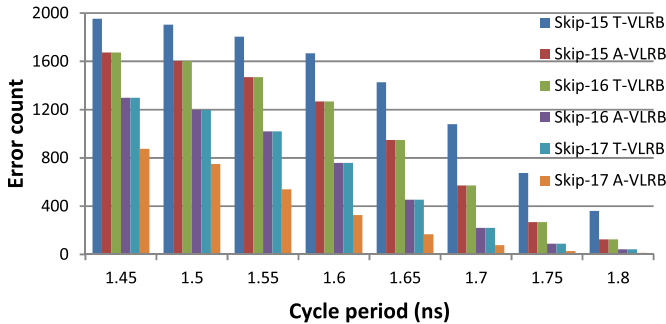


Fig. 22. Comparison of the error count between the T-VLRB and the VLRB with AHL (A-VLRB) in 32×32 multipliers.

and Fig. 24 compares the average latency in 32×32 multipliers. The threshold for the aging indicator is also set to 10%.

It can be seen the latency of variable latency bypassing with adaptive hold is equal or better than that of the variable-latency bypassing multiplier. The improved ratio is large when the cycle period is short because more timing violations occur in shorter cycle periods.

D. Area Comparison

Fig. 25 compares the normalized area of the AM, FLCB, A-VLCB, FLRB, and A-VLRB in 16×16 and 32×32 multipliers. The data are normalized to the area of the AM. In the 16×16 multiplier, the area of the A-VLCB and A-VLRB is 22.9% and 23.5% higher than FLCB and FLRB. In the 32×32 multiplier, the area of the A-VLCB and A-VLRB is 12.3% and 5.7% higher than that of the FLCB and FLRB, respectively. This is because when a fixed-latency bypassing multiplier is changed to a variable-latency bypassing multiplier, additional circuits are needed for AHL and Razor flip-flops to ensure the

correct operations of the multiplier after degradation. Note that the increased area overhead ratio of the 32×32 A-VLCB and A-VLRB is much smaller than that of the 16×16 A-VLCB and A-VLRB. This is because AHL and Razor flip-flops both occupy a smaller area ratio in larger multipliers.

E. Latency, Power, and Energy-Delay Product (EDP) Comparison of FLCB, A-VLCB, FLRB, and A-VLRB Over Seven Years

Fig. 26(a)–(c) compares the latency, power, and EDP of the AM, FLCB, FLRB, A-VLCB, and A-VLRB in 16×16 multipliers from year 0 to year 7. The average latency is normalized to the latency of AM at year 0. To make comparison simpler, the cycle period of the A-VLRB and A-VLCB is set to 1.2 ns, and the skip number is 7, and therefore, no timing violations occur, and the average latency of the A-VLCB and A-VLRB are similar.

In Fig. 26(a), it can be seen the average latency is increased due to the aging effect. The latency of the AM, FLCB, and FLRB increase by 15.2%, 14.36%, and 14.83%, respectively. However, the average latency of the A-VLCB and A-VLRB only increase by 2.76% and 3.47%, respectively. The increased latency ratio of the variable-latency multiplier is much less than that of the fixed-latency multiplier. This is because the variable-latency bypassing multiplier is less sensitive to the increased delay caused by the aging effect.

In addition, it can be seen the average latency of the AM is larger than the adaptive variable-latency multiplier after two years, and the average latency of the 16×16 A-VLCB and A-VLRB is 32.6%–50.1% and 28.8%–45.4% lower than that of the FLCB and FLRB, respectively. Therefore, significant latency reduction can be achieved using adaptive variable-latency multipliers.

Fig. 26(b) compares the power of 16×16 AM, FLCB, FLRB, A-VLCB, and A-VLRB over seven years. To make comparison fair, the power of AM, FLCB, and FLRB includes the power of flip-flops at the input and output, and the power of A-VLCB and A-VLRB includes the power of flip-flops at the input and the power of Razor flip-flops at the output. It can be seen that the power consumption decreases progressively because the transistor threshold voltage increases due to the aging effect. It can also be seen that the AM has the largest average power and that the average power of the fixed-latency multiplier is less than its corresponding variable-latency multiplier (the power of FLCB is 12.1%–12.6% less than that of A-VLCB, and the power of FLRB is 7.1%–12.3% less than that of A-VLRB, on average). This is because the fixed-latency multiplier uses the bypassing technique (discussed in Section II) to reduce power consumption. Compared with the fixed-latency multiplier, the variable-latency multiplier has higher power due to more complicated circuits. However, the variable-latency multiplier still has less power than that of the AM because it uses both the clocking gating and a bypassing power reduction technique.

