Article

Margin Elimination in a 55 nm Near-Threshold Microcontroller with Adaptive Prediction Capability and Voltage Scaling

Runze Yu 1, Zhenhao Li 1, Xi Deng 1, Zhaoxu Wang 1, Haoming Zhang 2 and Zhenglin Liu 1,*

1 School of Integrated Circuits, Huazhong University of Science and Technology, Wuhan 430000, China; runze_yu@hust.edu.cn (R.Y.); zhenhao@hust.edu.cn (Z.L.); d202180903@hust.edu.cn (X.D.); d202281043@hust.edu.cn (Z.W.)
2 Wuhan Top-AI Semiconductor Co., Ltd., Wuhan 430000, China; zhang_haoming@top-ai.cn
* Correspondence: liuzhenglin@hust.edu.cn

Abstract: This paper presents an innovative approach for error prediction (EP) tailored to near-threshold operations, addressing the energy-efficient requirements of digital circuits in applications such as IoT devices and wearables. The novel EP technique combines the benefits of error prediction and detection, effectively addressing critical issues associated with each method by enabling adaptive prediction capability and voltage scaling. More specifically, the presented EP method requires no modifications to the processor pipeline and mitigates the generation of false-positive errors, ensuring stable operation of the system at high-efficiency points. The effectiveness of this strategy is demonstrated through its implementation in a near-threshold 32-bit microprocessor system with a modest 5.82% area overhead. Silicon measurements validate the adaptive EP system from 0.59 to 0.66 V (4–32 MHz) and confirm its removal of all voltage margins. Here, the EP technique reduces the energy consumption by 18.6–25.1% with respect to the signoff margins and it allows the system to operate without energy overhead compared to its ideal non-margined critical operation point, with less than a 5% throughput loss.

Keywords: near-threshold computing (NTC); adaptive voltage scaling; Error Detection and Correction (EDaC); energy-efficient; margin elimination

1. Introduction

The demand for energy-efficient digital circuits has surged, driven by burgeoning applications such as Internet of Things (IoT) devices and wearables, prompting the emergence of near-threshold computing (NTC) [1,2]. Despite its merits, NTC brings about considerable variations in path delay due to process, voltage, and temperature fluctuations and the resulting design margins [3]. Although these margins ensure reliable low-voltage operation, they create a formidable energy overhead that overthrows part of the energy savings obtained by NTC.

Further energy savings are possible by reducing the design margins. The simplest recourse for margin alleviation entails the use of a replica delay line for on-chip performance monitoring [4]. Nevertheless, replica systems are inadequate in negating margins for local variations, such as intra-die discrepancies, local resistive (IR) drops, and localized temperature hotspots that elude capture.

To overcome all forms of variations, Error Detection and Correction (EDaC) systems employ in situ monitoring on critical paths. They typically identify timing errors by monitoring the data transitions of sequential elements within a specified Detection Window (DW) and correct occurring timing errors. The work presented in [5] introduces an EDaC approach for ultra-low-voltage microprocessors, employing error detecting latches (EDLs) to detect timing errors through monitoring virtual node potentials during the high clock phase. This approach implements a non-stall, single-cycle error correction method by boosting the supply voltage. However, it encounters limitations, including reduced...
per-stage detection capability and an 8.3% increase in area overhead, attributable to the integrated EDAC strategy. Ref. [6] uses the same error detection method as [5], utilizing EDLs to monitor virtual node potentials during the high clock phase. For error correction, a body swapping approach dynamically adjusts the transistor body bias, enabling efficient single-cycle error rectification. Nevertheless, the methodologies described in [5,6] face limitations, including reduced detection capabilities per stage and additional area overhead, stemming from hold time constraints imposed by an excessively wide DW. The approach in [7] leverages innovative current-based detection within latches, using a minimal increase in transistor count to monitor computational accuracy and efficiently identify timing errors through variations in current flow. Error correction is achieved by a one-cycle clock gate at the root node of the clock. However, this method’s reliance on current-based detection may limit its effectiveness under rapid dynamic voltage fluctuations, representing a trade-off between efficiency gains and robustness against environmental variations. Ref. [8] details a sophisticated mechanism for error detection, equipping flip-flops with the capability to detect timing errors in situ during the high clock phase. A unique body-swapping-based correction technique is proposed to bypass the need for instruction replay or pipeline stalling. Ref. [9] employs soft-edge flip-flops combined with in-latch transition detection and set-dominant error latches, precisely identifying timing errors immediately during the high clock phase. Error correction is achieved by a one-cycle clock gate at the root node of the clock. However, this method’s reliance on current-based detection may limit its effectiveness under rapid dynamic voltage fluctuations, representing a trade-off between efficiency gains and robustness against environmental variations.

