Article

Run-Time Thermal Management for Lifetime Optimization in Low-Power Designs

Daniele Rossi ^{1,*} and Vasileios Tenentes ^{2,†}¹ Department of Information Engineering, University of Pisa, 56122 Pisa, Italy² Department of Computer Science and Engineering, University of Ioannina, 451 10 Ioannina, Greece; tenentes@cse.uoi.gr

* Correspondence: daniele.rossi1@unipi.it

† These authors contributed equally to this work.

Abstract: In this paper, the magnitude of the temperature and stress variability of dynamic voltage and frequency scaling (DVFS) designs is analyzed, and their impact on the bias temperature instability (BTI) degradation and lifetime of DVFS designs is assessed. For this purpose, a design-time evaluation framework for BTI degradation was developed, which considered the statistical workload and die temperature profiles of DVFS operating modes. The performed analysis showed that, together with high stress variability, DVFS designs exhibited even higher temperature variability, depending on the workload and utilized operating modes, and the impact of temperature variability on lifetime could be up to $2\times$ higher than that due to stress. In order to account for temperature variability on aging detrimental effects, a thermal management run-time system is proposed that honors the desired lifetime constraints by properly selecting temperature constraints that govern the utilized operating modes. The proposed run-time system was applied on the largest benchmark circuit from the IWLS 2005 suite, Ethernet circuit, synthesized with the 32 nm CMOS technology. The proposed system was verified to obtain lifetime and performance estimation and the trade-off with up to 35.8% and 26.3% higher accuracy, respectively, when compared to a system that ignores temperature variability and accounts for average temperature only. The proposed framework can be suitably utilized for tuning run-time throttling policies of low-power designs, thus allowing designers to optimize lifetime–performance trade-offs, depending on the requirements mandated by specific applications and operating environments.

Keywords: bias temperature instability; dynamic voltage and frequency scaling; dynamic thermal management; low-power design; lifetime

Citation: Rossi, D.; Tenentes, V. Run-Time Thermal Management for Lifetime Optimization in Low-Power Designs. *Electronics* **2022**, *11*, 411. <https://doi.org/10.3390/electronics11030411>

Academic Editor: Leonardo Pantoli

Received: 30 December 2021

Accepted: 26 January 2022

Published: 29 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As the technology node shrinks, electronic systems become more prone to aging phenomena, jeopardizing their reliability. In particular, scaling to the 32 nm technology node and below leads to a change in the nature of reliability degradation effects. Indeed, due to aging phenomena, reliability degradation switches from sudden functionality issues to progressive degradation of the electrical characteristics and performance of electronic system components [1–4].

The dominating aging phenomenon for nanometer devices is bias temperature instability (BTI) [2,5,6], whose main effect is to increase the transistor threshold voltage, depending on the operating conditions (voltage, temperature, etc.), technology parameters, and workload [2,7,8]. Moreover, it should be considered that temperature and stress ratio values may vary from gate to gate and even from transistor to transistor in the same gate, thus inducing a great variability on aging degradation. If the resulting performance degradation exceeds circuit time margins, it may lead to circuit failure, thus reducing the lifetime of electronic systems.

Great effort has been devoted to the modeling of BTI effects and to develop techniques to counteract them [6,9–15]. Design strategies adopted to tackle the negative effects of BTI aging and enhance reliability include IC over-designing, larger time slacks, and monitoring of critical paths to counteract these effects at run-time [2,8,13,16]. On the other hand, all these solutions may have a large impact on system performance and escalate the cost of reliability.