Moreover, the power of the 16×16 A-VLRB is larger than that of the 16×16 A-VLCB. This is because the row-bypassing multiplier is more complicated than the

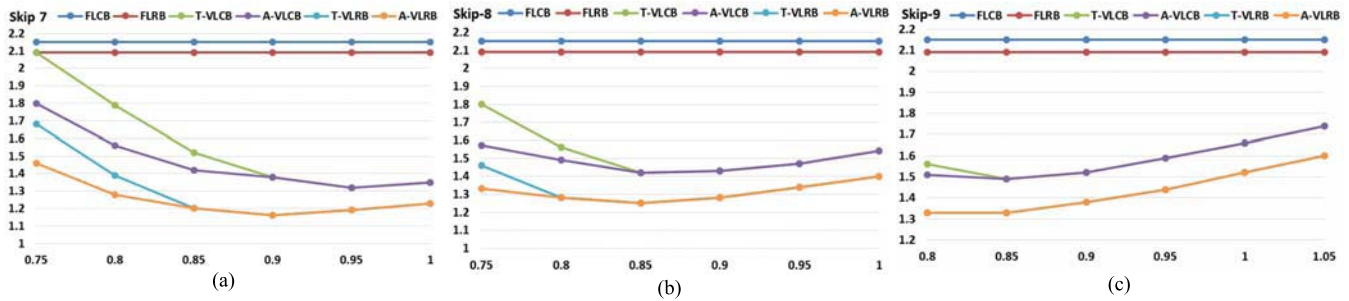


Fig. 23. Average latency comparison between 16×16 adaptive and traditional variable-latency multipliers. (a) Skip-7. (b) Skip-8. (c) Skip-9. The x -axis is the cycle period, and the y -axis is the average latency.

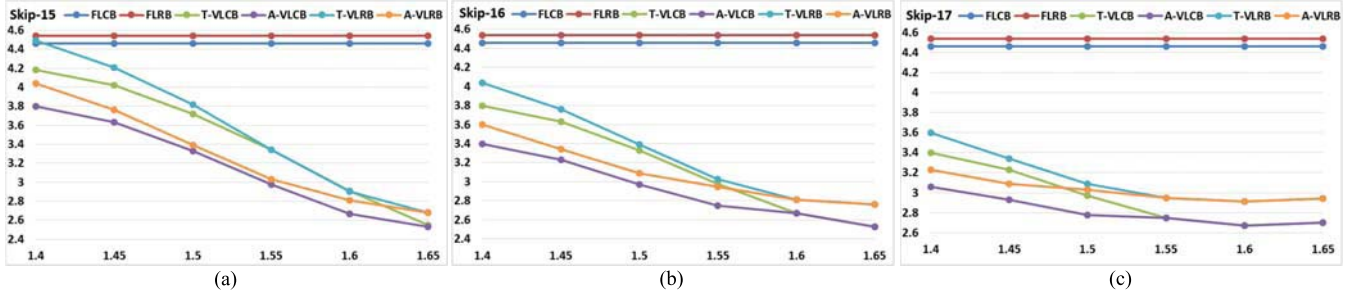


Fig. 24. Average latency comparison between 32×32 adaptive and traditional variable-latency multipliers. (a) Skip-15. (b) Skip-16. (c) Skip-17. The x -axis is the cycle period, and the y -axis is the average latency.