Therefore, despite in situ monitoring of EDaC systems’ ability to eliminate margins, these approaches introduce additional overhead: (1) the energy and area overhead introduced by the wide DW (indicating strong error-aware capacity) in terms of hold buffers and the implementation of the error detection circuits and (2) the energy and performance overhead introduced by extra error correction methods.

To mitigate these overheads, error prediction (EP) systems are introduced by monitoring the activities of critical combinational cells [10–12]. They anticipate impending timing errors before the rising edge of the root clock (clk_root) and correct them through a one-cycle clock gate. Therefore, EP systems do not require architectural-level correction mechanisms, but ensuring the accuracy of prediction is a critical issue. The work in [10] detects transitions occurring in the second half of the clock cycle halfway through the data path. As such, it predicts the onset of an error, gating the next clock cycle to prevent the error from occurring. However, accurately pinpointing half-points in the data path is difficult, necessitating timing margins similar to [4] to prevent actual errors. These margins can become substantially large due to complex variations in the near-threshold region. Therefore, a prediction method [11] based on completion detection has been proposed to combine high error-aware capacity with minimal area and energy overhead, requiring no extensive alterations to the existing processor architecture. However, the method discussed in the study [11] overlooked a critical issue: the clock latency between the root node and the critical endpoints. This oversight presents challenges in selecting the monitored cells. Moreover, the severe delay variations at low voltages introduce frequent false errors disrupting the accuracy of prediction systems, introducing significant energy overhead and speed degradation.

In order to overcome the drawbacks in the existing EDaC and EP techniques, this work aims to combine the benefits of predicting errors before the clock edge (i.e., no hold constraints, high error-aware capacity, and one cycle correction) and detecting errors after the clock edge (i.e., no margins) into a novel EP concept. The concept, initially introduced in our prior work [13], involves the incorporation of transition detection (TD) cells deeply inserted in the critical paths. These cells are designed to convert critical data transitions within an adjustable prediction window (PW) just prior to the root clock’s rising edge into prediction timing error signals. These signals are then utilized to prevent potential timing errors through a cycle of clock gating. The width of the prediction window is modulated by
lightweight error detection circuits integrated at critical endpoints, ensuring the equivalence of predicted timing errors to actual timing errors and mitigating the occurrence of frequent false errors. This strategy has only been validated through post-simulation results across various corners. Upon further analysis, we identify the following primary problems with the strategy in terms of its implementation and validation approach: (1) The method for adjusting the global PW width is overly simplistic and introduces additional false errors. For instance, a wider PW triggered by the critical paths may identify transitions in monitored cells within less critical paths as prediction timing errors, leading to frequent false errors. (2) The adjustment of the PW width relies on runtime error detection outcomes, necessitating unavoidable processor architecture-level error correction. This increases design complexity and introduces speed reductions. (3) The integration of error detection, correction, and prediction mechanisms imposes a significant area overhead of 5.82% on the processor. (4) The post-simulation results under various corners only partially reflect the feasibility of the method. The impact of variations within actual chips cannot be fully validated, such as temperature, voltage drop, and aging.

