On the Efficiency of Voltage Overscaling under Temperature and Aging Effects

Hussam Amrouch, *Member, IEEE,* Seyed Borna Ehsani, Andreas Gerstlauer, *Senior Member, IEEE,* Jörg Henkel, *Fellow, IEEE*

Abstract—Voltage overscaling has received extensive attention in the last decade as an attractive paradigm for systems in which resulting timing errors and thus a loss in accuracy can be accepted in exchange for an increase in energy efficiency. At the same time, the delay of a circuit is, in turn, and in addition to voltage, also subject to temperature and aging. Existing work has largely studied voltage overscaling in isolation. This ignores interdependencies with temperature and aging, which can lead to wrong or misleading conclusions. In this work, we are the first to model the *combined impact* of voltage, temperature and aging on the delay of circuits towards investigating the actual existing trade-offs between efficiency and accuracy provided by voltage overscaling. We show that analyzing voltage in isolation overestimates timing errors and thus underestimates the voltage scaling potential. We further develop an approach that leverages interdependencies to optimize energy, delay and accuracy trade-offs. We precisely translate the individual and combined impact of voltage-, temperature-, and aging-induced delay increase into corresponding probability of error (P_{error}). This reveals that the same amount of timing increase results in different error probabilities depending on the origin (i.e. voltage, temperature or aging). For the same timing increase, voltage reductions result in the smallest P_{error} compared to temperature or aging, while also reducing temperature- and aging-induced delay increases themselves. This allows voltage reduction to be employed as an effective means to minimize delay, reduce energy and thus maximize efficiency under a given upper bound on error probability. We apply our approach to multipliers in GPUs exploring the trade-off between efficiency and accuracy. We demonstrate how only accounting for voltage scaling alone leads to a considerably larger P_{error} (74% on average) than in reality. Our investigation also shows that for the same P_{error} constraint, optimizing for combined voltage, temperature and aging effects results, on average, in 116% better energy-delay product (EDP) compared to state of the art.

Index Terms—Voltage Overscaling, Approximate Computing, Efficiency, Accuracy, Guardbanding, Reliability, Aging, BTI, Temperature

1 Introduction

Voltage overscaling has been extensively researched as an effective means to increase efficiency of circuits that can tolerate errors to some degree. Voltage (V_{dd}) reduction provides quadratic savings in dynamic power along with an exponential saving in static power [8]. However, reducing V_{dd} without proportionally decreasing frequency leads to errors due to timing violations because of the unsustainable clock circuits will be operated at. With increasing severity of energy, variability and reliability concerns in current and future technologies, many better-than-worst-case design methods have been proposed to cope with such errors [6]. Furthermore, many systems, especially in the embedded domain, can inherently accept errors [7] or employ fault-tolerance mechanisms, such as redundancy or error correction, to protect against different sources of runtime faults. In all cases, systems are designed to tolerate a certain level of component errors, which voltage overscaling can trade off in exchange for significant efficiency gains.

Hussam Amrouch and Jörg Henkel are with the Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. (E-mail: amrouch, henkel@kit.edu). Seyed Borna Ehsani is with the Paul G. Allen school of computer science and engineering, University of Washington, USA. (Email: behsani@cs.washington.edu). Andreas Gerstlauer is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, USA. (E-mail: gerstl@ece.utexas.edu). This work is partially supported by a Humboldt Research Fellowship and the German Research Foundation (DFG) priority program "Dependable Embedded Systems" (SPP 1500). Authors would like thank Behnam Khaleghi from University of California, San Diego for his work at the gate-level analysis and library creation. We also thank Souvik Mahapatra from IIT Bombay and his team for the valuable support in aging modeling. Corresponding author: Hussam Amrouch

In addition to voltage, the delay of circuits during operation is also subject to different kinds of degradations. Their impact can range from millisecond scale, such as in temperature effects [1], [2], to significantly larger time scales, e.g. in aging effects [3]–[5]. Both temperature and aging will *naturally and inevitably* take place at runtime, which aggravates the occurrence of timing violations caused by voltage overscaling. At the same time, reductions in V_{dd} will lead to reductions in temperature- and aging-induced delay degradations. As such, there is a counteracting relationship between the inherent (primary) and temperature- and aging-driven (secondary) delay effects of voltage scaling.

Existing work to trade off efficiency with accuracy using voltage overscaling has traditionally studied voltage scaling in isolation. Similarly, error and delay trade-offs have been explored for different aging or temperature degradations *individually* [2], [4]. However, to the best of our knowledge, no existing work has looked at the *combined* impact of voltage, temperature and aging on timing, errors and energy efficiency when aiming for voltage overscaling. This ignores interdependencies and leads to inaccurate observations and conclusions in accurately evaluating the full potential, trade-offs, benefits and limitations of voltage overscaling.

1.1 Our Scope and Contributions

In this work, we present a novel approach to investigate voltage overscaling and timing error behavior under the joint impact of voltage, temperature and aging effects. This

first and foremost necessitates to correctly model the relation between voltage, temperature and aging and their *combined* effects on delay, energy and probability of error. Using such models, we precisely evaluate actual trade-offs and demonstrate, for the first time, a methodology to maximize energy-delay efficiency under an overall error goal while accurately considering joint effects. Our results show that studying degradations in isolation or simply using delay increase as a metric to quantify degradations do not accurately capture real behavior at the system level, where the full potential behind voltage overscaling is larger than what a traditional isolated analysis assumes.

We focus in this work on errors due to timing violations (i.e., errors caused by degradation-induced circuit delay increases due to temperature, voltage and aging effects), which are traditionally protected against by means of timing guardbands. Other error sources, such as soft errors, are orthogonal to this work and need to be protected against by other means such spatial and/or temporal redundancies. In general, they can be independently treated and analyzed, where results can be combined with our analysis to obtain overall error behavior as a function of voltage.

Our novel contributions within this paper are as follows:
(1) We link the physical level where degradation effects originate all the way up to the system level where errors finally occur to model precise energy-delay-error trade-offs under joint voltage scaling, temperature and aging effects. To achieve that, we create degradation-aware cell libraries that account for *combined* impact of voltage, temperature and aging. This allows designers to accurately obtain timing behavior of circuits under voltage overscaling.
(2) We further introduce methods to analyze detailed circuit behavior and break down root causes of timing errors in circuits optimized for both energy and performance. We demonstrate how different degradations (e.g., temperature, aging and voltage) cause different delay increases and how they, for the same maximum delay increase, result in different delay characteristics and thus probabilities of error, where voltage degradations contribute larger delays but have fewer impact on errors than temperature or aging ones. This confirms complex existing relationships between different degradations, delays and errors that necessitate accurate modeling and a holistic analysis.
(3) We propose a design methodology that builds on our models and observations to trade off accuracy with both energy and delay efficiency. For a certain error constraint, it determines the minimum sustainable V_{dd} under the joint impact of voltage, temperature and aging. Our approach leads to a significant efficiency increase by simultaneously a) minimizing temperature and aging guardbands and b) maximizing energy savings.

1.2 Preliminaries and Background

V_{dd} reductions provide quadratic and exponential savings in dynamic and static power, respectively [8], as follows:

$$P_{dynamic} = \alpha C V_{dd}^2 f \quad (1)$$

$$P_{static} = V_{dd} \times I_{off} \; ; I_{off} \approx e^{(V_{dd}-V_{th})} \quad (2)$$

Here, α is the activity factor, C is the load capacitance, f is the operation frequency, and I_{off} is the leakage current.

At the same time, reducing V_{dd} directly increases the delay of a transistor, gate or circuit (see the linear relation in Eq. 3) [8]. In addition, it also reduces the transistor drain current (I_{on}), which, in turn, leads to an additional increase in the gate delay (see the quadratic dependency of I_{on} on V_{dd} in Eq. 4) [8]:

$$\tau_d = \frac{C \cdot V_{dd}}{4}\left(\frac{1}{I_{onN}} + \frac{1}{I_{onP}}\right) \quad (3)$$

$$I_{on} \approx \frac{\mu}{2} \cdot (V_{dd} - V_{th})^2 \quad (4)$$

When the gate delays become larger, the overall critical path delay of a circuit (t_{cp}) enlarges as well. Hence, unless the clock period is adjusted accordingly, errors caused by timing violations will appear due to the unsustainable frequency:

$$f = \frac{1}{t_{cp}} \; ; t_{cp} = \sum_{i \in CP} \tau_d(i) \quad (5)$$

$$V_{dd} \searrow \Rightarrow \tau_d(i) \nearrow ; V_{dd} \searrow \Rightarrow I_{on} \searrow \Rightarrow \tau_d(i) \nearrow$$

$$\tau_{d(i)} \nearrow \Rightarrow t_{cp} \nearrow \Rightarrow \text{Errors due to timing violations!}$$

Here, τ_d (i) is the delay of gates that contribute to the critical path (CP). I_{onN} and I_{onP} are the drain currents of nMOS and pMOS transistors in the "on" state (i.e. saturation region). μ and V_{th} are the transistor carrier mobility and threshold voltage, respectively.

Temperature and aging effects: Both short- (i.e., temperature) and long- term (i.e., aging) reliability degradations share similar characteristics. They both originate from physical effects and then propagate all the way up to the system level, where they ultimately cause errors due to timing violations. Temperature and aging alter the key parameters of MOSFET transistors, such as V_{th} and μ. This, in turn, reduces the drain current (I_d) when the transistor is in the "on" state (see Eq. 4). Hence, aged transistors or transistors under high temperatures become slower. Thus, the delay of standard cells increases. As a result, the delay of critical paths in a circuit enlarges, leading to timing violations because the clock frequency becomes unsustainable (see Eq. 3 and 5). To compensate temperature- and aging- induced delay increases, a timing guardband needs to be added on top of the maximum delay of a circuit under nominal conditions [4]. A guardband ensures that timing is always met under all conditions [4]. Such timing guardbands consist of multiple components: a guardband to cope with temperature effects (T_{gb}) [2], estimated at the maximum temperature that can be reached during the operation of circuit (e.g., 125°C), and a guardband to cope with aging effects (A_{gb}) [4], [5], estimated at the end of the projected lifetime (e.g., 10 years). However, including timing guardbands directly leads to efficiency losses requiring circuits to be clocked at a lower frequency to ensure reliability:

$$f_{gb} = \frac{1}{t_{cp} + T_{gb} + A_{gb}} \Rightarrow f_{gb} < f \Rightarrow \text{efficiency loss!} \quad (6)$$

In our work, we target aging mechanisms that alter the electrical properties of transistors (e.g., threshold voltage, carrier mobility, etc.), which lead to delay increases. In our experiments, we demonstrate our approach using Bias Temperature Instability (BTI) because it is the dominant aging

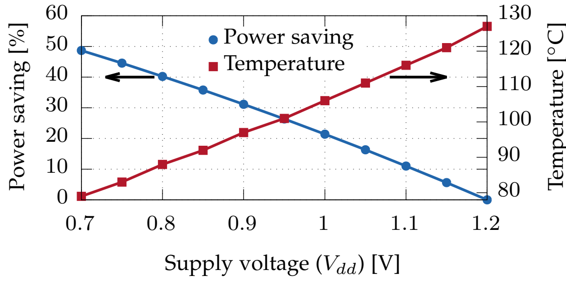


Fig. 1. Impact of voltage reductions on reducing the overall power and thus temperature. Beyond 0.7V, further reductions in power and temperature occur, but the induced delay increases become significant.

