Mitigation of Aging Effects Through Selective Time-borrowing and Alternative Path Activation

Andrei Silva
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
andreisantos@ufmg.br

Frank Sill
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais
franksill@ufmg.br

ABSTRACT
Integrated digital circuits are frequency capped by its heavily constrained paths between flip-flop stages. These so-called critical paths are highly susceptible to delay fluctuations leading designers to use guard-banding in order to avoid timing violations. Several effects can cause these variations, whereas aging is of rising importance. Many works have addressed this issue through monitoring of critical paths or techniques for error detection. The consequent error correction, however, requires the interruption of the circuit’s operation for restoration of correct values, resulting into a performance drop. This work proposes two strategies tackling this problem without cycle loss: A time-borrowing approach that redistributes the slack between a critical path and its most constrained fan out paths; and the technique Alternative Path Activation (APA) which applies duplication of the most time-constrained fan out paths of a critical path. Simulations were conducted for several test circuits indicating its feasibility. Further, the proposed approach was implemented in ARMv2 based processor core, resulting in an enhanced robustness against aging induced delay variations of 7.2\% at the cost of 2.5\% area and 2.7\% power overhead.

CCS CONCEPTS
• Computer systems organization → Reliability; Availability; Redundancy;

KEYWORDS
Aging; Timing faults, Time-borrowing, Processor

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SBCCI ’17, August 28-September 01, 2017, Fortaleza - Ceará, Brazil
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5106-5/17/08... $15.00
https://doi.org/10.1145/3109984.3110006

1 INTRODUCTION
Advances in microelectronics, microfabrication and design methodologies have led current integrated circuits to high complexity levels in computing and electronics systems. The driving force behind the tendency is the reduction of the technology sizes which already reached nanometer scale. The consequences of this increase in miniaturization, however, is now becoming very apparent. This includes, amongst others, problems such as Electromigration, radiation induced soft errors and threshold voltage (Vth) drift due to aging. Hence, there is rising need for solutions on design level to mitigate these degradation effects [1].

For digital circuits, aging effects manifest themselves mainly in the form of increased delay of logic elements, resulting into an increase of the time a signal requires to traverse a logic path. In Critical Paths (CPs), these delay shifts can cause severe timing violations, leading the entire system to fail or to experience unexpected behavior. Conservative approaches dealing with these problems usually entail worst case guard-banding, i.e., the addition of safety margins to supply voltage, cell sizes and maximum clock frequency. However, this results into pessimistic circuits with increase in terms of area and power consumption and reduction in performance [2].

A technically more sophisticated solution is the monitoring and treatment of aging effects in digital circuits. In [3], the authors conduct critical paths analysis and selection using machine learning and statistical methods in order to keep track of the circuit performance. An adaptive supply voltage scaling technique is proposed in [4] that reduces circuit aging due to Negative-Bias Temperature Instability (NBTI). The authors in [2] propose a self-tuning technique that adjusts circuits parameters progressively over time to deal with the pessimistic guard-banding. However, all these techniques are not able neither to detect nor to correct timing errors.

Approaches for detection and correction of timing errors are usually based on delay aware flip-flops that detects timing violations via shadow latches and emit an error signal. The work of [5] introduces the Razor Flip-Flop (RFF), a sequential element capable of detecting timing violation. The RFF has been explored and expanded through other works such as [6], [7], [8] and [9]. The main drawback of the proposed solutions, however, is the required processing of the detected timing violation on higher system layers, leading to reduced performance, also known as cycle-loss.

This work proposes two combined approaches for mitigation of aging induced timing violations without cycle-loss. The first technique applies time-borrowing in order to redistribute the slack between the CPs and the logical paths that follow in the next sequential stage. The second approach, named Alternative Path Activation technique (APA), deals with CPs that are followed by Highly Constrained Fan Out Paths (HCFOP), preventing any time-borrowing. The paper is organized as follows. Section 2 presents the state-of-the-art regarding monitoring and handling of timing violations in sequential circuits. Section 3 presents the proposed structures.
and its characteristics. Section 4 focuses on the integration of the proposed structures into a design flow, while section 5 discusses the simulation results. Finally, Section 6 concludes this paper.