In [2], a workload-dependent stress ratio computation framework was presented, which considered structural correlations within a logic circuit. It was shown that different workloads can induce a propagation delay variation up to 11%. Therefore, paths that are not critical at Time 0 may become critical over time. This result highlights that it is difficult to identify critical paths to be monitored at design time in order to detect possible timing issues possibly occurring during the circuit lifetime. Nevertheless, the approaches in [2,16] did not account for BTI aging variability induced by different temperatures at which different gates within a circuit may operate. In this regard, it is worth noting that workloads leading to a similar stress ratio distribution may exhibit different thermal profiles. Indeed, according to the BTI models in [6,7,17], the stress ratio does not depend on the frequency of the input signals, but only on the total amount of time during which a transistor is ON. Instead, the actual switching frequency may play an important role when the temperature effect on aging is accounted for. As is well known, dynamic power dissipation increases linearly with switching frequency, and as a result, the temperature turns out to be very sensitive to the operating frequency. In addition, it should be considered that identical blocks undergoing the same workload that are placed in different areas of a SoC will turn out to have the same stress ratio distribution, but might be characterized by very different thermal profiles. On top of this, it is worth noting that in designs implementing dynamic voltage and frequency scaling (DVFS) techniques to reduce power consumption [18,19], the different operating modes considerably impact the circuit thermal profile, together with electric stress applied to transistors [20–23]. Consequently, the BTI aging of DVFS designs depends critically on the applied voltage and frequency scaling policies [24,25]. As for DVFS designs, they usually operate under the control of a dynamic thermal management (DTM) system, which is responsible for monitoring online the temperature of the circuit using on-chip sensors. According to the temperature value, it then selects its active operating mode in order to honor predefined constraints, such as performance and thermal design power (TDP), which is the maximum amount of heat generated by a component (e.g., CPU, GPU, or system on a chip) that the cooling system is designed to dissipate under any workload. As a result, electric stress and operating temperature, which are the main parameters activating BTI aging [5–7], turn out to be critically dependent on the DTM policies applied to our system. It is therefore expected that predefined thermal constraints used for DTM affect the BTI aging of DVFS designs.

As mentioned earlier, the previous BTI-aware design techniques in [2,16] did not account for the variability of BTI aging induced by different temperatures characterizing the many gates within a critical path. Additionally, the effect of DTM policies was completely overlooked. Therefore, the previous approaches in [2,16] failed to accurately estimate the BTI effect on both the performance and lifetime. These limitations are overcome in the approach proposed in this manuscript. In particular, the aim of this manuscript was twofold:

- To prove for the first time that a fine-grained temperature distribution, together with a fine-grained stress distribution allow a much more accurate evaluation of the BTI aging effects on circuit performance and lifetime;
- To propose a run-time framework that, when applied to low-power designs adopting DVFS, allows designers to tune run-time throttling policies to trade-off lifetime and performance.

First, a detailed discussion on how the proper stress ratio can be evaluated in multi-input CMOS gates is provided. It is shown that the stress condition of a transistor in these gates may depend not only on its input voltage value, but also on the input values

of the other transistors, as well as on the previous applied input patterns. This analysis highlights that the stress ratio for a transistor can be overestimated when considering only its input voltage, neglecting the gate structure and overall input statistics. Then, by means of HSPICE simulation considering a 32 nm high-k, CMOS technology from [26], the impact of temperature-induced BTI variability on circuit lifetime is shown to be higher than that of stress-induced BTI variability. In particular, by considering simple logic gates and three different input signal probabilities P_{IN} (0.25, 0.5, 0.75), with a constant operating temperature ($T = 75^\circ\text{C}$), it is discussed that stress-induced variability on BTI can lead to a lifetime estimation variability exceeding 42% for a four-input NAND gate over the average value computed considering $P_{IN} = 0.5$, yet the induced variability on the propagation delay being very small (2.3% after 10 y of operation). Similarly, temperature-induced variability on BTI aging was assessed. An operating temperature varying from 70°C to 90°C was considered, which can lead to a 95% lifetime estimation variability in the case of two-input and three-input NAND gates over the average value considered before. As a result, temperature-induced variability on aging can lead to a higher lifetime variability than that induced by stress.

To properly account for this variability in lifetime estimation at design time, in this manuscript, a simulation framework for the BTI degradation analysis of DVFS designs accounting for workload and actual thermal profiles is proposed. They were generated considering a statistically probable workload and DTM constraints by means of the HotSpot tool [27]. The proposed approach allowed us to obtain the fine-grained stress ratio for every transistor in a circuit, as well as a fine-grained temperature profile. In particular, for each and every transistor in the considered benchmark, a unique model accounting for the specific stress ratio and operating temperature was produced, and a fine-grained stress ratio and temperature-aware aged library was generated. The developed simulation framework was applied to the Ethernet circuit from the IWLS 2005 benchmark suite [28]. The obtained results showed that, depending on the considered DTM constraints, the margin-based design can underestimate or overestimate the lifetime of DVFS designs by up to 67.8% and 61.9%, respectively. In addition, the proposed framework allows designers to explore the most appropriate DTM constraints according to a trade-off between long-term reliability (lifetime) and performance with up to 35.8% and 26.3% higher accuracy, respectively, against a system that ignores the effects of temperature variability on BTI and uses the average temperature to estimate the impact of BTI aging on circuit performance and lifetime. The proposed framework can be used for tuning run-time throttling policies of low-power designs, thus allowing designers to optimize lifetime–performance trade-offs, depending on the requirements mandated by specific applications and operating environments.