In this work, a more granularly adjustable and cost-effective error prediction strategy is proposed. The adaptive prediction capability is achieved through an adaptive scaling circuit without architecture-level error correction, mitigating false errors and ensuring prediction accuracy while minimizing the introduced energy, area and correction overheads. Furthermore, considering that environmental changes may cause frequent corrections, away from the efficiency operating point [14], an adaptive voltage scaling circuit based on error rates is introduced into the system. These two adaptive scaling circuits work together to ensure a stable chip operation at an ultra-low-margin high-efficiency working point.

Implementing the improved EP strategy targets an ultra-low-power cortex M0 processor with 5.82% area overhead, operating within a speed range of 4 to 32 MHz and a voltage range from 0.59 to 0.66 V. The chips’ performance across various conditions is analyzed to verify adaptive core voltage and prediction capability scaling. The measurement results illustrate that chips operate stably with no energy overhead and minimal performance degradation, contingent on the preset error rate. For instance, when executing the CoreMark benchmark at 16 MHz and 25 °C, a standard TT corner chip with a preset 5% error rate exhibited a 24% voltage margin reduction and a 43.8% energy consumption decrease compared to the traditional worst-case signoff.

The remainder of this work is structured as follows. Section 2 explains the presented EP strategy with adaptive scaling circuits, while Section 3 translates it into a near-threshold implementation of the Cortex-M0 system. Finally, Section 4 analyzes the measurement results, Section 5 compares this work with previous works, and Section 6 concludes this article.

2. Presented Adaptive Error Prediction Strategy

The presented adaptive error prediction strategy includes the adaptive scaling circuits, enabling adaptive voltage and prediction capability scaling to ensure stable operation at an ultra-low-margin, high-efficiency point.

2.1. Concept

As shown in Figure 1, error prediction through activity monitoring relies on the following two operations. First, a transition detection (TD) cell, highlighted in yellow, consisting of a delay chain and an XOR-gate, detects the toggling activity of the monitored cell by outputting a signal with pulse width \( T_{H} \). Next, the Dynamic-OR (DYNOR) TREE (highlighted in blue), consisting of multi-level DYNOR cells, evaluates all TD pulses within the PW (highlighted in green) at the end of the root clock cycle and reduces them to the final prediction signal (p_error). These DYNOR cells employ dynamic CMOS structures, precharging intermediate nodes during the low PW period and evaluating them based on the TD pulses during the high PW period. The timing diagram in Figure 1 clearly illustrates how the asserted p_error signal is generated and enables error correction through a one-cycle clock gate by the Internal-Clock-Gate (ICG) cell. The simple error correction
method avoids extensive modifications to the processor architecture and allows a lower voltage operation.

Prediction control signal: $TD_{en}[\_\_\_]$

**Figure 1.** Illustration of the presented EP technique. Activity at TD node shows how an error is predicted and corrected.

However, the prediction strategy’s accuracy is notably sensitive to variations, leading to considerable design margins, particularly at near-threshold voltages. To tackle this challenge, a critical data path is positioned on a timeline, as illustrated in Figure 2, for a quantitative margin analysis. Firstly, a zero-margin condition for the data path is established as the point where final activity barely fails to meet the destination endpoint’s setup time. Further, when the path meets the zero-margin condition, the time difference between the output pulse of the rightmost TD cell (called basic TD cell) and the rising edge of the PW can be defined as the introduced margin:

$$T_{margin} = T_{latency} + T_H + W_{PW} - T_{interval} - T_{setup}. \quad (1)$$

where $T_{latency}$ represents the clock network delay, $W_{PW}$ is the width of the PW, $T_{interval}$ signifies the time difference between the activity of the basic TD cell and the destination endpoint, and $T_{setup}$ corresponds to the required setup time.