2 PRIOR WORKS
Timing violations of logical paths have been the focus of several works related to reliability issues. Current works vary from monitoring of circuits parameters with error handling at the system’s level to hardware solutions and low level handling of timing violations.

In [3], the authors devise a method to select a sample from a large pool of critical paths in a circuit using machine learning algorithms and linear algebra. Since monitoring requires sensors and delay testing techniques, the selection of these paths to represent the entire circuit becomes a crucial endeavor to reduce area and power overhead.

The authors in [4] propose the Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) techniques to compensate for the drift in the transistor parameters due to aging, maintaining the circuits performance through the control of the body bias of both transistors flavors. The methods showed little area and power overhead while maintaining nominal circuit operation. The work of [2] also proposes the use of parameters tuning throughout the circuit’s lifetime, dealing with supply voltage, operation frequency and cooling systems.

The Razor Flip-Flop (RFF) is a well-known and referenced device presented in [5] for delay aware sequential elements. The RFF is composed of a main FF and a shadow latch that samples the input with a delayed clock. An XOR gate compares the output of the main FF with the shadow latch and, if both are different, an error signal is generated. An extra clock cycle is then used to restore the correct data into the main FF.

The problem of the RFF is its dependency on an entire clock cycle to restore the correct value to the output. In [6], the TIMBER FF (TFF) uses a similar structure that does not require the extra cycle to restore the correct value to the FF. It uses instead a time-borrowing approach, with the time required to restore the correct data taken from part of the next clock cycle. This approach yields good results, taken that the logical path that follows the TFF has sufficient time to deal with the delayed data in the same cycle.

The authors in [9] propose a solution that enhances the time-borrowing approach using the Error Avoidance Flip-Flop (EAAF). In this work, the authors create a slack-aware redistribution technique that analyses the fan out stages of the critical paths and places the EAAF accordingly. Their results show a high critical path coverage and improvement of the timing margin by 11% in an industrial processor.

3 PROPOSED APPROACH
We propose two techniques that tackles the delay shifts caused by aging in sequential circuits. The first is based on a Time-Borrowing Flip-Flop (TBFF), a cell to be used in heavily constrained logic path for time balancing between critical paths and its fan outs sequential stages; and the second is a strategy that works through the activation of an alternative logic path during the detection of timing violations.

3.1 Selective Time-borrowing
The main idea of the TBFF is the delayed sampling of the result of a CP without any error detection with focus on aging induced delay faults, i.e., it is assumed that a delayed value is always the correct value. That means, unlike Timber FF [6] and EAAF [9], the TBFF always works with a delayed clock. Consequently, there is no need for a control system, leading to great reduction of area overhead, since no shadow cell, comparator and multiplexer is needed. On the downside, there is lower applicability against transient faults, e.g., due to radiation effects [10].

Figure 1 depicts the structure of the TBFF. It samples the data on the D pin with a \( N \cdot t_{buffer} \) seconds delay in relation to the system’s clock. This causes the output (Q pin) to be also delayed by this time, culminating in the time-borrowing from the next stage. This delayed sampling of the data creates a longer clock period for the critical path, increasing its available time (slack).

Figure 2: Comparison of TBFF and type-D FF (DFF) for delayed input data

The timing diagram in Figure 2 compares the behavior of a standard type-D FF (DFF) with the proposed TBFF for delayed data at the input D of both FF. While the DFF fails to sample the correct data, i.e., the output stays at “0”, the output of the TBFF reads the correct value “1”.

An important premise for this method is that the critical path precedes only non-critical ones. That is, the logical paths processing the output of the CP have to have enough slack to process the delayed data (at least \( N \cdot t_{buffer} \) seconds). This premise calls for a timing analysis of the circuit for automatic placement of the TBFFs. Such flow is presented in the following Section 4.

The advantage of the TBFF is its independence from external control and reduced routing overhead, diminishing the risk of congestions [11]. Each TBFF has its own set of delaying buffers, dimensioned accordingly to the most time constrained paths that follows.
Since only the most critical paths in a circuit are expected to receive these devices, no major overhead in area and consumption is expected. It should be noted, though, that the increase of the buffer delay due to aging must be considered during slack estimation.