The remainder of this paper is organized as follows. Section 2 gives a background on the causes of BTI along with current strategies to tackle it. In Section 3, through HSPICE simulations, we assess the impact on the long-term reliability (lifetime) of stress-induced and temperature-induced BTI variability. Section 4 presents the proposed simulation framework. In Section 5, we then provide simulation results and discuss how the proposed simulation framework can be used to trade-off circuit lifetime and performance. Section 6 concludes the paper.

2. Background

Bias temperature instability (BTI) degradation originates from the creation of charges at the Si–dielectric interface. During the stress phase, the Si–H bonds at the Si–dielectric interface break. The broken bonds act as interface traps, while the released hydrogen, in the form of both atoms (H) and molecules (H_2), diffuses toward the gate [5]. During the stress phase (transistor ON), the concentration of interface traps increases, which leads to an increase in the transistor threshold voltage V_{th} [5]. During the recovery phase (transistor OFF), the hydrogen diffuses back and recombines with the Si dangling bonds, annealing

them [5]. As a result, V_{th} decreases back towards its initial value. However, the recovery is only partial, and so, a net increase in V_{th} is experienced by transistors over time.

BTI has been extensively modeled using a number of methods, one of which is the reaction diffusion model [6]. This allows the threshold voltage increase of a transistor to be estimated as a function of technology parameters and operating conditions. Negative BTI (NBTI) is observed in pMOS transistors, and it usually dominates over the positive BTI (PBTI) observed in nMOS transistors [6,7]. In [7,29], an analytical model was proposed that allows designers to estimate long-term, worst-case threshold voltage degradation. It is:

$$\Delta V_{th} = \chi K \sqrt{C_{ox}(V_{gs} - V_{th0})} e^{-\frac{E_a}{kT}} (\alpha t)^{1/6}. \quad (1)$$

The parameter C_{ox} is the oxide capacitance, t the operating time, k the Boltzmann constant, T the device temperature, and E_a a fitting parameter ($E_a \simeq 0.08$ eV [7]). The parameter K lumps technology and environmental parameters and has been estimated to be $K \simeq 2.7 \text{ V}^{1/2} \text{ F}^{-1/2} \text{ s}^{-1/6}$ by fitting (1) with the experimental results reported in [30]. The coefficient χ equals 0.5 for PBTI and one for NBTI. Finally, stress ratio α is the fraction of the operating time during which a MOS transistor is ON (under stress). It is $0 \leq \alpha \leq 1$, with $\alpha = 0$ if the transistor is always OFF (recovery phase), while $\alpha = 1$ if it is always ON (stress phase). This value depends on input statistics (workload) and logic gate structure, as will be clarified in Section 3.

3. Analysis of BTI Aging Variability

In this section, the impact of BTI variability on propagation delay and lifetime is assessed by considering different stress ratios and operating temperatures. It is highlighted that the two most influential parameters of BTI degradation and its variability, namely stress and operating temperature, depend mostly on the workload of the circuit and, as for the temperature, also on the device's location in the layout of the design. Moreover, since designs with dynamic voltage and frequency scaling (DVFS) are controlled by a dynamic thermal management (DTM) system, the policies followed by the DTM system strongly influence the power consumption of the DVFS design, inducing a temperature variability that should be considered for an accurate BTI aging estimation.

3.1. Stress Tables for Logic Gates Using Input Probabilities

To accurately estimate aging during the timing analysis of a design, only the time each transistor is under stress should be accounted for. Since the workload may not be known at the design phase, signal probabilities need to be considered, which are strongly influenced by the structural correlations of a logic design [2]. To clarify this aspect, let us consider a simple NOT gate. The dependency of the stress condition on the signal probability is straightforward for an inverter. Denoting by P_{IN} the probability of the input IN to be at logic one value, the values of the stress ratios for its composing nMOS and pMOS transistors are: $\alpha_n^{not} = P_{IN}$ and $\alpha_p^{not} = 1 - P_{IN}$.