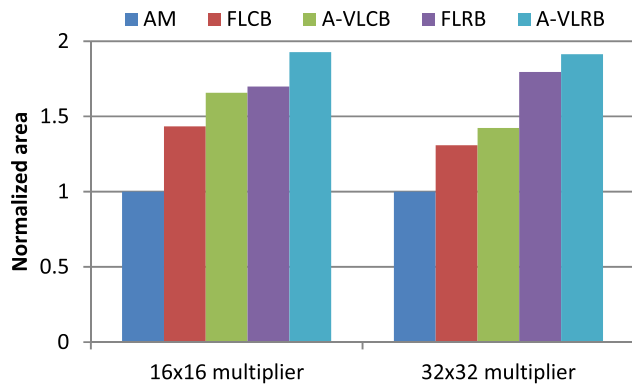


Fig. 25. Area overhead in transistors for AM, FLCB, A-VLCB, FLRB, and A-VLRB in 16×16 and 32×32 multipliers.

column-bypassing multiplier and because the area overhead of the row-bypassing multipliers is larger than that of the column-bypassing multipliers, which results in more power consumption.

Fig. 26(c) compares the EDP of AM, FLCB, A-VLCB, FLRB, and A-VLRB in 16×16 multipliers over seven years. The EDP also decreases progressively because the transistor threshold voltage increases due to the aging effect. It can be seen that the EDP of the A-VLRB is higher than that of the AM in year 0 and lower than the AM after year 2. The average reduction is 3.6%. The EDP of the A-VLCB is a little higher than that of the AM in year 0 and lower than that of the AM after year 1. The average reduction is 10.1%. Hence, variable-latency multipliers can achieve the lowest average EDP mainly because variable-latency multipliers have an average latency that is similar to that of the AM but characterized by lower power than the AM.

Fig. 27(a)–(c) compares the average latency, power, and EDP between the AM, FLCB, FLRB, A-VLCB, and A-VLRB in the 32×32 multipliers from year 0 to year 7. The average latency is normalized to the latency of the AM at year 0. The cycle period of the A-VLRB and A-VLCB is set to 2.3 ns, and the skip number is 7. Therefore, no timing violation occurs, and the average latency of the A-VLCB and A-VLRB is similar.

In Fig. 27(a), it can be seen that the average latency is increased each year also due to the aging effect. The average latency of the AM, FLCB, and FLRB increases by 15.0%, 14.9%, and 14.9%, respectively. However, the average latency of the A-VLCB and A-VLRB only increases by 1.3% and 0.98%, which proves the variable-latency bypassing multiplier is also less sensitive to the aging effect.

Similar to Fig. 26(a), the average latency of the variable-latency bypassing multiplier is significantly less than that of the fixed-latency multiplier (the average latency of the A-VLCB is 31%–50.7% less than that of the FLCB, and the average latency of the A-VLRB is 33.2%–53.6% less than that of the average latency of the FLRB). The average latency of the variable-latency bypassing multipliers is higher than that of the AM in year 0 and lower than that of the AM after year 2.

Fig. 27(b) compares the power of the 32×32 AM, FLCB, FLRB, A-VLCB, and A-VLRB over seven years. To make comparison fair, the power of the AM, FLCB, and FLRB includes the power of flip-flops at the input and output, and the power of A-VLCB and A-VLRB includes the power of flip-flops at the input and the power of Razor flip-flops at the output. It can be seen that the power consumption decreases progressively each year due to the aging effect and increased