In Figure 2, the gray shaded area represents negative $T_{margin}$ values, signifying that the TD pulse cannot transition into an asserted $p_error$ signal. This situation could lead to uncorrected timing errors under the zero-margin condition, potentially resulting in system failures. Therefore, maintaining a small positive margin under all conditions is crucial. This can be achieved by selecting the appropriate basic TD cell. However, static timing analysis results for critical paths under all conditions reveal a range for basic TD cell selection (highlighted in blue). The rightmost basic TD cell is consistently selected to provide accurate predictions with significant margins across all conditions.
To address this challenge, the adaptive error prediction is enhanced by selectively activating the appropriate basic TD cells for different paths in a chip. Initially, it is necessary to identify a selectable range of basic TD cells for all paths that require monitoring, based on the outcomes of static timing analysis. These critical paths are then organized into N groups according to their levels of criticality for more precise control. It is advantageous to finely control the prediction capability of each path group on their criticality during actual chip operation. Since critical paths do not continuously toggle, this finely controlling method is more beneficial for reducing the margins retained by prediction strategies.

The control D Flip-Flop's (TD_en) width for each group is set based on the maximum number of basic TD cells within that group, allowing for uniform control over the enablement of basic TD cells within each group. For illustration, refer to Figure 3, which displays group 1 as an example. This group encompasses three pathways of comparable criticality. Within a delineated dashed box, each TD cell is governed by a precise bit in TD_en[1][3:1], regulating its capacity to navigate through a DYNOR gate and generate the eventual p_error signal.

Figure 2. Analysis of prediction design margins on a critical path timeline.

Figure 3. Illustration of uniform control over the enablement of basic TD cells within each group.
This methodology effectively prevents the occurrence of false errors by ensuring that only relevant TD cells are activated based on the criticality and specific requirements of each path. It demonstrates a sophisticated use of static timing analysis to dynamically adjust error prediction mechanisms, thus improving the reliability and efficiency of chip operations.

2.2. Adaptive Scaling Circuits

The adaptive scaling process and signal interactions within the system are depicted in Figure 4 to demonstrate false-error-free operations without the need for architecture-level error correction. The flowchart’s left section outlines the power-on prediction capability and voltage scaling process. After the initial power-on phase, the processor initiates a predefined critical paths reversal loop. The voltage scaling circuit steadily decreases the voltage at regular intervals (labeled as the timing signal in Figure 4), as controlled by a timer, until an asserted detected timing error (d_error) signal is triggered. The d_error signal’s assertion signifies that critical paths have experienced a recent timing error, complying with the zero-margin condition outlined in previous subsection. It can be generated by a lightweight error detection circuit, akin to traditional EDaC systems in [5,9,15], monitoring the activities of timing elements within the DW after the clock’s rising edge. It is worth noting that our system’s error awareness is solely determined by the prediction strategy, eliminating the need for a wide DW and the associated hold buffers overhead. The high-level d_error signal’s origin from a specific group of paths is identifiable. Further elaborating, the scaling circuit methodically increases the corresponding TD_en signal for the specific group from 0 to n – 1 bits. This incremental adjustment persists until a confirmed p_error signal indicates a minor positive margin within the paths belonging to this group. Until the reversal loop for all critical paths is executed to completion, the initial scaling phase concludes, allowing the processor to resume execution.

![Flowchart and signal interactions of adaptive scaling circuits.](image-url)

Figure 4. Flowchart and signal interactions of adaptive scaling circuits.

The chip’s performance is further analyzed as environmental conditions change from worse to better. In harsher conditions, timing in data paths will become tighter, and successive TD cell insertions ensure prediction accuracy and system stability. In better conditions, the equivalence between predicted and actual errors typically persists due to the same influence trends. However, delay variations across individual combinational cells in the data path may lead to an increase in margin between predicted and actual errors, resulting
in false-positive prediction errors. In both cases, this would result in frequent clock gate events, leading to significant performance losses and deviations from the energy-efficient point. In Figure 4, the error rate monitor circuit counts p_error occurrences over a specific period and triggers in-run prediction capability and voltage re-scaling after overflow events.