In the case of a multiple-input gate, a more accurate analysis is required. Indeed, as highlighted in (1), the BTI threshold voltage degradation experienced by a transistor depends on the difference between its gate and source voltage V_{GS} . For parallel transistors connected between the power supply (either V_{dd} or ground) and the output, straightforward considerations similar to those for the NOT gate apply. Instead, in the series transistors of a multiple-input gate (pull-down nMOS network in NAND gates and pull-up pMOS network in NOR gates), the voltage at the source node may depend on the status of the other transistors in the series connection. Therefore, as introduced in [31], the stress of each transistor of a multiple-input logic gate depends not only on its input voltage (0 V or V_{dd}), but also on the status of the other transistors of the logic gate.

As an example, consider a two-input NAND gate and denote by MN1 the nMOS transistor whose drain is connected to the output node and by MN2 the nMOS transistor whose source is connected to ground. When both MN1 and MN2 are ON ($IN1 = IN2 = 1$),

it is $V_{GS1} = V_{GS2} = V_{dd}$. Therefore, according to (1), both transistors are under stress. If now MN2 is turned off ($V_{GS2} = 0$), it undergoes a recovery phase. Moreover, since the source parasitic capacitance of MN1 is charged up to $V_{dd} - V_{th}$, thus resulting in $V_{GS1} = V_{th}$, also transistor MN1 undergoes a recovery phase, although its input is equal to logic one. Analogous considerations hold true for a series of pMOS transistors of the pull-up network in a two-input NOR gate, where MP1 is the transistor connected to the output and MP2 to V_{dd} .

The stress (s) and recovery (r) status of all transistors composing two-in NAND and NOR gates are reported in Table 1, referred to as the *stress table*, for all input combinations. The last row of the stress table shows the average stress ratio for all transistors, considering the input patterns as equally likely, which therefore present a signal probability $P_{IN1} = P_{IN2} = 0.5$.

Table 1. Stress/recovery condition for 2-in NAND and NOR gates as a function of the input patterns.

IN1	IN2	2-IN NAND				2-IN NOR			
		MP1	MP2	MN1	MN2	MP1	MP2	MN1	MN2
0	0	s	s	r	r	s	s	r	r
0	1	s	r	r	s	s	r	r	s
1	0	r	s	r	r	r	r	s	r
1	1	r	r	s	s	r	r	s	s
Stress ratio α		0.5	0.5	0.25	0.5	0.5	0.25	0.5	0.5

The values of the stress ratio $\alpha = \frac{t_s}{LT}$, where t_s is the time during which a transistor is under stress and LT is the circuit lifetime, can be generalized as a function of the input probability P_i ($i = 1, 2$), as reported in Table 2. Hence, in a NOR gate, for the upper transistor MP1 to be stressed, its input has to be logic zero, whereas for the lower transistor MP2, both inputs have to be zero. Analogous considerations hold true for a two-input NAND gate.

Table 2. Stress table for 2-in basic gates as a function of the input signal probabilities.

Stress Ratio	2-IN NAND	2-IN NOR
α_{n1}	$P_1 P_2$	P_1
α_{n2}	P_2	P_2
α_{p1}	$1 - P_1$	$(1 - P_1)(1 - P_2)$
α_{p2}	$1 - P_2$	$1 - P_2$

This analysis is extended to three- and four-input gates. In this case, since the source node of some of the series transistors are internal nodes and not connected to either the ground (nMOS) or V_{dd} (pMOS), their stress conditions depend on the voltage at the internal nodes, which in turn may depend on the inputs applied during the previous clock cycle. In Table 3, the expression of the transistor stress ratio for three-input basic gates at clock cycle i is reported.

Table 3. Stress table for 3-in basic gates at clock cycle i as a function of the input signal probabilities.

Stress Ratio	3-IN NAND	3-IN NOR
α_{n1}	$P_1 P_2 P_3$	P_1
α_{n2}	$P_2 P_3 + P_2(1 - P_1)(1 - P_3)P_{3,(i-1)}$	P_2
α_{n3}	P_3	P_3
α_{p1}	$1 - P_1$	$(1 - P_1)(1 - P_2)(1 - P_3)$
α_{p2}	$1 - P_2$	$(1 - P_2)(1 - P_3) + P_1 P_3(1 - P_2)(1 - P_{3,(i-1)})$
α_{p3}	$1 - P_3$	$1 - P_3$