### 3.2 Alternative Path Activation

The approach Alternative Path Activation (APA) applies a modified flip-flop that is capable of detecting timing violations in a critical path. The Alternative Path Activation Flip-Flop (APAFF) detects the delayed data by a second FF (shadow FF) sampling the input D pin with a delayed clock, much like the TFF presented in [6]. The difference here is that the delayed clock is generated internally in the cell similar to the TBFF. If the data sampled from the shadow FF differs from the one sampled by the main FF, an error signal is generated.

The diagram in Figure 3 shows the architecture of the APAFF, with the Shadow FF sampling the D input with a N buffer delayed clock. The XOR generates the error signal whenever both FFs differs in output state. The Shadow FF was chosen in place of a Latch in order to prevent metastability issues.

The great advantage of this method is its parallel characteristic. Since both possible results of the critical path (Q and Q') are being processed at the same time by exact copies, the error correction has to provide only the selection of the path with the correct data. Thus, the data of the critical path can arrive with a maximum delay that is identical to the delay of the copy minus the delay required for error detection. The down-side of this technique is the required area overhead.

Unconstrained fan out paths, i.e. paths with sufficient slack, can receive its input from a multiplexer connected to both Q and Q' that is located directly after the APAFF. Hence, these paths don’t require a copy, reducing the area overhead.

![Figure 3: Architecture of an Alternative Path Activation Flip-Flop (APAFF)](image)

This technique is applied to critical paths followed by Heavily Time Constrained Fan Out Paths (HCFOP), i.e. paths with delays equal or close to the critical paths delay, which diminishes or even prevents the use of time-borrowing. Two possible ways of using this technique were devised: Alternative Parallel Processing (APARP) and the Alternative Pre-Processed (APREP).

#### 3.2.1 The Alternative Parallel Processing Topology

In the APARP topology, the APAFF has its input connected at the end of a critical path and its both outputs (Q and its inverse Q') connected to exact duplicates of the HCFOP or parts of the HCFOP (see Figure 4). The outputs of both copies are connected to a multiplexer (MUX). In normal operation, the error flag is at “0”, signaling the MUX that the copy connected to output Q is bearing the correct value. During an error situation, the APAFF will change the error flag to “1”, switching the MUX to the copy of the circuit that processes the Q’ result.

Unconstrained fan out paths, i.e. paths with sufficient slack, can receive its input from a multiplexer connected to both Q and Q’ that is located directly after the APAFF. Hence, these paths don’t require a copy, reducing the area overhead.

![Figure 4: Structure of the APARP topology](image)

The timing diagram depicted in Figure 5 illustrates the sequence of the APARP topology. During normal operation, the APAFF works as a common FF, sampling the input at the system clock. In the third cycle, a delayed value is detected by the Shadow FF, setting the error pin to “1”. The error signal controls the MUX at the end of the double HCFOP's and switches the output of the path that processed output Q to one that processed output Q’, correcting the signal arriving at the next flip-flop stage.

#### 3.2.2 The Alternative Pre-Processed Topology

The APREP topology uses data preprocessing to speed up the HCFOP. The duplicated logical paths are inputted by fixed “0” and “1”. The APAFF then uses the result of the CP to select which duplicate bears the correct value. Figure 6 depicts the architecture of the APREP approach. Here, the copies of the HCFOP have its input connected to “0” and “1”, respectively, while paths with sufficient slack receive the data directly from output Q.
3.2.3 Example. Both topologies require the HCFOPs to be wrapped with MUXs controlled by the APAFF so that the corrected value is passed on to other cells connected to the HCFOP. Figure 7a depicts an example circuit, with a HCFOP composed of six cells traced by the red arrow.

Figure 7b presents the design after application of the APARP technique. The HCFOP has been duplicated and multiplexers (indicated by red circles) have been inserted wherever the HCFOP fans out to other cells not belonging to the HCFOP. All multiplexers are connected to the APAFF (blue circle) error pin at the beginning of the HCFOP (not shown).

\[
\tau_{\text{balanced}} = \frac{\tau_{\text{CP}} + \tau_{\text{HCFOP}}}{2}
\]

(1)

With \(\tau_{\text{CP}}\) and \(\tau_{\text{HCFOP}}\) the arrival times for CP and HCFOP, respectively.