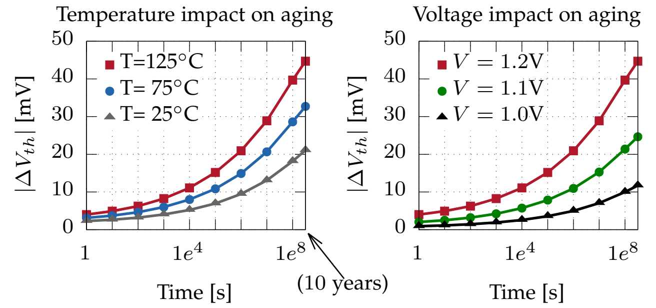


Fig. 2. Impact of reducing temperature and voltage on mitigating the aging-induced degradation, quantified by threshold voltage increase (ΔV_{th}) in 45nm pMOS transistors. Note that the x-axis has a log scale.

mechanism in planar and FinFET technologies [3]. However, our approach is not limited to BTI, and other aging phenomena that also impact the delay of circuits, such as Hot Carrier Injection (HCI), can be straightforwardly included as will be explained in Section 3.2. Aging mechanisms that do not influence transistor delay itself but instead lead to permanent errors, e.g., Time-Dependent Dielectric Breakdown (TDDB) and/or electromigration in interconnections, are orthogonal to our work and, similar to soft errors, can be treated separately and then combined with our analysis methodology to obtain overall error probabilities.

Interdependencies between temperature, aging and voltage: Reductions in V_{dd} also affect circuit delays and are equally protected against using corresponding guardbands in traditional designs. In addition, however, V_{dd} also directly affects temperature and aging. As explained earlier, reducing V_{dd} leads a considerable saving in power, which, in turn, results in lower operating temperatures due to reductions in on-chip power densities. In Fig. 1, we present an example of the potential saving in power when V_{dd} is scaled down from 1.2V to 0.7V along with the resulting reduction in temperature obtained from our experimental setup (details in Section 6.1).

Lower temperatures result in increased carrier mobilities μ [1], which reduces circuit delays. Finally, the mechanisms underlying defect generation in aging also have a very strong dependency on both voltage and temperature [3], [9]. A reduction in V_{dd} strongly decelerates aging not only due to the smaller electric fields akin to the lower V_{dd} but also due to the lower temperature. Using a physics-based aging model [3], Fig 2 demonstrates the dependency of BTI-induced degradation (i.e. ΔV_{th}) on V_{dd} and temperature T. Note that BTI degradation over time follow a power law with an approximately $1/6$ exponent [3], [58]. As shown in Fig 2, reductions in V_{dd} or T result in less aging (i.e. smaller ΔV_{th}). In theory, this allows temperature and aging effects along with their guardbands to be mitigated using voltage reduction. However, operating circuits at lower than their nominal voltage comes with a considerable delay increase and thus trade-off itself, as explained earlier.

In summary: Reducing V_{dd} increases circuit delays, which leads to timing errors when clock frequencies are not scaled in proportion. In reality, circuits are already protected against worst-case temperature- and aging-induced delay increases using timing guardbands. As a result, as long as circuits are not operated at worst temperature or age, these guardbands will also absorb some of the delay increases and hence timing errors due to voltage overscaling. In addition, scaling voltages will also reduce temperature and aging degradations, thus reducing combined delay increases and hence further mitigating timing errors. In fact, under scaled voltages, maximum temperature and aging will also be reduced, leaving extra guardbands to absorb voltage scaling delays even under worst-case conditions. Thus, in other words, voltage can be scaled further and thus energy be minimized[1] compared to an isolated analysis that does not consider temperature and aging under a given upper bound on probability of errors. Alternatively, guardbands can be narrowed to match improved worst-case conditions and thus increase performance when scaling voltage.

In all cases, voltage overscaling increases circuit delays but mitigates degradations, where combined effects at different voltage and hence energy levels depend on the exact trade-off between voltage- versus temperature- and aging-induced delays and errors. This provides the opportunity and need to investigate trade-offs between errors and energy-delay efficiency in the context of combined voltage, aging and temperature degradations.

In this work, we consider planar MOSFET technology at the 45nm node for all our quantitative results, background analysis and experiments. Cell library characterizations (details in Section 3.2) are done using HSPICE simulations for the typical-typical corner in which the combined effects of voltage, temperature and aging are modeled. Since we target performance-critical designs operated at or near their nominal base frequency, we limit characterizations to the super-threshold region above 0.7V. As such, all results presented in this paper are specific to this technology setup. However, interdependencies between voltage, temperature, aging and delay will exist in other technology nodes as well. Our proposed method for cell library characterization and joint error analysis/optimization is general and can be applied using other base technology libraries and models. Note that in other technology nodes, some trends may reverse (e.g., temperature increases might lead to larger transistor drain current [10] and thus smaller delays/faster speed in smaller nodes due to the reverse temperature dependence [1]) while others may stay the same (e.g., impact of BTI [3]). However,

1. Lowering V_{dd} reduces energy in the super-threshold region until an optimal energy point that occurs at or near the threshold voltage.

our proposed characterization, analysis and optimization flow will still apply.

The rest of this paper is organized as follows: Section 2 summarizes existing related work. Section 3 demonstrates the limitations of existing methods in quantifying degradation-induced errors and explains our novel method using degradation-aware cell libraries. Afterwards, we investigate and compare in Section 4 the resulting delays and errors under *individual* and *combined* impact of voltage, temperature and aging degradations. Then, we present in Section 5 how accuracy and efficiency can be traded off using voltage overscaling under the combined impact of voltage, temperature and aging degradations. Section 6 demonstrates our evaluation results and comparisons against state of the art. Finally, Section 7 concludes our paper.

2 Related Work

The impact of voltage reductions alone on timing errors has been extensively studied for energy improvements in both fault-tolerant and -intolerant systems [6], [7], [12]. In fault-tolerant approximate computing [7], voltage overscaling has been employed to save energy while accepting resulting timing errors, using circuit modifications [13]–[17], error correction [18]–[20] or combinations of techniques [21] to minimize the impact of timing errors. By contrast, better-than-worst-case design techniques, like Razor [6], aim to correct rather than accept errors from voltage overscaling or reduced guardbands. Recently, [22] employed voltage overscaling to trade off energy with quality in the context of coarse-grain reconfigurable architectures. They also showed that voltage reductions can separately minimize aging effects, but did not consider how aging in turn affects errors.

Unlike voltage, only few works have studied the impact of temperature- and aging-induced degradations on timing errors. To accurately model how timing errors due to aging or temperature *standalone* propagate to faults at the system level, aging- and temperature-aware cell libraries were proposed [2], [5]. Alternatively, [23] proposed a microarchitectureal stuck-at-0 fault model that avoids expensive gate-level simulations. However, such a simplified model (at the microarchitecture level) cannot accurately capture the impact of various effects of how timing errors originating from lower levels (i.e. device and gate levels) are masked and propagate. Our analysis (details in Section 3) demonstrates that the same delay increase caused by different sources of degradation (temperature, voltage, aging) can result in significantly varied error probabilities at higher levels.

As an alternative to voltage scaling, [24] and [25] recently proposed to trade off efficiency with accuracy by means of timing guardband reduction without voltage reduction (i.e. increasing the performance in exchange for a certain amount of incurred probability of error (P_{error}) by narrowing guardbands and running the design at a faster clock). Similar to voltage overscaling, such a technique provides an efficiency improvement under a specific P_{error} constraint. To avoid catastrophic errors and actually minimize efficiency losses, [24] and [25] then applied approximate circuit design principles. However, they either necessitate permanent modifications in the design functionality or they deal with aging or temperature effects standalone.

There are several techniques to minimize aging effects using voltage scaling as a knob, such as [26], [26]–[35]. The aim of such techniques is to sacrifice performance while reducing peak temperatures or decelerating aging effects. Facelift [28], for instance, adjusts the V_{dd} of CPUs in a multi-core system in such a way that aging rate becomes balanced among various cores. Such techniques do not study the combined impact of degradations. They also quantify aging by ΔV_{th} and do not translate that degradation into the corresponding circuit's delay increase or errors. While, ΔV_{th} is a proper metric to quantify aging at the transistor level, it cannot capture aging effects either at the circuit or system level. In fact, estimating accurately the required timing guardband and hence the resulting performance and efficiency losses is what matters for designers at the circuit level when quantifying degradation effects. More importantly, translating the degradation effects into the corresponding P_{error} is what designers at the system level ultimately require when trading off efficiency and accuracy.

3 Quantifying Degradation Errors

As explained, even though temperature, aging and voltage reduction impact the electrical characteristics of MOSFETs in different ways, all of them similarly manifest themselves as a degradation in the speed of gates. Thus, the delay of the critical paths of a circuit enlarges and consequently timing violations occur unless a sufficient guardband is employed. Every violated path leads to an error in one of the components of the circuit. For example, in arithmetic units (e.g., multipliers, adders, etc.), timing violations manifest themselves as errors in the performed computations. One can calculate the probability of error (P_{error}) resulting from the studied degradation (e.g., voltage, temperature, aging) using a statistical analysis. *Achieving such an analysis represents concisely our goal in this section.* Note that state of the art (e.g., [4], [5]) quantifies the impact of degradation effects solely with respect to delay increase. While such a metric properly represents the impact of the studied degradation *at the circuit level*, it cannot describe how the incurred timing violations will later cause errors *at the system level*. Instead, we translate degradations into P_{error}, which can serve as a more meaningful metric to trade off accuracy (i.e. accuracy loss) with efficiency at the system level.

To achieve that, we create standard cell libraries containing the detailed delay and power information for every sequential and combinational gate under the studied degradation effects. We then employ these libraries to perform detailed degradation-aware timing analysis and gate-level simulations of various components, such as arithmetic units, under different degradations and timing constraints to compute the resulting P_{error} using a statistical analysis under various scenarios.

3.1 Limitations of Existing Methods

State of the art typically quantifies the impact of degradation effects –due to voltage overscaling, temperature increase, or aging– on circuit delay by analyzing how the delay of few logic gates [4], [28] or the delay of a ring oscillator (RO) [36] will be increased when that particular degradation

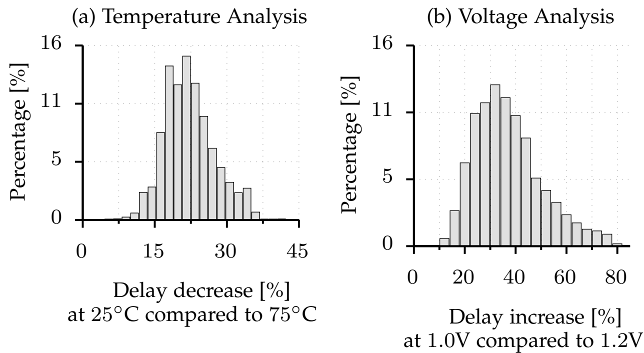


Fig. 3. Distributions of temperature- and voltage-induced gate delay increases in a 45nm standard cell library. The delay of gates can be affected differently under the same change in T or V_{dd} [37].

occurs. Based on such an abstracted analysis, the predictions are later generalized towards studying the impact of the degradation effect on the delay of the entire circuit. Other state of the art [13] extracts the critical path of a circuit (CP) and then analyzes it using SPICE simulations to accurately quantify how the studied degradation will increase the delay of that path. In all cases, existing state of the art aimed at analyzing how the delay of circuit is affected by degradation effects, always studies an individual degradation (i.e. voltage, temperature or aging) standalone and none of the existing work investigated the *combined impact* of multiple degradation effects, as is our focus.

To demonstrate why the aforementioned methods to study the impact of degradation effects are inaccurate, we analyze in the following how changes in temperature and voltage can affect the delay of standard cells as well as the delay of paths within a circuit. A similar analysis with respect to aging effects, instead of voltage and temperature, is available in our previous work [5].

Standard cells under degradation effects: Fig. 3 [37] shows the impact of temperature and voltage changes on the delays of gates within a 45nm standard cell library [38] obtained using HSPICE simulations. As shown in Fig. 3(a), the same temperature reduction of 50°C differently impacts the delay of different gates. The delay decrease can be merely 0.6% and up to 49%. Analogously, a voltage decrease of 0.2V can differently impact the delay of gates. As shown in Fig. 3(b), a variance in gate delay increases between 2.5% and 81% can be observed.