As can be seen, all parallel-connected transistors (pMOS for the NAND gate and nMOS for the NOR gate) exhibit the same average stress ratio α equal to $1 - P_i$ for the pMOS and P_i for the nMOS transistors. Instead, the stress ratios of series transistors strongly differ. Let us identify the series transistors considering their distance from the output node as follows: transistors n1 (NAND) and p1 (NOR) are connected to the gate output, whereas transistors n3 (NAND) and p3 (NOR) are connected to ground and V_{dd} , respectively. Moreover, the stress ratios of transistors n3 (NAND) and p3 (NOR) depend on the respective input probability only, since the value of V_{gs} depends merely on the respective input voltage. Differently, transistors n1 (NAND) and p1 (NOR) turn out to be under stress only when all the series transistors are on. Let us clarify this point by analyzing the nMOS series transistors in a 3-in NAND gate, as depicted in Figure 1. When all three inputs are at a high logic value (Figure 1a), the source node of transistor n1 is completely discharged to ground ($V_{s1} = 0$), and as a result, transistor n1 is under stress ($V_{gs1} = V_{dd}$). On the other hand, when at least one of the inputs IN2 and IN3 is at low logic value (Figure 1b), the n1 source is charged up to $V_{s1} = V_{dd} - V_{tn}$. As a result, it is $V_{gs1} = V_{tn}$, and according to (1), transistor n1 does not experience any stress condition, yet has its input at a high logic value. Similar considerations hold for the pMOS transistor p1 in a 3-in NOR gate.

Figure 1. A 3-in NAND gate: (a) transistor n1 is under stress; (b) transistor n1 is not under stress.

Another interesting consideration can be drawn for transistors n2 (NAND) and p2 (NOR), for input configurations $IN1\ IN2\ IN3 = 0\ 1\ 0$ and $1\ 0\ 1$, respectively. In this case, in fact, the source nodes of the two transistors are in a high-impedance state, and their values depend on the input configuration at previous clock cycle ($i - 1$), as shown in the expressions of α_{n2} (NAND) and α_{p2} (NOR) in Table 3. Therefore, the nMOS transistor n2 (pMOS transistor p2) turns out to be under stress only if the node was discharged (charged) during the previous clock cycle. This analysis can be easily extended to four-input gates. It is worth noting that leakage affecting the voltage values of different nodes was not considered in the performed analysis, as this goes beyond the purpose of identifying stress conditions as a function of input probabilities.

3.2. Electrical and Thermal Simulation Flows and Setup

In order to account for the effective electric stress applied to a transistor, the *effective stress ratio*, denoted by α_{eff} , can be defined as follows:

$$\alpha_{eff} = \alpha^{\frac{1}{6}} \sqrt{\frac{V_{gs} - V_{th0}}{V_{dd} - V_{th0}}} \quad (2)$$

Consequently, (1) can be re-written as follows:

$$\Delta V_{th} = K' \alpha_{eff} e^{-\frac{E_a}{kT}} t^{1/6}, \quad (3)$$

where α_{eff} accounts for the BTI dependency on the workload (input statistics), the exponential term accounts for the operating temperature, and the constant $K' = \chi K \sqrt{C_{ox}(V_{dd} - V_{th0})}$ lumps all technology parameters and the operating voltage.

Figure 2 depicts the developed flow for evaluating the impact of the logic gate input signal probabilities on the stress ratios of their transistors, considering also the operating conditions (temperature and voltage) and lifetime. The obtained data for each gate were utilized to generate an “aged” library, which was then used to simulate complex circuits with all transistors mapped to the proper aging. The details are discussed in Section 4.

Figure 2. Flow for workload and operating condition-aware gate model characterization.

The impact of temperature variations on aging and the evaluation of the effect of different DVFS operating modes on the system thermal profile were assessed by devising a simulation flow that takes data about the physical synthesis of a circuit and its input statistics as the input and then generates the power analysis of each considered DVFS mode as the output. This information, together with the circuit layout feed the HotSpot tool [27], which performs a thermal analysis of each operating mode. The block diagram of the developed flow is shown in Figure 3.

Figure 3. Flow for obtaining thermal maps after physical synthesis.

By exploiting the proposed simulation flows to evaluate the electric stress and thermal profile of a circuit, designers can accurately estimate the aging degradation of a circuit and assess its impact on the lifetime and reliability, as well as explore the reliability and performance trade-offs by selecting different DVFS operating modes. A 32 nm high-k metal-gate, CMOS technology from [26] was considered for analysis and validation throughout the paper.