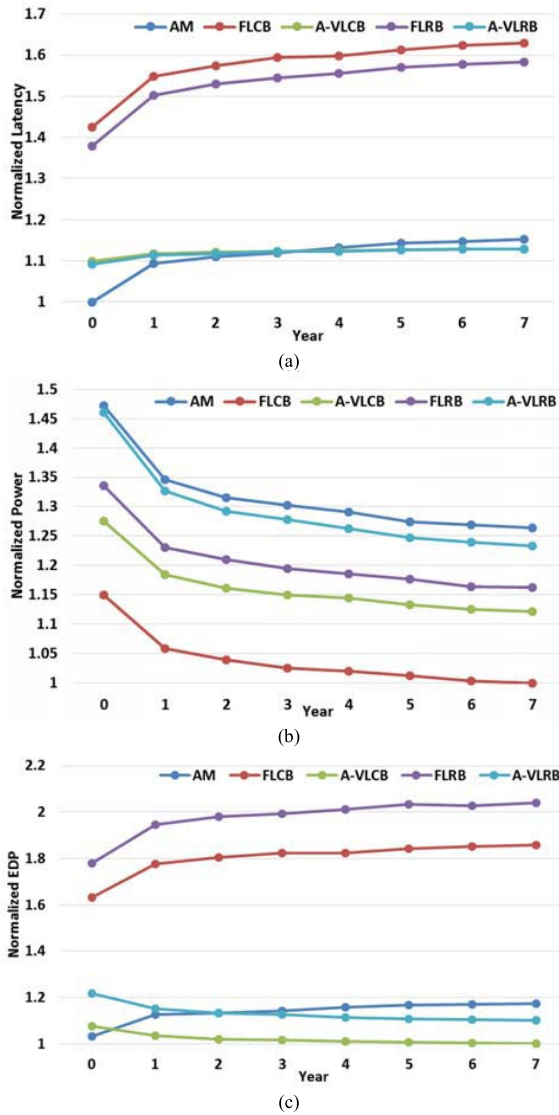


Fig. 26. Normalized delay power and EDP comparison over seven years in the 16×16 multiplier. (a) Normalized latency. (b) Normalized power. (c) Normalized EDP.

transistor threshold voltage. Similarly, it can be seen that the AM has the largest average power and that the average power of the fixed-latency multiplier is less than its corresponding variable-latency multiplier (the power of the FLCB is 14.8%–15.2% less than that of A-VLCB, and the power of the FLRB is 9.1%–14.5% less than that of the A-VLRB). Moreover, the power of the 32×32 A-VLRB is larger than that of the 32×32 A-VLCB. This is because the area overhead of the row-bypassing multipliers is larger than that of the column-bypassing multipliers. Greater area overhead incurs more power consumption.

Fig. 27(c) compares the EDP of the FLCB, A-VLCB, FLRB, and A-VLRB in 32×32 multipliers over seven years. The EDP decreases progressively because the transistor threshold voltage increases due to the aging effect. It can be seen that the EDP of the A-VLRB is higher than that of the AM, and lower than that of the AM after year 2. The average reduction is 1.1%. The EDP of the A-VLCB is a little higher

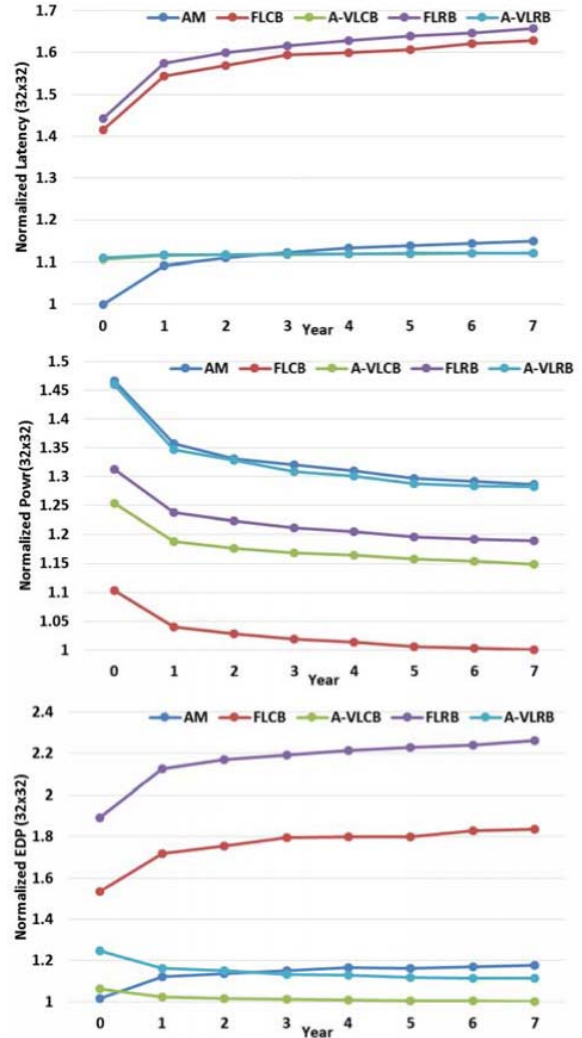


Fig. 27. Normalized delay power and EDP comparison over seven years in the 32×32 multiplier. (a) Normalized latency. (b) Normalized power. (c) Normalized EDP.

than that of the AM in year 0 and lower than that of the AM after year 1. The average reduction is 10.45%. In summary, the variable-latency multiplier can achieve the lowest average EDP compared to the AM and fixed-latency bypassing multipliers.