In addition, the same gate itself can be differently influenced by the same T and/or V_{dd} change. This is due to the operating conditions ($OPCs$) of a gate (i.e. output load capacitance and input signal slew of gate) determining how the delay changes of pMOS and nMOS transistors within the gate can magnify or cancel each other towards impacting the overall gate's delay. Fig. 4 demonstrates the delay decrease of a single Inverter (INV_X1) in isolation when the temperature decreases from 75°C to 25°C under 7×7 fixed input signal slews and output load capacitances (similar to what is done in any typical standard cell library, e.g. [38]). We also used HSPICE simulations to accurately measure the Inverter's delay under every OPC. As shown in Fig. 4, the delay decrease of the Inverter due to such a 50°C reduction is inconsistent and it strongly depends on the OPC. Similar observation can also be made when the

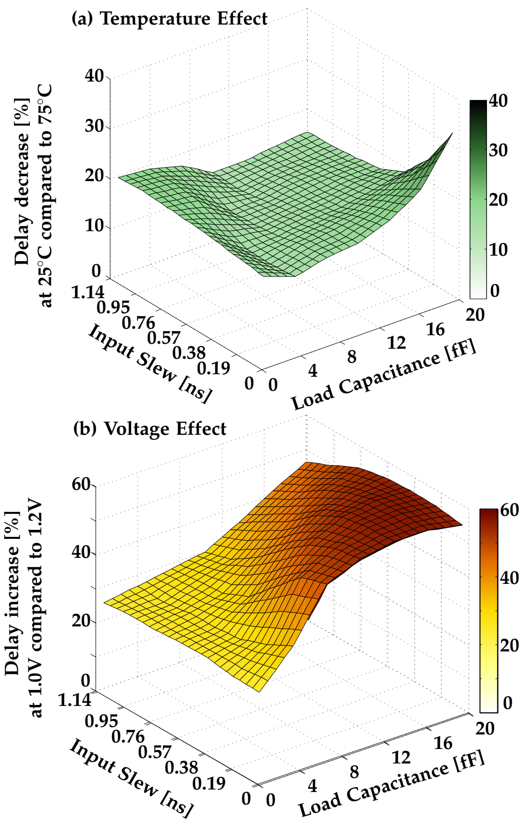


Fig. 4. Role of $OPCs$ in determining the impact of a temperature and voltage change on the delay of an *Inverter*.

voltage decreases from 1.2V to 1.0V. Note that other more complex gates (e.g., D-FF consisting of > 25 transistors) can have even a more complex OPC dependency and timing behavior. Furthermore, in practice, when embedded in a larger circuit, the OPC of a gate will itself change with temperature and voltage. A gate's OPC is determined by the also temperature- and voltage-dependent characteristics (such as capacitance) of its surrounding gates within the netlist. This makes the actual relationship between temperature, voltage and delay increases even more complex.

The above analysis demonstrates that the role of $OPCs$ cannot be neglected when modeling the effects of degradations on the delay of gates. Analyzing solely the delay of one or few gates under a single operation condition ($OPCs$), as done in state of the art, is insufficient to accurately model the impact of degradation effects on the delay of a circuit.

Impact of degradations on the critical paths of circuits: Since the same degradation can differently impact the delay of gates and even the delay of the same gate itself, the prospect that a path that formerly (i.e. before the degradation) was critical will not remain critical any more needs to be considered. We show in Fig. 5 a motivational case study, which represents the scenario of exploiting a temperature reduction to adjust and reduce voltage correspondingly. When the temperature drops from 85°C to 25°C the CP has been switched from path_1 to path_2 (see Fig. 5(a, b)). Then, when the voltage drops from 1.2V to 0.98V, the CP switches again and it becomes path_3 (see Fig. 5(c)). Note that both path_1 (before degradation) and path_3 (after T

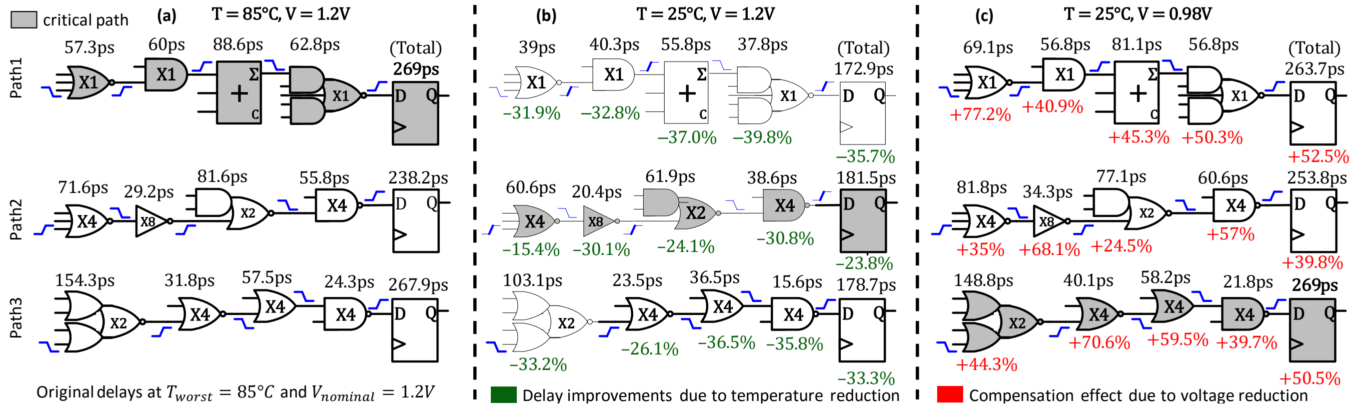


Fig. 5. Impact of temperature and voltage changes on altering the CP of a circuit. (a) shows how path$_1$ was formerly critical at $T = 85°$C. (b) shows how path$_2$ became critical when the temperature is reduced to $T = 25°$C. (c) shows how a voltage decrease from 1.2V to 0.98V compensates the gained delay improvement from the temperature reduction but makes path$_3$ now critical. Note that the presented results show a realistic example in which all gates' delay have been measured by HSPICE simulations under different (T, V) cases at the 45nm technology node.

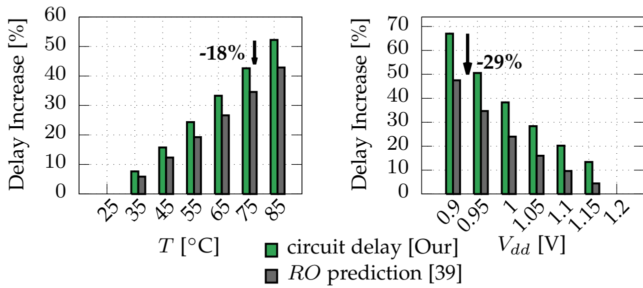


Fig. 6. Evaluating how accurate an RO-based approach in predicting the impact of T increase (relative to the baseline of 25°C) and V_{dd} increase (relative to the baseline of 1.2V).

and V_{dd} changes) have a similar delay at the end even though they have an entirely different internal structure. This demonstrates further why looking to initial CPs might be misleading when it comes to analyzing the circuit's delay under degradation effects. Concisely, considering solely the initial CP is not sufficient as potentially other paths may become critical. Furthermore, considering the top x% of CPs may not also be a practical solution because determining an x that is guaranteed to include all paths that may become later critical is not feasible.

Hence, analyzing the timing of the entire circuit's netlist (i.e. not only the original critical path or a few critical paths) is indispensable to accurately modeling the effects of degradation effects on the circuit's delay.

Accuracy of ring oscillator-based predictions: A ring oscillator (RO) typically consists of a feedback loop of an odd number of inverters to produce an oscillation frequency predicting delay increases/decreases. Degradation effects due to a change in voltage, temperature or aging directly influence the delay of inverters and thereby the number of oscillations. With careful calibrations, a delay increase/decrease can be predicted based on the changes in oscillations. However, the same temperature or voltage change can differently impact the gates' delay, as demonstrated in Fig. 3. This brings up the critical question of "*how good is the RO in predicting the impact that degradation effects actually have on the critical path of circuit?*". To investigate that, we study a typical RO of 15 stages [40]. We compare the predicted delay increase with the RO against the actual delay increase of the CP of a 32-bit multiplier. The latter has been obtained by the Synopsys Timing Analysis tool along with our created degradation-aware cell library that corresponds to the targeted temperature or voltage case. Fig. 6 presents the comparison results for both temperature and voltage scenarios.

As observed, in both scenarios, there is a noticeable error in RO-based predictions reaching 18% and 29% for the temperature and voltage scenarios, respectively. In fact, such large errors demonstrate why one cannot rely on RO-based predictions to accurately quantify the impact of degradation effects (i.e. voltage, temperature or aging) on circuits' delay.

In summary: All in all, accurately analyzing the timing behavior of circuits by means of existing EDA tool flows (using their mature algorithms evolved over decades) along with detailed cell libraries that contain accurate delays of standard cells under combinations of degradation effects (voltage, temperature, aging) is indispensable – *This holds even more when considering the combined, instead of the individual, impact of these degradations due to the complex interactions between them at the physical, transistor and gate levels.*

3.2 Our Degradation-Aware Cell Libraries

In our work, we leverage the concept of degradation-aware cell libraries for aging and temperature effects from [2], [5]. However, we extend the process of library creation to further consider the effects of voltage reduction and the combined impact of these degradation effects. To explore a wide design space, we target the following operating ranges: For temperature, we consider a range from room temperature (25°C) to a worst-case temperature of 125°C in steps of 1°C. For aging, we estimate ΔV_{th} at different temperatures, voltages and lifetimes (see Section 6) and use that to create corresponding aging-aware cell libraries. We considered BTI aging only because it is the dominant aging mechanism [3]. However, to additionally consider other aging mechanisms that also impact the transistors' delay like Hot Carrier Injection (HCI), a combined model of BTI and HCI [11] can be employed during the cell library

creation process in which the overall V_{th} increase, caused by BTI and HCI together, is estimated. Finally, regarding voltage, we consider the range of 1.2V to 0.5V in 50mV steps. We generate cell libraries that model both *individual* and *joint* combinations of voltage-, temperature- and aging-induced degradations. All cell libraries have been created based on the 45nm open-cell library from Nangate [38] and the Predictive Technology Model (PTM) of transistors at the 45nm technology node [41], where HSPICE circuit simulations [42] are employed to characterize the delay and leakage/dynamic power of every gate under the individual and combined effects of temperature, aging and voltage degradations. Resulting degradation-aware cell libraries are compatible with existing commercial tool flows. Therefore, we can directly employ them to perform logic synthesis and timing analysis without any modifications. Our created cell libraries under the combined impact of voltage, temperature and aging are available at [43].

3.3 Probability of Error Calculation

After creating degradation-aware cell libraries, we employ them to perform detailed gate-level simulations and translate degradation-induced delay increases into corresponding system-level P_{error}. Fig 7 demonstrates our implemented flow to translate degradation effects in an RTL component (e.g., adder, multiplier, etc.) into corresponding P_{error}. In this work, we calculate P_{error} as, effectively, the error rate (ER). However, our approach is not limited to a specific error metric. Other metrics, e.g. worst-case or mean error, can be analogously considered when computing errors. We first synthesize the RTL of components to be characterized. We then perform a static timing analysis for the generated gate-level netlist under the desired degradation using the corresponding cell library. This provides us with the detailed information in standard SDF format of the delay of every gate within the netlist under the effects of the studied degradation. Finally, we perform simulations of the gate-level netlist for a given clock frequency with full, reduced or no timing guardband. Timing violations will represent themselves as errors in the outputs of the circuit. Any such error is considered an error of the component in this cycle, and we compute P_{error} as the fraction of simulated cycles with errors. In our implementation, we employ a Synopsys tool flow to perform the required RTL synthesis and static timing analysis (STA). We use ModelSim from Mentor Graphics for all gate-level simulations.