On the right side of the flowchart, the processor initially responds to the interrupt by entering the critical paths reversal loop. Then, periodic d_error signal monitoring helps identify the underlying causes of excessive error rates. A low-level d_error indicates over-pessimistic predictions, causing frequent undesired clock gate events and increased margins. In this scenario, the corresponding group’s TD_en signal is decreased until a low-level p_error is observed. A high-level d_error signal denotes substantial path delay degradations, leading to significantly reduced throughput. The scaling circuits then orchestrate voltage increments to complete the prediction capability scaling. With this, the re-scaling phase concludes, allowing the processor to resume execution.

2.3. TD Cells Redundancy Method

The alternative method leverages the redundancy stemming from the overlapping outputs of TDs that monitor adjacent cells along a path. This overlapping effect is due to the requirement for TDs to maintain a high signal throughout the propagation delay of the path’s slowest logic cell, ensuring that no internal activity remains undetected. Given that many cells function more rapidly than the slowest one, the outputs of TDs along a path tend to overlap. A TD becomes expendable and can be removed if its entire high signal phase is encompassed by overlaps from adjacent TDs on the path. This redundancy approach can further reduce the area overhead introduced by the strategy without compromising the error aware capability effectively.

3. Implementation Details

As shown in Figure 5, the proposed EP strategy is implemented in a near-threshold 32-bit microprocessor system. It consists of a CORTEX-M0 core, error detection and prediction circuit, memory, error rate monitor circuit, adaptive scaling circuits, clock generate module, regulated Low Dropout Regulator (LDO), and other modules. The proposed error rate monitor and adaptive scaling modules are synthesized along with the processor as conventional digital logic, while the error detection and prediction circuits are inserted post-synthesis using Engineering Change Order (ECO) commands. The diagram illustrates that specific digital modules, including the processor, operate in a speed range of 4 to 32 MHz, with voltage sweeping in near-threshold region in 30 mV increments by the LDO, while the rest maintain a standard voltage. To facilitate the ultra-low-voltage implementation, the operation of all standard cells is initially verified at the concerned voltage. Cells with functional errors or extremely large delays are excluded, and retained cells are recharacterized at these voltages to obtain new timing libraries. Subsequently, an initial low-power synthesis and place-and-route (P&R) flow transforms the microprocessor’s RTL into silicon. The error prediction strategy is then integrated into the system using ECO commands after static statistical timing analysis. However, there might a potential impact on the timing of the original paths due to the added load, multiple iterations (four in this paper) are performed until setup and hold timing requirements are satisfied. Finally, chip post-simulation and sign-off verification are conducted.

In our implementation, when integrating EP strategies, the focus lies on selecting appropriate monitored paths to ensure sufficient error detection capability while minimizing area overhead as much as possible. There is the fact that the maximum clock frequency is determined by critical paths. Thus, only data paths to the most critical endpoints need to be monitored allowing a limited overhead. The amount of critical endpoints is determined by the chance of false-positive monitoring, e.g., the chance that non-monitored path propagates slower than all monitored paths owing to various variations, as described in Equation (2).

\[
    P_{false} = P(\exists p_i : T_{prop,p_i} > \max(T_{prop,q_j}))
\]

(2)
where \( p_i \) represents all non-monitored paths and \( q_i \) represents all monitored paths. When the false monitoring occurs, there is a probability that non-monitored paths may fail while monitored paths do not, thus causing a system operation failure. To avoid this occurrence, enough timing slack (12.2% clock period in this paper) is covered with 343 out of 4025 endpoints being monitored. The probability of such an event is determined by the delay distribution of a subset of paths from 1000 Monte Carlo (MC) simulations at 600 mV, which is less than \( 1 \times 10^{-15} \) in this paper, decreasing with increased voltage. Hence, 343 out of 4025 endpoints were monitored by error detection circuits from Monte Carlo simulations and grouped into 47 sets according to their similar criticality for more precise control over prediction capabilities. Then, 225 monitored cells were inserted into monitored paths for error prediction with sufficient error aware capability and thus forming a correct error propagation path. Furthermore, all basic TD cells are connected to control registers of the adaptive scaling module. Then, all TD cells covered by the detection range of adjacent TD cells are eliminated according to the static timing analysis. The strategy integration incurred a 5.82% increase in chip area, as depicted in Figure 6.