V. CONCLUSION

This paper proposed an aging-aware variable-latency multiplier design with the AHL. The multiplier is able to adjust the AHL to mitigate performance degradation due to increased delay. The experimental results show that our proposed architecture with 16×16 and 32×32 column-bypassing multipliers can attain up to 62.88% and 76.28% performance improvement compared with the 16×16 and 32×32 FLCB multipliers, respectively. Furthermore, our proposed architecture with the 16×16 and 32×32 row-bypassing multipliers can achieve up to 80.17% and 69.40% performance improvement compared with the 16×16 and 32×32 FLRB multipliers. In addition, the variable-latency bypassing multipliers exhibited the lowest average EDP and achieved up to 10.45% EDP reduction in 32×32 VLCB multipliers. Note that in addition to the

BTI effect that increases transistor delay, interconnect also has its aging issue, which is called electromigration. Electromigration occurs when the current density is high enough to cause the drift of metal ions along the direction of electron flow. The metal atoms will be gradually displaced after a period of time, and the geometry of the wires will change. If a wire becomes narrower, the resistance and delay of the wire will be increased, and in the end, electromigration may lead to open circuits. This issue is also more serious in advanced process technology because metal wires are narrower, and changes in the wire width will cause larger resistance differences. If the aging effects caused by the BTI effect and electromigration are considered together, the delay and performance degradation will be more significant. Fortunately, our proposed variable latency multipliers can be used under the influence of both the BTI effect and electromigration. In addition, our proposed variable latency multipliers have less performance degradation because variable latency multipliers have less timing waste, but traditional multipliers need to consider the degradation caused by both the BTI effect and electromigration and use the worst case delay as the cycle period.