The probability of error analysis is input-dependent. Therefore, we employ a system-level simulator to extract the input data while applications are being executed. To determine the ultimate application-level effect of degradations, the output of gate-level simulations under the effects of degradation-induced errors (from Fig. 7) can, in turn, be injected back into the system-level simulation to quantify how the Quality-of-Service (QoS) is affected, as will be shown in our final experimental validations (see Section 6.4).

3.4 Degradation Effects

We employ our framework to analyze the resulting delay increase and P_{error} when voltage overscaling is performed without taking temperature and aging effects into account.

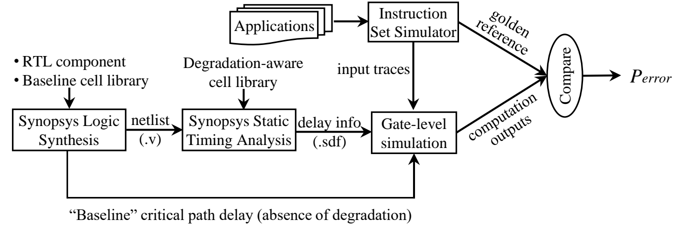


Fig. 7. Our proposed methodology for translating degradation effects into error probability. Degradation-aware cell libraries are employed either under *individual* or *combined* effects of temperature, aging and voltage.

We call this technique as done in the existing state of the art, Temperature- and Aging-Unaware Voltage Overscaling (T/A-Unware VOS). We demonstrate the obtained results in Fig. 8 in comparison with our proposed work (T/A-Aware VOS) in which voltage overscaling is performed under the combined impact of voltage, temperature and aging degradations. All analyses presented here and later in Sections (4, and 5) are done for a 32-bit multiplier with input data from the "gaussian" benchmark being executed on top of a GPU (implementation details in Section 6.1). Fig. 8(a) shows the normalized increase of delays from voltage scaling alone versus combined delay increases when considering a circuit at maximum temperature and aging, which are in turn both affected by scaled voltages. As can be seen, primary delay increases due to voltage scaling always dominate delay decreases due to the secondary impact of voltage on temperature and aging, i.e. combined delays in both cases increase monotonically with reduced voltages, but the opposing effect of voltage on temperature and aging helps to reduce combined delay increases by more than 10%.

We further analyze the impact of voltage-dependent delay increases on probabilities of error. In our T/A-Aware VOS, temperature and aging effects are thereby protected against using a timing guardband (see Eq. 6) that corresponds to worst-case temperature (125°C) along with worst-case aging at the end of lifetime (10 years) experienced at worst-case, i.e. nominal voltage. As can be noticed in Fig. 8(b), the resulting P_{error} at every reduced voltage step in our T/A-Aware VOS is always lower than the resulting P_{error} in the case of T/A-Unware VOS. This is because (1) delay increases are smaller as shown in Fig. 8(a) and (2) since maximum temperature and aging are reduced, the included timing guardband can help mitigate timing violations induced by voltage overscaling, which leads to less timing errors and thus smaller P_{error}.

This demonstrates that the full potential behind voltage overscaling is larger than what it has been often assumed because the resulting P_{error} when accounting for temperature and aging effects (which will naturally and inevitably occur in conjunction with voltage overscaling) is in reality smaller than what state of the art used to estimate.

It is noteworthy that our analysis shows that delay increases do not directly translate into equivalent error probabilities, i.e. *that there is a non-obvious and non-monotonic relationship between delay increase and P_{error}*. Furthermore, we consider here conservative timing guardbands covering worst-case temperature and aging degradations. However, due to the interdependencies between voltage, temperature

and aging, reductions in voltage will mitigate both temperature and aging. Hence, smaller guardbands can be used. In Sections (4 and 5), we focus on further breaking down the impact of voltage, temperature and aging on delay and energy, and the impact of delay and timing guardbands on P_{error} when aiming for voltage overscaling.

4 Delays and Errors under Voltage, Temperature and Aging Degradations

We apply our characterization to first investigate the effects of voltage, temperature and aging on a 32-bit multiplier *individually*. Fig. 9 shows the critical path increase along with the resulting P_{error} when no timing guardband is employed, i.e. when components are operated at their nominal "baseline" frequency obtained from synthesis in the absence of any degradations. To calculate P_{error}, gate-level simulations are performed for >30M inputs obtained by tracing the operands of multipliers in a Graphics Processing Unit (GPU) while executing a representative application (see Section 6). As Fig. 9 shows, similar delay increases caused by the three different degradations (voltage, temperature, and aging), result in different P_{error}. *Therefore, using delay increase as a metric is misleading when it comes to quantifying the impact of degradations at the system level.*

To better highlight the relationship between degradation-induced delay increase and the resulting P_{error}, we summarize these trade-offs for the multiplier under temperature, aging and voltage effects in Fig. 10. As can be seen, for the same delay increase, voltage reduction leads to the smallest P_{error} in general. This is due to the fact that even when the same amount of delay increase is induced, the distribution of delays and hence the number of violated paths varies based on the source of degradation (voltage, temperature or aging). As shown in Fig. 10, with a delay increase of 15% in the CP, voltage reduction results in only 10 violated paths, while aging and temperature lead to 33 and 34 violated paths, respectively. This is mainly due to the fact that different degradations impact the transistor parameters (e.g. V_{th}, μ, I_{on}, etc.) differently, as explained earlier in Section 1. *This indicates that when trading off delay increases due to voltage, temperature, aging, allowing for timing violations due to voltage reduction has the least impact on errors. This further motivates voltage overscaling as an effective and attractive means to trade off efficiency with accuracy—especially when multiple degradation effects are considered.*

At the same time, as discussed before, increases in circuit delays due to voltage scaling (see Eq. 3) are offset by reductions in temperature and aging. This allows trading off temperature and aging guardbands (i.e. performance) for V_{dd}-induced delay increases (i.e. energy) under voltage reductions and a given error constraint. Within this context, it is not beneficial to operate circuits under better-than-worst-case temperature and aging guardbands. As discussed above (Fig. 10), compared to temperature and aging, timing violations due to voltage reduction lead to relatively smaller P_{error}. Conversely, for a given upper bound on P_{error}, larger timing violations due to voltage reduction can be tolerated compared to temperature and aging. For example, for a $P_{error} = 0.08$, Fig. 10 indicates that only an around 5% violation of aging or temperature guardbands

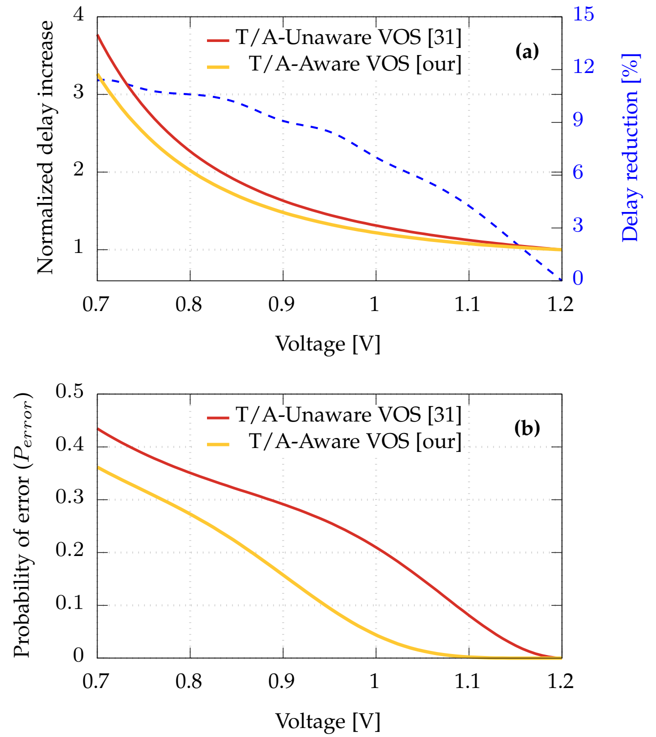


Fig. 8. Delay and P_{error} due to temperature- and aging-unaware voltage overscaling (T/A-Unware VOS) as done in prior art vs. temperature- and aging-aware overscaling (T/A-Aware VOS) proposed in this work.

are sustainable. By contrast, an up to 20% delay increase due to voltage reduction can be tolerated. As shown in Fig. 9(c), this corresponds to a V_{dd} reduction from 1.2V to 1.05V, which will in turn reduce temperature and aging by more than 6% and 50%, respectively.

These reductions in maximum temperature and aging can then be leveraged to narrow their guardbands accordingly. Thus, instead of operating at a given voltage with 5% loosened/violated guardbands, the circuit can be scaled to a lower V_{dd} resulting in smaller energy and worst-case guardbands while meeting the same P_{error} constraint. Fig. 11 summarizes these relationships for the 32-bit multiplier. For each P_{error} constraint, it shows the minimally achievable voltage, power and associated reductions in temperature and aging that meet the constraint.

5 Trading Off Accuracy With Efficiency

We address the problem of maximizing energy-performance efficiency for a given accuracy constraint (represented as an upper bound on P_{error} for each component in a design) under the joint impact of voltage, temperature and aging. As previously explained, both temperature and aging strongly depend on voltage. Therefore, voltage reduction provides a threefold gain: (1) It results in considerable savings in both dynamic and leakage power (see Eq. 1); (2) It minimizes the required timing guardband to protect against temperature effects because lower voltage directly leads to lower on-chip temperatures (see Fig. 1); and (3) It minimizes the required timing guardband to protect against aging effects because lower voltage together with lower temperature leads to a considerable reduction in aging-induced degradation (see

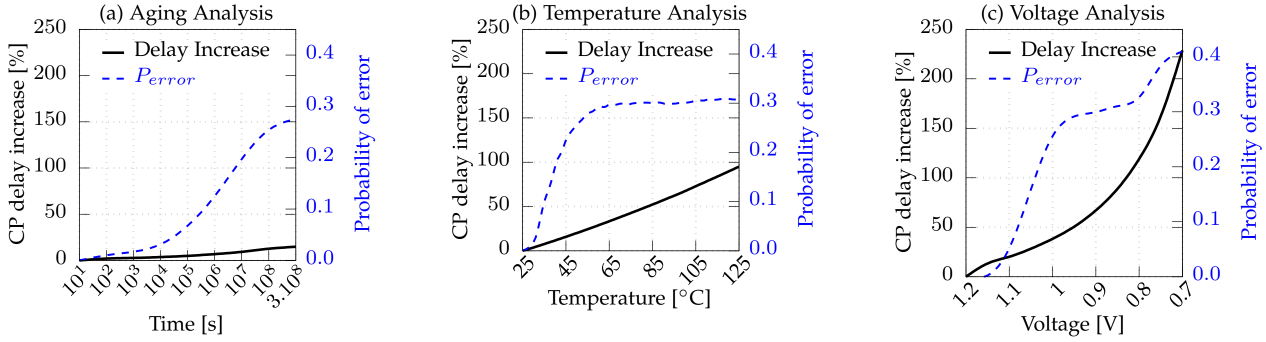


Fig. 9. Impact of aging-, temperature- and voltage- induced degradation on the critical path (CP) delay of a 32-bit multiplier with the resulting P_{error}.

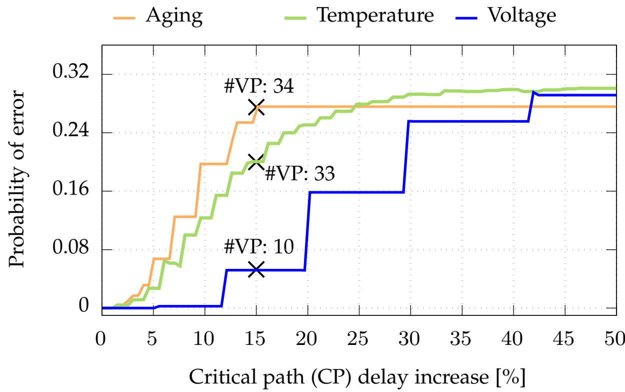


Fig. 10. Relationship between timing violations due to different degradation effects and the resulting P_{error} in a 32-bit multiplier. The same delay increase leads to different P_{error} due to a different distribution of delays and hence varied number of violated paths (#VP). V_{dd} reduction, in general, has the least impact on P_{error}.