Hence, the clock of the TBFF that connects to the input of CP shall be delayed by the difference between \(\tau_{\text{CP}}\) and \(\tau_{\text{balanced}}\). Thus, the buffer amount \(N_{\text{buffers}}\) results from:

\[
N_{\text{buffers}} = \left\lfloor \frac{(\tau_{\text{critical}} - \tau_{\text{balanced}})}{\tau_{\text{buffer}}} \right\rfloor
\]

(2)

Where \(\tau_{\text{buffer}}\) refers to the delay created by a single buffer cell.

Since the time balancing operation increases the slack, other paths may become more critical than the one that received the TBFF. Hence, this procedure must be repeated until the worst path of the entire circuit is a balanced one, that is, the circuit can no longer benefit from the redistribution of slack.

When a critical path is found with no possible slack redistribution between CP and HCFOP, the APA technique is applied. The flow for such procedure is depicted in Figure 8.

When dealing with aging influence on a circuit, other logical paths not detected as critical during synthesis and simulation phase may become constrained due to increase delay. In order for these paths to be covered by the techniques a set of “almost critical” paths may become candidate for the TBFF or APA placement in case they become constrained during operation. The number of paths for the set, however, has to be considered due to increase in area and power overheads.

It is also possible to devise three approaches for the buffer insertion phase: Via direct insertion into each FF, via global delayed clock to be distributed among the FFs that uses the same delay, or via clock tree synthesis. The difference between area overhead and issues due to placement and routing congestions, however, requires further analysis.

5 RESULTS

This sections presents results of simulations for two test circuits and a ARMv2 based processor. All results are based on a commercial 130nm technology library. The circuits have been simulated with...
the tool Cadence Virtuoso, while the processor was synthesized by Cadence RTL compiler.

5.1 Test Circuits

The first test design is a two-stage sequential circuit composed of inverter chains (INVCHAIN circuit). The first stage is composed of 20 inverters serially connected, creating the critical path to be tested. For the TBFF test, the second stage made up of 6 inverters, simulating a non-critical fan out from the first stage. For the APA test, on the other hand, the second stage is formed by 18 inverters, simulating the HCFOP. The diagram of the test circuits for the TBFF and APA is depicted in Figures 9a and 9b, respectively.

The maximum delay of the 20-inverters chain was simulated with 442 ps and of the 6-inverter chain with 135 ps. Thus, the balanced time results to 289 ps, which is achieved by the insertion of 3 buffers, having a delay of 50 ps each, at the TBFF.

In order to create timing violations, the clock frequency of the circuits was successively increased. The results revealed a maximum frequency increase from 1.55 GHz in the normal circuit to 1.93 GHz if time-borrowing via TBFF is applied. Thus, the circuit can deal with an aging induced delay increase of up to 24.5 % with 15.8 % and 12 % of area and power overhead, respectively.

The diagram in Figure 10 shows the simulation results for the INVCHAIN at a clock frequency of 1.93 GHz. The D pin of the FF at the start of the first stage is set to “1” before the first simulation clock. The data is processed by the 20-inverter chain to be sampled by the second FF at the next clock cycle. At 1.93 GHz, the circuit with a standard type-D FF (DFF) misses the 20-inverters result, passing to the next stage a “0” value, both indicated with a red “X”.

In contrast, the TBFF successfully passes on the “1” to the fan out path (shown by the “TBFF CP Result” and “DFF CP Result”). The FO Result are the fan out response of the circuit, represented in Figure 9 as the Test Output pins. By the third clock, the correct result is “1”, which the TBFF circuit achieved to obtain.

In the next step, the APARP topology was applied for the 18-inverter chain and the circuit was simulated with successively increasing clock frequency. The results indicate a maximum frequency increase to 1.74 GHz. That means, the circuit is fully functional up to an increased delay of the critical path of 12.2 %. The area and power overhead is of 54 % and 74.8 %, respectively.