![Figure 5.](image1.png)

**Figure 5.** The 32-bit microprocessor system with EP strategy overview. The different colored dashed boxes represent the different voltage domains.

![Figure 6.](image2.png)

**Figure 6.** The components of EP integration overhead.

### 4. Silicon Measurements

For improved testability, the processor provides a configuration option as a baseline design with the EP strategy turned off within the same die. Figure 7a presents the overall test platform, where the chip operates inside a temperature chamber. The results are sent to a PC via Universal Asynchronous Transmitter (UART), and an energy monitoring board powers the chip. The regulated LDO’s low-voltage output to the core is monitored via a pad-connected oscilloscope. Figure 7b illustrates our chip’s micrograph.
The tests are conducted by running the CoreMark benchmark at various temperatures and record the current drawn by the energy monitoring board, the results printed by the PC, and core voltage fluctuations monitored by an oscilloscope. Figure 8 depicts voltage changes in a standard TT corner chip at room temperature (25 °C) and 32 MHz. The core voltage gradually decreases to 0.66 V following the initial power-on reset in phase II, ultimately stabilizing for CoreMark execution. Subsequently, the temperature of the chamber is gradually changed with different preset error rates. Test results show that the core voltage stays constant at 0.66 V as the temperature increases, owing to the gradually relaxing timing. Also, the core voltage stays constant when an over 5% error rate is set but introduces substantial performance degradation due to the exponential relationship between transistor current and voltage in the near-threshold region. Then, with a 5% error rate setting, when the temperature of the chamber is gradually lowered to −5 °C, the voltage increases to 0.7 V in phase IV, ensuring stable operation and exemplifying the efficacy of the adaptive voltage scaling circuit.

To demonstrate our system’s margin reduction, the chip runs CoreMark at frequencies from 4 to 32 MHz under three voltage scaling conditions:
1. Critical voltage scaling ($V_{\text{critical}}$), representing non-margined critical operation just above the error threshold.
2. Signoff voltage scaling ($V_{\text{signoff}}$), representing traditional baseline signoff voltage at SS corner, −40 °C, 10% voltage margin.
3. EP system voltage scaling at a 5% error rate ($V_{\text{EP-5\%}}$), representing our system’s stable operating voltage at 5% error rate.
In Figure 9, the yellow shaded area indicates a 24% reduction in voltage margin compared to signoff conditions at 16 MHz, which aligns with the 43.8% reduction in energy consumption shown in the green area. And the area between \( (V_{EP} - 5\%) \) and \( (V_{critical}) \) (depicted as the shaded slope region) represents the margin reduction compared to zero-margin operation, achieved by running at a 5% error rate. Then, it can be concluded that there is no additional energy margin compared to critical zero-margin operation, with less than a 5% throughput loss.

![Figure 9. Voltage and energy scaling over frequency at three operation conditions.](image)

Furthermore, five representative chips of different types are selected and their stable core voltage at \(-30\, ^\circ\text{C}, 25\, ^\circ\text{C}, \) and \(70\, ^\circ\text{C}\) are presented in Figure 10. Irrespective of chip variations, our system effectively adapts voltage and prediction capability to ensure stable operation. Notably, decreasing temperatures coincide with an increase in core voltage. The red arrows in the Figure 10 signify the margin reduction relative to the signoff conditions.

![Figure 10. The stable core voltage of five representative samples at different temps.](image)

5. Comparison with Previous Works

An extensive comparative analysis is conducted with other research works, and the results are summarized in Table 1. The studies in [9,16] employed traditional EDaC strategies, which involve substantial area overhead to ensure robustness. In [9], the approach enabled time borrowing to avoid architectural-level timing error corrections, but it also increased...
design complexity and required significant margins for stable operation in multi-level cascading. In [16], timing errors were corrected through instruction replay, introducing additional design complexity and resulting in a speed decrease.