Fig. 2). *Thus, voltage overscaling not only improves energy but also mitigates the delay increase from aging and temperature leading to smaller associated guardbands and increased performance.*

In contrast to the isolated analysis in Figs. 9 and 10, P_{error} in the final system will be determined by *a combination of voltage reduction and corresponding temperature and aging degradations*, as demonstrated in Fig. 11. Fig. 12 shows our flow for finding an operating point of a design that maximizes efficiency for a given accuracy constraint under the combined impact of voltage reduction, aging and temperature effects. We first determine the minimum voltage that meets a P_{error} below the predetermined accuracy constraint for each component in the design. Starting from the nominal voltage, we iteratively reduce V_{dd} and translate the reduced V_{dd} in each step into a resulting P_{error} (under the combined voltage, temperature and aging effects) for each component. We keep reducing V_{dd} as long as all the resulting P_{error} remains below the predetermined per-component constraints, as described in Section 3.3 and Fig. 7. At every reduced V_{dd} step, the resulting power savings are evaluated and used to estimate a corresponding peak temperature reduction as well as the associated temperature guardband (T_{gb}). Furthermore, based on the resulting temperature along with the V_{dd} level, aging-induced degradations and the associated aging guardband (A_{gb}) are estimated. Converting temperature and aging into corresponding guardbands is performed using static timing analysis with degradation-aware cell libraries as described in Section 3 and Fig. 7. Details on how power, aging and temperature are estimated are presented in Section 6 along with details on the experimental setup. Finally, based on the voltage level along with the resulting temperature and aging, timing errors under the *combined* effects of voltage, temperature and aging are analyzed with estimated temperature and aging guardbands included[2]. This provides us with the overall P_{error} at that voltage level. Fig. 11 presented earlier shows an example of power, temperature and aging savings along with the minimum V_{dd} over a range of P_{error} constraints for a 32-bit multiplier in a GPU. After determining the minimum V_{dd}, the resulting power savings together with the reduced temperature and aging guardbands are used to evaluate the efficiency improvement.

To evaluate the effectiveness of our temperature- and aging-aware voltage overscaling (T/A-Aware VOS) technique, we implemented the following two versions:

(1) *Fixed T/A-Aware VOS*: The circuit here is protected against temperature-induced and aging-induced degradations through employing *fixed* timing guardbands (A_{gb} and T_{gb}) that correspond to the worst-case temperature (125°C) and worst-case aging at nominal voltage. In such a case, the employed guardbands are conservative and *constantly applied throughout* even through temperature and aging effects might become less at lower voltages.

(2) *Adaptive T/A-Aware VOS*: The circuit here is protected against temperature-induced and aging-induced degradations through employing *adaptive* timing guardbands (A_{gb} and T_{gb}) that correspond to the actual worst-case temperature and aging degradations induced at the new reduced voltage level. In such a case, the employed guardbands are not conservative as they will be *small, yet sufficient* to protect against temperature and voltage effects at lower V_{dd}.

In Fig. 13 we summarize the evaluation results for efficiency improvement represented by (a) energy improvement, (b) delay improvement, and (c) the improvement in energy-delay product (EDP) of the 32-bit multiplier. As can be noticed, *fixed T/A-Aware VOS* technique provides (under the same P_{error} constraint) better results than an *adaptive T/A-Aware VOS* w.r.t. energy improvements (see Fig. 13(a)).

2. Estimated temperature and aging guardbands are used to determine the final frequency for the design as shown in Eq. 6.

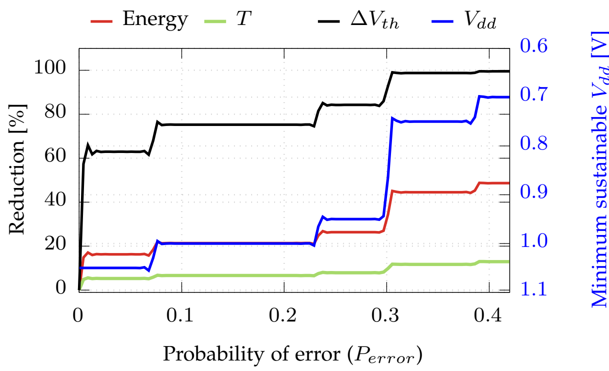


Fig. 11. Relationship between P_{error} and resulting reductions in temperature (T), aging (ΔV_{th}) and power along with the minimum selected (sustainable) V_{dd}, shown on the right y-axis. The higher the accepted P_{error} the lower the V_{dd} and thus the higher the reduction. Both temperature and aging have a strong dependency on voltage. Thus, reducing V_{dd} effectively mitigate both of them.

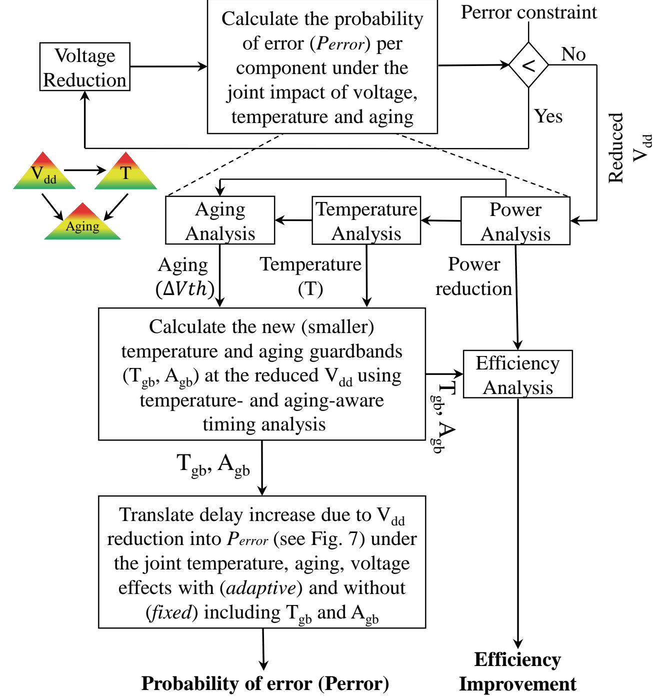


Fig. 12. Our design flow for finding the minimum V_{dd} and associated power savings and reductions in temperature and aging guardbands that can be sustained for a given P_{error} constraint.

The delay improvement obtained due to timing guardband adaptation is demonstrated in Fig. 13(b). As can be seen, the higher the P_{error}, the larger the gain in the delay improvement. This is because when we loosen the P_{error} constraint (i.e. accepting more errors), a lower V_{dd} will be selected leading to a further mitigation for temperature and aging effects. Hence, narrower timing guardbands (A_{gb} and T_{gb}) will be needed resulting in a larger delay improvement. Finally, Fig. 13(c) shows the EDP improvement. As can be seen, *adaptive T/A-Aware VOS* achieves better results under the same P_{error} constraint than *fixed T/A-Aware VOS*, which is mainly due to the delay gains from at similar energy when using guardband adaptation (see Fig. 13(b)).
Concisely, under the joint effect of temperature, aging and voltage, combining voltage reductions with adaptively exploiting associated temperature and aging guardband reductions will provide the best trade-off between accuracy and efficiency. *Not only does it maintain close-to-maximum energy savings, it also minimizes the performance loss (caused by temperature and aging guardbands), leading to further improvements in overall efficiency.*

6 Evaluation and Comparisons

We evaluate our approach by applying it to the multipliers in a General Purpose Graphics Processing Unit (GPGPU). In our experiments, we assume that the critical paths of GPUs [46] typically pass through the multipliers. To further validate this assumption, we synthesized an open-source RTL of GPU [56] using our baseline 45 nm standard cell library, where we observed that the multipliers are in fact on the critical paths. Similar to existing approaches in the voltage overscaling and approximate computing domain [7], we assume that control blocks/stages that can not tolerate errors are separated and protected against degradations through traditional schemes, such as stronger gates [44].

Note that our approach is not specific to multipliers or GPUs and can be applied to other units and general systems. In practice, our technique provides designers with a new degree of freedom in which accuracy is traded off with efficiency. We use GPUs as an example since they are typically power-hungry and suffer from high temperatures due to elevated power densities. It is noteworthy that arithmetic units in GPUs account for up to 20% of total power [45]. Furthermore, applications running on top of GPUs, such as multimedia, image/graphics or video processing, can often inherently tolerate errors and therefore allow for a certain P_{error} to be accepted without otherwise applying fault-tolerance or error correction mechanisms. Note that our approach equally applies to applications that are not error-tolerant, but in such cases other techniques to cope with faults [47]–[51] (e.g., redundancy) need to be additionally included.

6.1 Experimental Setup

We employ GPGPUSim [52] as a cycle-accurate instruction-set simulator for a typical NVIDIA Fermi GPU architecture. The simulated architecture consists of 16 streaming multiprocessors (SMs) each with 32 multipliers of 32-bit width. We apply the same P_{error} constraint to all multipliers and find the smallest V_{dd} that satisfies all constraints. The simulator enables us to trace the activities of microarchitecture components while executing applications. We select various representative benchmarks from the CUDA SDK (a.k.a Ispass) [52] and Rodinia [53] GPU benchmark suites. We examined a wide variety of applications, where we excluded those that cannot tolerate any errors. We connect GPGPUSim with GPUWattch [45] to extract the overall power traces of each application over its execution time. For consistency with our characterization flow at the device, gate and circuit levels (details in Section 3), we configure GPUWatch to target a 45nm technology node with operating voltages from 1.2V to 0.5V. Extracted power profiles of each application are then employed to obtain the result-

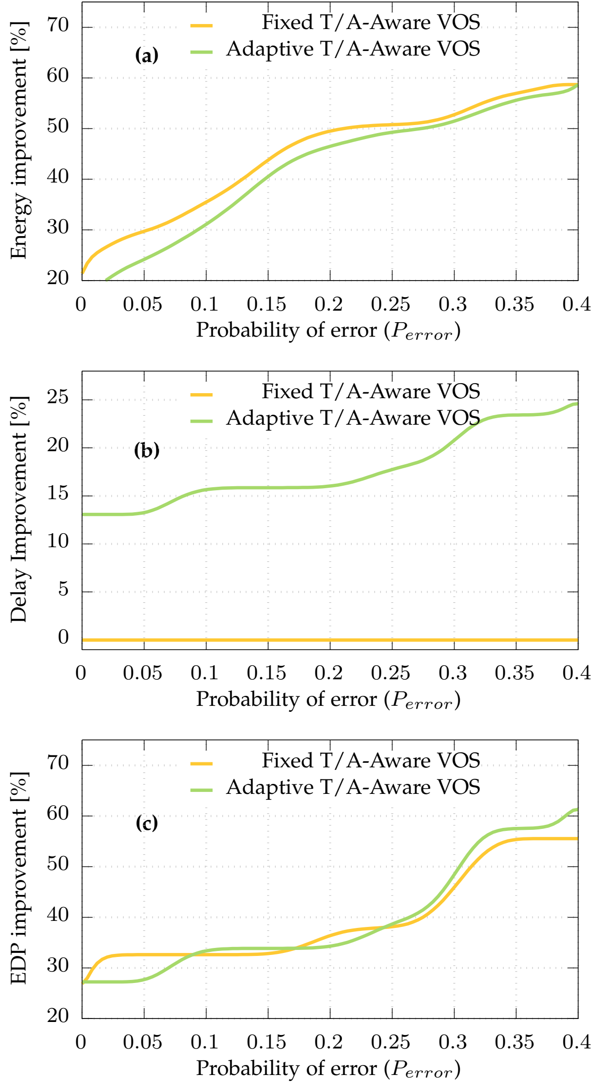


Fig. 13. Comparison between *fixed* and *adaptive* temperature- and aging-aware (T/A-Aware) VOS when accuracy is traded off with efficiency under the combined impact of voltage, temperature and aging.