REFERENCES

- [1] Y. Cao. (2013). *Predictive Technology Model (PTM) and NBTI Model* [Online]. Available: <http://www.eas.asu.edu/~ptm>
- [2] S. Zafar *et al.*, "A comparative study of NBTI and PBTI (charge trapping) in SiO₂/HfO₂ stacks with FUSI, TiN, Re gates," in *Proc. IEEE Symp. VLSI Technol. Dig. Tech. Papers*, 2006, pp. 23–25.
- [3] S. Zafar, A. Kumar, E. Gusev, and E. Cartier, "Threshold voltage instabilities in high-k gate dielectric stacks," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 1, pp. 45–64, Mar. 2005.
- [4] H.-I. Yang, S.-C. Yang, W. Hwang, and C.-T. Chuang, "Impacts of NBTI/PBTI on timing control circuits and degradation tolerant design in nanoscale CMOS SRAM," *IEEE Trans. Circuit Syst.*, vol. 58, no. 6, pp. 1239–1251, Jun. 2011.
- [5] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pMOS NBTI effect for robust nanometer design," in *Proc. ACM/IEEE DAC*, Jun. 2004, pp. 1047–1052.
- [6] H. Abrishami, S. Hatami, B. Amelifard, and M. Pedram, "NBTI-aware flip-flop characterization and design," in *Proc. 44th ACM GLSVLSI*, 2008, pp. 29–34.
- [7] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "NBTI-aware synthesis of digital circuits," in *Proc. ACM/IEEE DAC*, Jun. 2007, pp. 370–375.
- [8] A. Calimera, E. Macii, and M. Poncino, "Design techniques for NBTI-tolerant power-gating architecture," *IEEE Trans. Circuits Syst., Exp. Briefs*, vol. 59, no. 4, pp. 249–253, Apr. 2012.
- [9] K.-C. Wu and D. Marculescu, "Joint logic restructuring and pin reordering against NBTI-induced performance degradation," in *Proc. DATE*, 2009, pp. 75–80.
- [10] Y. Lee and T. Kim, "A fine-grained technique of NBTI-aware voltage scaling and body biasing for standard cell based designs," in *Proc. ASP-DAC*, 2011, pp. 603–608.
- [11] M. Basoglu, M. Orshansky, and M. Erez, "NBTI-aware DVFS: A new approach to saving energy and increasing processor lifetime," in *Proc. ACM/IEEE ISLPED*, Aug. 2010, pp. 253–258.
- [12] K.-C. Wu and D. Marculescu, "Aging-aware timing analysis and optimization considering path sensitization," in *Proc. DATE*, 2011, pp. 1–6.
- [13] K. Du, P. Varman, and K. Mohanram, "High performance reliable variable latency carry select addition," in *Proc. DATE*, 2012, pp. 1257–1262.
- [14] A. K. Verma, P. Brisk, and P. Jenne, "Variable latency speculative addition: A new paradigm for arithmetic circuit design," in *Proc. DATE*, 2008, pp. 1250–1255.
- [15] D. Baneres, J. Cortadella, and M. Kishinevsky, "Variable-latency design by function speculation," in *Proc. DATE*, 2009, pp. 1704–1709.
- [16] Y.-S. Su, D.-C. Wang, S.-C. Chang, and M. Marek-Sadowska, "Performance" optimization using variable-latency design style," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 10, pp. 1874–1883, Oct. 2011.
- [17] N. V. Mujadiya, "Instruction scheduling on variable latency functional units of VLIW processors," in *Proc. ACM/IEEE ISED*, Dec. 2011, pp. 307–312.
- [18] M. Olivieri, "Design of synchronous and asynchronous variable-latency pipelined multipliers," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 4, pp. 365–376, Aug. 2001.
- [19] D. Mohapatra, G. Karakonstantis, and K. Roy, "Low-power process-variation tolerant arithmetic units using input-based elastic clocking," in *Proc. ACM/IEEE ISLPED*, Aug. 2007, pp. 74–79.
- [20] Y. Chen, H. Li, J. Li, and C.-K. Koh, "Variable-latency adder (VL-Adder): New arithmetic circuit design practice to overcome NBTI," in *Proc. ACM/IEEE ISLPED*, Aug. 2007, pp. 195–200.
- [21] Y. Chen *et al.*, "Variable-latency adder (VL-Adder) designs for low power and NBTI tolerance," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 11, pp. 1621–1624, Nov. 2010.
- [22] M.-C. Wen, S.-J. Wang, and Y.-N. Lin, "Low power parallel multiplier with column bypassing," in *Proc. IEEE ISCAS*, May 2005, pp. 1638–1641.
- [23] J. Ohban, V. G. Moshnyaga, and K. Inoue, "Multiplier energy reduction through bypassing of partial products," in *Proc. APCCAS*, 2002, pp. 13–17.
- [24] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Impact of NBTI on the temporal performance degradation of digital circuits," *IEEE Electron Device Lett.*, vol. 26, no. 8, pp. 560–562, Aug. 2005.
- [25] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Negative bias temperature instability: Estimation and design for improved reliability of nanoscale circuit," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 26, no. 4, pp. 743–751, Apr. 2007.
- [26] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pMOS NBTI effect for robust nanometer design," in *Proc. 43rd ACM/IEEE DAC*, Aug. 2006, pp. 1047–1052.
- [27] D. Ernst *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proc. 36th Annu. IEEE/ACM MICRO*, Dec. 2003, pp. 7–18.



Ing-Chao Lin (M'09) received the M.S. degree in computer science from the National Taiwan University, Taipei, Taiwan, and the Ph.D. degree in computer science and engineering from Pennsylvania State University, State College, PA, USA, 2007.

He was with Real Intent, Inc., Sunnyvale, CA, USA, from 2007 to 2009, and since 2009, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, where he is currently an Assistant Professor. His current research interests

include VLSI design and computer-aided design for nanoscale silicon, low-power reliable system design, and computer architecture.



Yu-Hung Cho received the B.S. degree in electronic engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, and the M.S. degree in computer science and information engineering from the National Cheng Kung University, Tainan, Taiwan, in 2010 and 2012, respectively.

He is currently with Welltrend Semiconductor, Inc., Kaohsiung, Taiwan. His current research interests include reliability-aware design for VLSI.



Yi-Ming Yang received the B.S. degree in computer science from the National Taichung University of Education, Taichung, Taiwan, and the M.S. degree in computer science and information engineering from the National Cheng Kung University, Tainan, Taiwan, in 2011 and 2013, respectively.

He is currently with VIA Technologies, Inc., New Taipei, Taiwan. His current research interests include reliability-aware design for VLSI.