Table 1. Summary and comparison with existing EDAc systems.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection method</td>
<td>Time borrow</td>
<td>Time borrow</td>
<td>Half-path prediction</td>
<td>Completion detection</td>
<td>Adaptive prediction</td>
</tr>
<tr>
<td>Correction method</td>
<td>Margin remained</td>
<td>Instruction replay</td>
<td>Margin remained</td>
<td>Clock gate</td>
<td>Clock gate</td>
</tr>
<tr>
<td>Failures 1</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Area overhead</td>
<td>76.92%</td>
<td>42%</td>
<td>10%</td>
<td>4.9%</td>
<td>5.82%</td>
</tr>
<tr>
<td>Technology</td>
<td>40 nm</td>
<td>28 nm</td>
<td>40 nm</td>
<td>28 nm</td>
<td>55 nm</td>
</tr>
<tr>
<td>Near-Vth @voltage 2</td>
<td>YES @0.29 V</td>
<td>YES @0.55 V</td>
<td>YES @0.44 V</td>
<td>YES @0.25 V</td>
<td>YES @0.59 V</td>
</tr>
<tr>
<td>Energy savings 3</td>
<td>75%</td>
<td>44.8%</td>
<td>50.5%</td>
<td>33%</td>
<td>43.8%</td>
</tr>
</tbody>
</table>

1 A YES indicates there are potential system faults at low voltage. 2 The lowest stable operating voltage achieved by lowering the voltage until timing errors become uncorrectable at 25 °C for TT corner chips. 3 Compared with the energy consumption at the signoff baseline frequency and voltage.

In contrast, refs. [10,11] implemented prediction strategies with minimal overhead but the potential for false errors at near-threshold voltage due to significant prediction mismatches. Our approach combines the advantages of both methods while mitigating their drawbacks. Specifically, our method introduces a 5.82% area overhead to enable adaptive prediction capability, eliminating the occurrence of false errors. Error correction is achieved through a straightforward clock gate mechanism, avoiding architectural-level alterations. Simultaneously, it reduces energy consumption by 43.8% at 16 MHz compared to signoff conditions and completely eliminates margins when compared to critical zero-margin points.

This comprehensive approach effectively balances the trade-offs between error prediction, correction, and energy efficiency, resulting in a highly efficient and reliable system.

6. Conclusions

In conclusion, our research introduces an innovative approach to enhance energy efficiency in near-/sub-threshold computing. The EP strategy, with its adaptive voltage and prediction capability scaling, effectively addresses the challenges in most near-threshold systems. The EP strategy leverages one-cycle clock gating to correct timing errors, enabling the system to function within a predetermined error tolerance. This approach permits the reduction of operating voltage without substantial performance degradation, thereby enhancing energy efficiency. The adaptive nature of the prediction capability in the EP strategy guarantees precise predictions across diverse conditions, thereby mitigating the occurrence of false positive errors. Its integration into a 32-bit microprocessor system demonstrates its efficiency in eliminating margins with minimal overhead. The significant energy savings from silicon measurements affirm its effectiveness in reducing consumption while incurring only minimal throughput loss.

Author Contributions: Conceptualization, R.Y. and Z.L. (Zhenglin Liu); methodology, R.Y. and X.D.; software, R.Y. and Z.W.; validation, R.Y., X.D. and Z.L. (Zhenhao Li); formal analysis, R.Y. and Z.W.; investigation, Z.L. (Zhenglin Liu); resources, Z.L. (Zhenglin Liu) and H.Z.; writing—original draft preparation, R.Y.; writing—review and editing, R.Y.; supervision, R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China (Grant No. 62274068 and No. 62202178).
Data Availability Statement: All data underlying the results are available as part of the article and no additional source data are required.

Conflicts of Interest: Haoming Zhang is employed by the Wuhan Top-AI Semiconductor Co., Ltd. The authors declare no direct or indirect commercial or financial relationships that could be construed as a potential conflict of interest regarding the research, authorship, and publication of this article. The processor utilized in this study was accessed through a research-focused licensing agreement, and all intellectual contributions and rights to the methodologies described are retained by the listed authors.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.