ing temperature profile using the thermal HotSpot simulator [54]. For aging, we employ a state-of-the-art physics-based aging model [3], which accurately estimates ΔV_{th} induced by BTI under the dependency of temperature and voltage. We consider peak temperature and estimate ΔV_{th} under worst-case activity factors (i.e. under a continuous BTI stress). This matches traditional worst-case analysis to determine timing and aging guardbands, respectively. In our error analysis and optimization, it results in a conservative analysis of P_{error}. In case of aging, note that in recent technology nodes, the dependency of BTI on activity factor is weaker [3], [55]. Furthermore, as explained earlier (Section 3.2), the effects of HCI aging mechanisms can be additionally considered be employing a physics-based aging model that estimates overall ΔV_{th} under the joint impact of BTI and HCI. Finally, we employ temperature- and aging-aware timing analysis using degradation-aware cell libraries presented in Section 3 to translate reductions in temperature and aging into the corresponding reductions in temperature

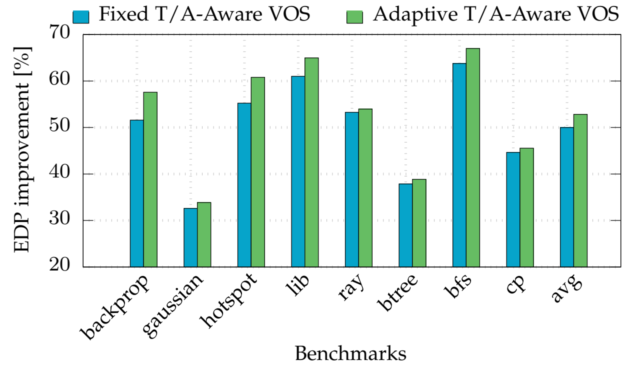


Fig. 14. Comparison between *fixed* and *adaptive* Temperature- and Aging-Aware VOS in terms of energy-delay product (EDP) improvement as a result of a P_{error} constraint of 0.15.

and aging guardbands (i.e. T_{gb}, A_{gb}). At every voltage step, resulting temperature and aging due to the reduction in voltage are estimated and a cell library is created under the joint impact of voltage, temperature and aging. In our experimental setup, we use predetermined temperature and aging levels for each V_{dd} extracted from offline analysis of examined applications.

As explained in Section 3, our approach to translate timing violations induced by voltage reduction into corresponding P_{error} needs the input operands of the studied arithmetic component (i.e. multiplier) in order to drive the required gate-level simulations and thus calculate P_{error} as shown in Fig. 7. To obtain such input traces, we modify GPGPUSim to trace the input operands of multipliers while the GPU executes an application. Finally, to evaluate how the resulting errors influence the final output of a running application, the complete output traces from the multipliers obtained from gate-level simulations are injected back into GPGPUSim. This enables us to see how timing violations originating at lower (i.e. device, gate and circuit) levels will finally translate into application-level errors.

6.2 Accuracy and Efficiency Trade-offs

To evaluate the achievable efficiency improvement at different P_{error} constraints, we quantify the energy and delay savings when our technique is applied to the multipliers of the GPU. Energy savings stem from power savings due to V_{dd} reduction. Delay savings stem from minimizing temperature and aging guardbands. Note that smaller timing guardbands mean less performance loss as Eq. 6 clarifies. The efficiency improvement, at a given P_{error} constraint, is calculated as a relative reduction in the energy-delay product (EDP) compared to a baseline in which the system operates at nominal V_{dd} with complete temperature and aging guardbands (i.e. no accuracy loss since the reliability is fully sustained and thus no timing errors will occur).

We first extend our evaluation presented in Fig. 13(c), in which we compare between *fixed* and *adaptive* T/A-Aware VOS, to cover more application scenarios. As can be noticed from Fig. 14, our *adaptive* T/A-Aware VOS technique always provides a higher efficiency (represented by EDP improvement). As such, we use it for the remaining evaluations.

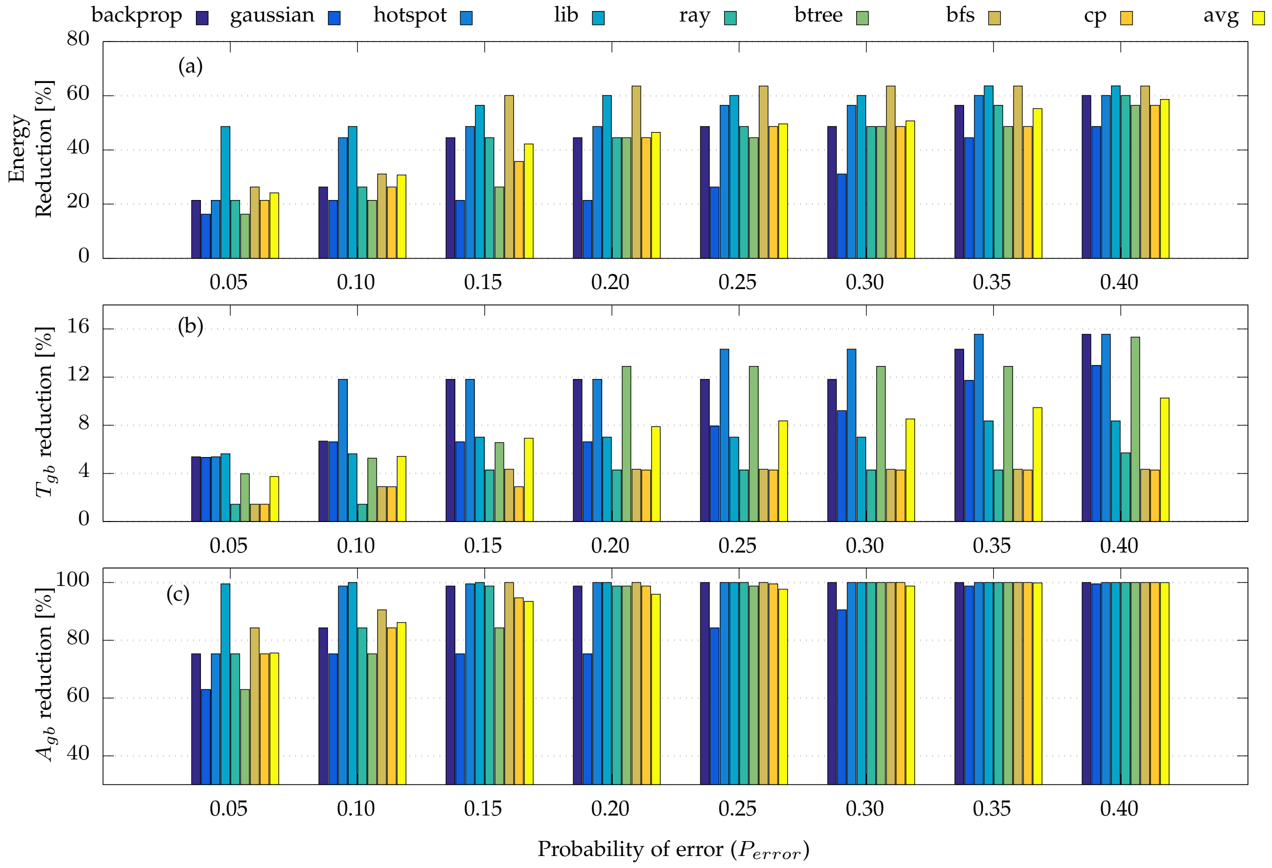


Fig. 15. Our achieved reductions in (a) energy, (b) temperature guardband (T_{gb}) and (c) aging guardband (A_{gb}) for various P_{error} constraints.

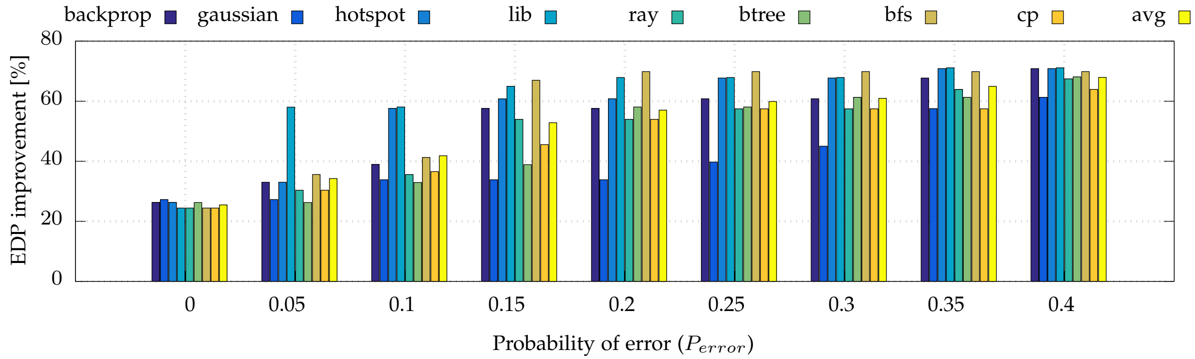
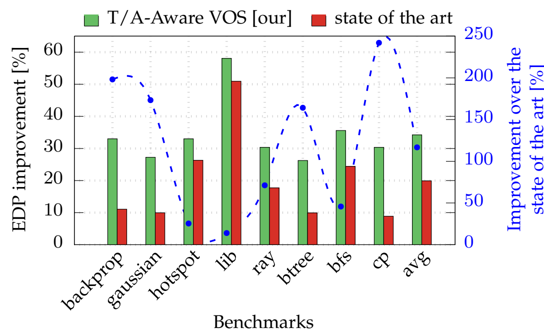


Fig. 16. Trade-offs between efficiency (represented by energy-delay product EDP) and accuracy (represented by varied P_{error} constraints) obtained by our adaptive temperature- and aging-aware voltage overscaling technique (adaptive T/A-Aware VOS). Accepting higher error probabilities (P_{error}) lead to larger efficiency improvements due to the lower selected V_{dd} level.

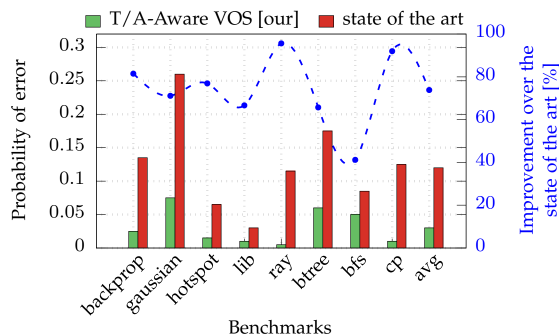
Fig 15 then further details the achieved energy savings along with reductions in temperature and aging guardbands for different P_{error} constraints using our adaptive T/A-Aware VOS technique. As expected, for a higher P_{error}, the achieved savings are increased due to the lower operating voltages being selected. For instance, when P_{error} is 0.2, the savings in energy, T_{gb} and A_{gb} reach on average 44%, 7.2% and 96%, respectively. Note that the high aging reductions are due to aging-induced ΔV_{th} having an exponential dependency on voltage and temperature as demonstrated in Fig. 2. Therefore, reductions in temperature and voltage from applying our technique lead to such aging timing guardband reductions.

Finally, Fig. 16 demonstrates the trade-off between accuracy and efficiency under the joint effect of temperature, aging and voltage. As can be noticed, a higher P_{error} leads to a higher EDP improvement. For example, a P_{error} of merely 0.05 provides an average 34% and up to 58% increase in EDP. For a higher P_{error} of 0.1 and 0.4, the EDP improvement reaches on average 41% and 67%, respectively.