The second test circuit used is a 4:2 compressor composed of three 8-bit Ripple Carry Adders (RCA). A compressor is a circuit composed by a tree of adders, usually employed in multiplication blocs. In a 4:2 topology, four numbers are reduced to two and then added together for the final result [12]. The test circuit employed has the two phases separated sequentially, with the final result ready after two clock cycles. Even though one of the great advantages of the compressor tree is the elimination of carry propagation, the

Figure 8: Flow for TBFF insertion and APA application

Figure 9: Test circuits simulated for both techniques

Figure 10: Simulation results for the INVCHAIN test circuit at 1.93 GHz
test circuit uses RCA in order to present a clear critical path in the circuit: The logical path that passes the carry signal.

For the TBFF test, the maximum delay of the critical path is 1,700 ps and of the related fan out full adder path is 118.5 ps. The balanced time is 909.0 ps, rendering the need for sixteen buffers. Due to application of TBFF, the maximum frequency for the compressor was increased from 518.1 MHz to 547.0 MHz. Hence, the circuit can be duplicated by the APA technique and the maximum frequency over the original maximum. The area and power was increased by 1.6% and 1.6% from APA

Table 1: Results for the simulations with TBFF. (SI - Slack Improvement, AO - Area Overhead, PO - Power Overhead)

<table>
<thead>
<tr>
<th>Circuit</th>
<th>SI</th>
<th>AO</th>
<th>PO</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVCHAIN</td>
<td>24.5%</td>
<td>15.8%</td>
<td>12%</td>
</tr>
<tr>
<td>COMPRESSOR</td>
<td>5.6%</td>
<td>4.7%</td>
<td>3.9%</td>
</tr>
</tbody>
</table>

Table 2: Results for the simulations with APA

<table>
<thead>
<tr>
<th>Circuit</th>
<th>SI</th>
<th>AO</th>
<th>PO</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVCHAIN</td>
<td>12.2%</td>
<td>54%</td>
<td>74.8%</td>
</tr>
<tr>
<td>COMPRESSOR</td>
<td>4.3%</td>
<td>5.9%</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

The overheads observed in the test circuits present large variations between values due mainly to the test circuits sizes. Since the INVCHAIN test circuit is much smaller than the compressor, both techniques create greater impact on the circuits overall structure, reflecting on greater area and power values. Consequently, the circuit complexity and size also reflects on the slack improvement obtained.

It should also be noted that in the INVCHAIN APA example the application of TBFF was not possible due to lack of sufficient slack in the HCFOP. Hence, both techniques should be applied in combination.

5.2 AMBER Processor

The Amber core is a microprocessor compatible with 32-bit ARM architecture and ARMv2a instruction set. It also counts with a number of peripherals such as UART, timers and Ethernet MAC [13]. After synthesis of the processor, critical paths were extracted and balanced with its respective fan out stages, following the flow presented in Section 4. The synthesized processor contained 8168 cells from the 130nm technology library, with the CPs extracted containing around 60 logical gates from one FF to another. In order for the slack improvement to be calculated, a 20ns time period was set for the Amber clock (50MHz).

The results obtained are summarized in the Table 3. At total, the flow detected the need for 30 critical paths to be balanced in order to improve the total circuit slack. Three paths were also detected at which the following stage was too constrained for the balancing to cause any major improvement, so the APA with APARP topology was implemented, resulting in greater robustness of this paths at the cost of an area overhead of 1.6%. The power overhead created by the two techniques amounted for 2.7 % in total.

6 CONCLUSIONS

This work presents two techniques that mitigate aging induced timing violations in integrated circuits. The first is a time borrowing based technique that applies a flip-flop (TBFF) with additional delay cells (buffers) to delay the data sampling of critical paths. The second approach is the Alternative Path Activation (APA), which is employed when a critical path is succeeded by equally constrained stages. These techniques were applied for simple test circuits indicating its feasibility. Further, the approach was implemented for an ARMv2a processor, leading to an acceptable aging induced delay tolerance increase of 7.2 % at an area overhead of 2.5 % and power overhead of 2.7 %.

7 ACKNOWLEDGMENTS

The authors would like to thank PPGEE/UFMG, CNPq and FAPE-MG for financial support.

REFERENCES