Another observation obtained from this analysis is that an efficiency improvement of 25% (on average) can be

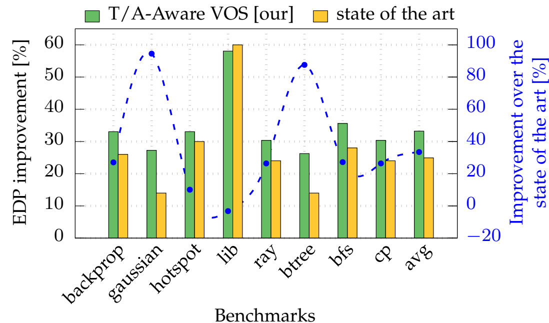


(a) For the same P_{error} constraint of 0.05, adaptive T/A-Aware VOS achieves higher efficiency improvement (represented by EDP).

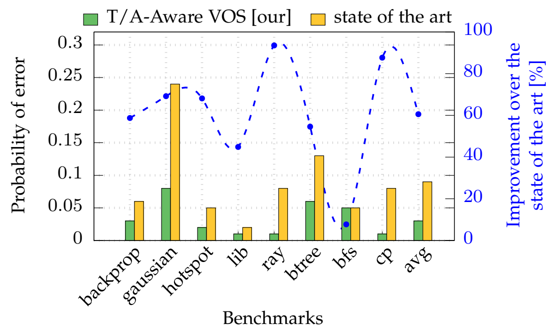


(b) For the same efficiency improvement of 30% (represented by EDP), our adaptive T/A-Aware VOS always results in smaller P_{error}.

Fig. 17. Comparison between state of the art voltage overscaling (e.g. [31]), which is temperature- and aging-unaware (T/A-Unaware VOS), and our voltage overscaling (adaptive T/A-Aware VOS) in which the combined impact of voltage, temperature and aging is considered. Left and right y-axis show the resulting P_{error} and the achieved improvement, respectively.



(a) For the same P_{error} constraint of 0.05, our adaptive T/A-Aware VOS results in 33% higher efficiency (represented by EDP).



(b) For the same efficiency improvement of 30% (represented by EDP), our adaptive T/A-Aware VOS results in 60% smaller P_{error}.

Fig. 18. Comparison between state of the art timing guardband reduction (TGR) [24], which aims at loosening guradbands to gain efficiency, and our voltage overscaling (adaptive T/A-Aware VOS) in which the combined impact of voltage, temperature and aging is considered. Left and right y-axis show the resulting P_{error} and the achieved improvement, respectively.

achieved even under a P_{error} constraint of 0 (i.e. when accuracy loss is not tolerated). This is because baseline guardbands are overly conservative. E.g., applications in reality never reach a maximum temperature of $125°$C (the maximum temperature observed across our evaluated applications is $107°$C) even at nominal voltage. Furthermore, baseline guardbands are set based on an isolated and additive analysis of worst-case temperature and aging degradations, where combined delays along each gate in a circuit path in reality never all reach worst-case temperature and aging contributions simultaneously. All combined, this allows guardbands and/or voltage to be scaled over the baseline without incurring any timing errors.

6.3 Comparisons with State of the Art

Voltage overscaling has long been used to trade off efficiency and accuracy [4], [26]–[35]. Specifically, approaches similar to [31] aim to reduce the chip's operating V_{dd} to reduce energy consumption and to mitigate aging degradation. However, all of the previous work has studied the effects of voltage overscaling on timing errors *in isolation* from temperature and/or aging effects. In other words, timing violations in state of the art voltage overscaling originate *solely* from the impact of voltage reduction on prolonging delays of circuit paths, and effects of temperature and aging were neglected assuming the circuit will be protected against them using traditional schemes, e.g. by including the required timing guardbands.

In the following analysis, we investigate the consequences of looking at the voltage reduction as a sole source of degradation instead of considering the *combined* impact of voltage, temperature and aging on timing errors. Such an analysis allows us to explore the full potential behind voltage overscaling when trading off efficiency with accuracy. Fig. 17(a) demonstrates a comparison between *T/A-Unaware VOS*, which represents state of the art voltage overscaling (e.g. [31]) in which temperature and aging effects are not considered, and our *adaptive T/A-Aware VOS* technique in which the combined impact of voltage, temperature and aging is considered. As shown in Fig. 17(a), under the same P_{error} constraint, adaptive T/A-Aware VOS always achieves higher EDP improvements compared to state of the art T/A-Unware VOS. The improvement reaches on average around 116% and up to 240%. As discussed in Section 3.3, Fig. 8, under the same P_{error}, T/A-Aware VOS will result in selecting smaller reduced voltage level leading to higher energy savings. In addition, the adaptation in timing guardbands leads to delay improvements (see Fig. 13(b)). In Fig. 17(b), we compare the resulting P_{error} for a given targeted EDP improvement (e.g. 30% in this analysis) for T/A-Unaware VOS and adaptive T/A-Aware VO. As shown, our technique results in an average of 73% (up to 95%) smaller P_{error} compared to the state of the art.

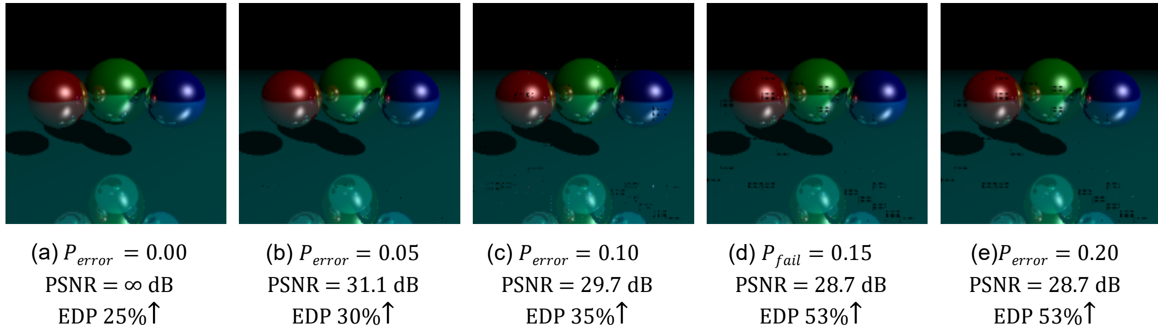


Fig. 19. Impact of P_{error} on the output images when the "RAY" benchmark is analyzed when our proposed T/A-Aware VOS technique is in use. Errors (obtained from detailed gate-level simulations) are injected into all 32 multipliers within the GPU streaming multi-processor.

This confirms that by looking solely at voltage alone and thus neglecting temperature and aging effects, a lower effectiveness of VOS will be perceived.

Apart from voltage overscaling, trading off efficiency with accuracy can be alternatively done by loosening the timing guardbands to gain performance [24](i.e. increasing the performance in exchange for a certain amount of incurred P_{error} by simply narrowing guardbands and running the design at a faster clock). In Fig. 18(a), we compare the EDP improvement due to our adaptive T/A-Aware VOS to a state-of-the-art timing guardband reduction (TGR) technique [24] under the same P_{error} constraint. As shown, our technique results in, on average, 33% and up to 90% better EDP compared to [24], when a constraint of $P_{error} = 0.05$ is targeted. Conversely, for a given targeted EDP improvement (e.g., 30% in this analysis), our technique results in 60% less P_{error} compared to TGR (Fig. 18(b)). This is because unlike TGR [24], where efficiency is only gained due to performance improvements, our technique profits from both performance and energy improvements.

6.4 Quality of Service Evaluation

Finally, we quantify the ultimate impact of errors generated by multipliers when our technique is implemented in GPUs. To visually demonstrate the impact of trading off accuracy with efficiency, we selected an image processing benchmark "RAY" from [52] as an example. Fig. 19 shows the final output images, along with their Peak to Signal Ratio (PSNR) and the achieved EDP improvement. Note that for human eyes, a PSNR above 30dB is typically considered an acceptable level [57]. As can be observed, for a P_{error} of up to 0.10, EDP is improved by 35% and image quality remains high, with a PSNR of nearly 30dB. For a larger P_{error} of 0.20, EDP improvement reaches 53% and the image quality marginally degrades, with a small PSNR reduction to merely 28.7dB (i.e. just 1.3dB below the accepted level).

7 Summary and Conclusions

In this work, we investigated trade-offs between accuracy and efficiency under the joint interactions of temperature, aging and voltage. In systems that can tolerate faults, voltage can be overscaled in exchange for improved efficiency. Our investigation revealed that evaluating voltage overscaling in isolation and solely based on its induced delay increase (as done in state of the art) is misleading, as there are opposing interactions between degradations, where voltage impacts temperature and aging with a non-obvious impact on combined delays and error probabilities. We accurately model, analyze and quantify joint voltage-temperature-aging interactions and trade-offs, and we propose a novel adaptive and temperature- and aging-aware voltage overscaling optimization to maximize energy-performance efficiency under a probability of error (P_{error}) constraint. Results from applying such an approach to multipliers in a GPU demonstrate that for a P_{error} of merely 0.05, efficiency can be improved by 30% with little to no impact on the final output in an image processing application. Our created degradation-aware cell libraries under the joint impact of voltage, temperature and aging are available at [43]. They can be directly used within existing EDA tool flows without any further modifications. Therefore, they can enable designers to accurately evaluate the trade-offs between accuracy and efficiency provided by voltage overscaling using their standard design flows.

References

[1] D. Wolpert and P. Ampadu, "Temperature effects in semiconductors," in *Managing Temperature Effects in Nanoscale Adaptive Systems*, 2012, pp. 15–33.

[2] H. Amrouch, B. Khaleghi, and J. Henkel, "Optimizing temperature guardbands," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017, pp. 175–180.

[3] N. Parihar, N. Goel, S. Mukhopadhyay, and S. Mahapatra, "Bti analysis toolmodeling of nbti dc, ac stress and recovery time kinetics, nitrogen impact, and eol estimation," *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 392–403, 2018.

[4] S. Arasu, M. Nourani, J. M. Carulli, and V. K. Reddy, "Controlling aging in timing-critical paths," *IEEE Design & Test*, vol. 33, no. 4, pp. 82–91, 2016.

[5] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Reliability-aware design to suppress aging," in *Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE*. IEEE, 2016, pp. 1–6.

[6] D. Ernst, S. Das, S. Lee, D. Blaauw, T. Austin *et al.*, "Razor: circuit-level correction of timing errors for low-power operation," *IEEE Micro*, vol. 24, no. 6, pp. 10–20, Nov 2004.

[7] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *ETS*, 2013.

[8] C. Hu, *Modern semiconductor devices for integrated circuits*. Prentice Hall, 2010.

[9] V. M. van Santen, J. Martin-Martinez, H. Amrouch, M. M. Nafria, and J. Henkel, "Reliability in super-and near-threshold computing: A unified model of rtn, bti, and pv," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 1, pp. 293–306, 2018.

[10] D. Jang, E. Bury, R. Ritzenthaler, M. G. Bardon, T. Chiarella *et al.*, "Self-heating on bulk finfet from 14nm down to 7nm node," in *Electron Devices Meeting (IEDM), 2015 IEEE International*. IEEE, 2015, pp. 11–6.

[11] H. Amrouch, V. M. van Santen, T. Ebi, V. Wenzel, and J. Henkel, "Towards interdependencies of aging mechanisms," in *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design*. IEEE Press, 2014, pp. 478–485.

[12] V. Chandra and R. Aitken, "Impact of voltage scaling on nanoscale SRAM reliability," in *Proceedings of the Conference on Design, Automation and Test in Europe*. European Design and Automation Association, 2009, pp. 387–392.

[13] K. He, A. Gerstlauer, and M. Orshansky, "Controlled timing-error acceptance for low energy IDCT design," in *DATE*, 2011.

[14] D. Mohapatra, G. Karakonstantis, and K. Roy, "Significance driven computation: a voltage-scalable, variation-aware, quality-tuning motion estimator," in *Proceedings of the International Symposium on Low Power Electronics and Design*, ser. ISLPED '09. ACM, 2009.

[15] G. Karakonstantis, D. Mohapatra, and K. Roy, "System level DSP synthesis using voltage overscaling, unequal error protection and adaptive quality tuning," in *IEEE Workshop on Signal Processing Systems (SIPS)*, 2009.

[16] A. B. Kahng, S. Kang, R. Kumar, and J. Sartori, "Slack redistribution for graceful degradation under voltage overscaling," in *Proceedings of the 2010 Asia and South Pacific Design Automation Conference*, ser. ASPDAC '10. IEEE Press, 2010, pp. 825–831.

[17] F. Kurdahi, A. Eltawil, K. Yi, S. Cheng, and A. Khajeh, "Low-Power Multimedia System Design by Aggressive Voltage Scaling," *IEEE Transactions on VLSI Systems*, vol. 18, no. 5, pp. 852–856, 2010.

[18] R. Hedge and N. R. Shanbhag, "Soft digital signal processing," *IEEE Transactions on VLSI Systems*, vol. 9, no. 6, pp. 813–823, 2001.

[19] B. Shim and N. R. Shanbhag, "Energy-efficient soft error-tolerant digital signal processing," *IEEE Transactions on VLSI Systems*, vol. 14, no. 4, pp. 336–348, 2006.

[20] G. V. Varatkar and N. R. Shanbhag, "Energy-efficient motion estimation using error tolerance," in *International Symposium on Low Power Electronics and Design*, ser. ISLPED '06, 2006.

[21] K. He, A. Gerstlauer, and M. Orshansky, "Circuit-level timing-error acceptance for design of energy-efficient dct/idct-based systems," *Transactions on Circuits and Systems for Video Technology (TCSVT*, vol. 23, no. 6, pp. 961–974, 2013.

[22] H. Afzali-Kusha, O. Akbari, M. Kamal, and M. Pedram, "Energy and reliability improvement of voltage-based, clustered, coarse-grain reconfigurable architectures by employing quality-aware mapping," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 480–493, 2018.

[23] M.-L. Li, P. Ramachandran, U. R. Karpuzcu, S. K. S. Hari, and S. V. Adve, "Accurate microarchitecture-level fault modeling for studying hardware faults," in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*. IEEE, 2009, pp. 105–116.

[24] H. Amrouch, B. Khaleghi, A. Gerstlauer, and J. Henkel, "Towards aging-induced approximations," in *Design Automation Conference (DAC), 2017 54th ACM/EDAC/IEEE*. IEEE, 2017, pp. 1–6.

[25] B. Boroujerdian, H. Amrouch, J. Henkel, and A. Gerstlauer, "Trading off temperature guardbands via adaptive approximations," in *International Conference on Computer Design (ICCD)*. IEEE, 2018.

[26] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of PMOS NBTI effect for robust nanometer design," in *Design Automation Conference, 2006 43rd ACM/IEEE*. IEEE, 2006, pp. 1047–1052.

[27] S. Gupta and S. S. Sapatnekar, "GNOMO: Greater-than-NOMinal Vdd operation for BTI mitigation." in *ASP-DAC*, 2012, pp. 271–276.

[28] A. Tiwari and J. Torrellas, "Facelift: Hiding and slowing down aging in multicores," in *Microarchitecture, 2008. MICRO-41. 2008 41st IEEE/ACM International Symposium on*. IEEE, 2008, pp. 129–140.

[29] J. Abella, X. Vera, and A. Gonzalez, "Penelope: The NBTI-Aware Processor," in *MICRO*, 2007, pp. 85–96.

[30] J. Kong, S. W. Chung, and K. Skadron, "Recent thermal management techniques for microprocessors," *ACM Comput. Surv.*, vol. 44, no. 3, pp. 13:1–13:42, 2012.

[31] L. Zhang and R. P. Dick, "Scheduled voltage scaling for increasing lifetime in the presence of NBTI," in *Proceedings of the 2009 Asia and South Pacific Design Automation Conference*. IEEE Press, 2009, pp. 492–497.

[32] P. Pop, K. H. Poulsen, V. Izosimov, and P. Eles, "Scheduling and voltage scaling for energy/reliability trade-offs in fault-tolerant time-triggered embedded systems," in *Proceedings of the 5th IEEE/ACM international conference on Hardware/software codesign and system synthesis*. ACM, 2007, pp. 233–238.

[33] U. R. Karpuzcu, B. Greskamp, and J. Torrellas, "The BubbleWrap many-core: popping cores for sequential acceleration," in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 2009, pp. 447–458.

[34] X. Chen, Y. Wang, Y. Cao, Y. Ma, and H. Yang, "Variation-aware supply voltage assignment for simultaneous power and aging optimization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 2143–2147, 2012.

[35] M. Cho, C. Tokunaga, M. M. Khellah, J. W. Tschanz, and V. De, "Aging-aware Adaptive Voltage Scaling in 22nm high-K/metal-gate tri-gate CMOS," in *Custom Integrated Circuits Conference (CICC), 2015 IEEE*. IEEE, 2015, pp. 1–4.

[36] R. Zheng, J. Velamala, V. Reddy, V. Balakrishnan, E. Mintarno *et al.*, "Circuit aging prediction for low-power operation," in *Custom Integrated Circuits Conference, 2009. CICC'09. IEEE*. IEEE, 2009, pp. 427–430.

[37] H. Amrouch, B. Khaleghi, and J. Henkel, "Voltage adaptation under temperature variation," in *2018 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, 2018, pp. 57–60.

[38] "45nm Open Cell Library," https://www.silvaco.com/products/nangate/FreePDK45_Open_Cell_Library.

[39] C. R. Lefurgy, A. J. Drake, M. S. Floyd, M. S. Allen-Ware, B. Brock *et al.*, "Active management of timing guardband to save energy in power7," in *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*, 2011.

[40] G.-Y. Wei and M. Horowitz, "A fully digital, energy-efficient, adaptive power-supply regulator," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 4, 1999.

[41] "Predictive Technology Model," http://ptm.asu.edu/.

[42] "Synopsys EDA Tools Flow," http://www.synopsys.com/.

[43] "Created Degradation-Aware Cell Libraries," Upon publication, download link will be made publicly available.

[44] S. Roy, D. Liu, J. Singh, J. Um, and D. Z. Pan, "Osfa: A new paradigm of aging aware gate-sizing for power/performance optimizations under multiple operating conditions," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 10, pp. 1618–1629, 2016.

[45] J. Leng, T. Hetherington, A. ElTantawy, S. Gilani, N. S. Kim *et al.*, "Gpuwattch: enabling energy optimizations in gpgpus," in *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3. ACM, 2013, pp. 487–498.

[46] C. Nvidia, "Nvidias next generation cuda compute architecture: Kepler gk110," *Whitepaper*, 2012.

[47] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "Exploiting structural duplication for lifetime reliability enhancement," in *ACM SIGARCH Computer Architecture News*, vol. 33, no. 2. IEEE Computer Society, 2005, pp. 520–531.

[48] V. Claesson, S. Poledna, and J. Söderberg, "The XBW model for dependable real-time systems," in *icpads*. IEEE, 1998, p. 130.

[49] H. Kopetz, *Real-time systems: design principles for distributed embedded applications*. Springer Science & Business Media, 2011.

[50] H. Kopetz, A. Damm, C. Koza, M. Mulazzani, W. Schwabl *et al.*, "Distributed fault-tolerant real-time systems: The Mars approach," *IEEE Micro*, vol. 9, no. 1, pp. 25–40, 1989.

[51] D. Zhu and H. Aydin, "Reliability-aware energy management for periodic real-time tasks," *IEEE Transactions on Computers*, vol. 58, no. 10, pp. 1382–1397, 2009.

[52] A. Bakhoda, G. L. Yuan, W. W. Fung, H. Wong, and T. M. Aamodt, "Analyzing cuda workloads using a detailed gpu simulator," in *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*. IEEE, 2009, pp. 163–174.

[53] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer *et al.*, "Rodinia: A benchmark suite for heterogeneous computing," in *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on*. Ieee, 2009, pp. 44–54.

[54] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron *et al.*, "Hotspot: A compact thermal modeling method-

ology for early-stage vlsi design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, 2006.
[55] A. Thirunavukkarasu, H. Amrouch, J. Joe, N. Goel, N. Parihar *et al.*, "Device to circuit framework for activity-dependent nbti aging in digital circuits," *IEEE Transactions on Electron Devices*, vol. 66, no. 1, pp. 316–323, Jan 2019.
[56] https://github.com/jbush001/NyuziProcessor.
[57] N. Thomos, N. Boulgouris, and M. Strintzis, "Optimized transmission of JPEG2000 streams over wireless channels," *IEEE TIP*, vol. 15, no. 1, pp. 54–67, 2006.
[58] M. Subrat, H. Wong, R. Tiwari, A. Chaudhary, N. Parihar, R. Rao, S. Motzny, V. Moroz, and S. Mahapatra, "Predictive TCAD for NBTI stress-recovery in various device architectures and channel materials," In Reliability Physics Symposium (IRPS), 2017 IEEE International, pp. 6A-3. IEEE, 2017.

Hussam Amrouch (S'11, M'15) is a Research Group Leader at the Chair for Embedded Systems (CES), Karlsruhe Institute of Technology (KIT), Germany. He is leading of the Dependable Hardware research group. He received his Ph.D. degree with distinction (Summa cum laude) from KIT (2015). His main research interests are design for reliability, thermal-aware VLSI design, modeling and mitigating aging effects at the device/circuit levels. He holds seven HiPEAC Paper Awards. He has three best paper nominations at DAC'16, DAC'17 and DATE'17 for his work on reliability. He currently serves as Associate Editor at Integration, the VLSI Journal. He is a member of the IEEE. ORCID 0000-0002-5649-3102

Seyed Borna Ehsani is currently a first-year PhD student at the Department of Computer Science and Engineering, University of Washington, WA, USA. He obtained his Bachelors degree in Computer Engineering in the summer of 2018, from Sharif University of Technology, Tehran, Iran. His primary field of interest is computer architecture, particularly energy-efficient hardware designs and their application in embedded and heterogeneous systems.

Andreas Gerstlauer (SM'11) is an Associate Professor in the Electrical and Computer Engineering Department at The University of Texas at Austin, TX, USA. He received the Ph.D. degree in Information and Computer Science from the University of California, Irvine (UCI), CA, USA, in 2004, and he was an Assistant Researcher with the Center for Embedded Computer Systems at UCI between 2004 and 2008. His research interests include system-level design automation, system modeling, design languages and methodologies, and embedded hardware and software synthesis. He has co-authored 3 books and over 100 refereed conference and journal publications. His work has received several best paper nominations from, among others, DAC, DATE and HOST, and two best paper awards at DAC'16 and SAMOS'15. He is the recipient of a 2016-2017 Humboldt Research Fellowship. He has been General and Program Chair for conferences such as MEMOCODE and CODES+ISSS, and he currently serves as Associate Editor for ACM TODAES and ACM TECS journals.

Jörg Henkel (M'95-SM'01-F'15) is the Chair Professor for Embedded Systems at Karlsruhe Institute of Technology. Before that he was a research staff member at NEC Laboratories in Princeton, NJ. He received his diploma and Ph.D. (Summa cum laude) from the Technical University of Braunschweig. His research work is focused on co-design for embedded hardware/software systems with respect to power, thermal and reliability aspects. He has received six best paper awards throughout his career from, among others, ICCAD, ESWeek and DATE. For two consecutive terms he served as the Editor-in-Chief for the ACM Transactions on Embedded Computing Systems. He is currently the Editor-in-Chief of the IEEE Design&Test Magazine and is/has been an Associate Editor for major ACM and IEEE Journals. He has led several conferences as a General Chair incl. ICCAD, ESWeek and serves as a Steering Committee chair/member for leading conferences and journals for embedded and cyber-physical systems. Prof. Henkel coordinates the DFG program SPP 1500 "Dependable Embedded Systems" and is a site coordinator of the DFG TR89 collaborative research center on "Invasive Computing". He is the chairman of the IEEE Computer Society, Germany Chapter, and a Fellow of the IEEE.