3.3. Stress-Induced BTI Variability

This subsection first discusses the stress-induced BTI variability generated by different signal probabilities reflected on the propagation delay of basic logic gates, thus logic paths, and their LT variability. In the case of a NOT gate, results are shown in Figure 4a,b for

normalized propagation delay and LT , respectively. The normalization factor is the delay exhibited by a logic gate at Time 0 (t_0).

Propagation delay was evaluated as the time interval elapsing between the instant at which the NOT input voltage experiences 50% of its full excursion, which is equal to $V_{dd}/2$, and the instant at which the NOT output voltage performs the correspondent 50% variation, which is still equal to $V_{dd}/2$. Denoting by V_{in} and V_{out} the input and output voltage of the NOT gate, respectively, the expression of propagation delay Pd can be formalized as follows:

$$Pd = t(V_{out} = V_{dd}/2) - t(V_{in} = V_{dd}/2). \quad (4)$$

It should be considered that all gates have been designed to be symmetric at t_0 ; therefore, either a $0 \rightarrow 1$ or $1 \rightarrow 0$ transition can be considered to evaluate the propagation delay for a fresh NOT. Instead, for aged gates, the propagation delay depends on the threshold voltage degradation, and thus on the input statistics. In this case, the propagation delay was measured in the worst-case conditions.

As for lifetime LT , it was evaluated as the time interval required for the propagation delay to degrade by 15% over the value at t_0 [32]. In doing so, it was assumed that the clock period was determined by considering the worst-case propagation delay increased by 15% in order to account for possible effects due to PVT variations. If the propagation delay exceeds this margin, then an incorrect signal propagation takes place, thus possibly causing a system failure. Therefore, LT at a generic operating time t is evaluated as:

$$LT = t - t_0 : Pd(t) = 1.15 \cdot Pd(t_0). \quad (5)$$

Figure 4, which was obtained considering input signal probability values $P_{IN} = 0.25, 0.5$, and 0.75 , clearly shows that, although the delay variation is very small, the LT variation exceeds two years.

Figure 4. NOT propagation delay with BTI aging: (a) delay variation with different input signal probabilities; (b) propagation delay and lifetime variability.

Figure 5 depicts the simulation results for the normalized propagation delay of basic logic gates NAND and NOR, with 2–4 inputs, for different values of the input probabilities. Namely, the two extreme cases $P_{in} = 0.25$ and $P_{in} = 0.75$, together with the average case $P_{in} = 0.5$, $\forall i = 2, 3, 4$ were considered for the input probabilities. For all simulated gates and input probabilities, the delay degradation after only 6 mo of operation exceeds 50% of the overall degradation in 10 y of operation. It is interesting to note that delay degradation is dominated by pMOS transistors' aging ($P_{in} = 0.25$) and increases with the number of inputs for the NAND gate, whereas this decreases for the NOR gate. Indeed, the stress probability of series pMOS transistors in NOR gates diminishes noticeably with the increase in the number of inputs, as highlighted by the stress ratio expressions reported in Tables 2 and 3. This consideration does not apply to the parallel pMOS transistors in the NAND gates.

Another important characteristic that is worth highlighting is the generally small variability of the delay degradation for different input probabilities and the noticeable

difference between the degradation trend for the NAND and NOR gates. In order to better assess this variability, the following two metrics can be defined:

$$var_{1.S} = \frac{\Delta delay|_{max.S}}{delay_{t0}}; \quad var_{2.S} = \frac{\Delta delay|_{max.S}}{delay(P_{in} = 0.5)}, \quad (6)$$

where $\Delta delay|_{max.S}$ is the maximum difference of the propagation delay for different input probabilities and is given by $\Delta delay|_{max.S} = delay(P_{in} = 0.25) - delay(P_{in} = 0.75)$. Therefore, $var_{1.S}$ represents the variability of the propagation delay for different input probabilities against the delay exhibited by the considered gate at t_0 ; $var_{2.S}$ is instead the variability of the propagation delay for different input probabilities against the average case ($P_{in} = 0.5$). Values for the normalized propagation delays and defined variability metrics are reported in the first set of columns of Table 4.

Figure 5. Normalized propagation delay for different input signal probabilities for basic gates (operating temperature $T = 75^\circ\text{C}$; NF: normalization factor): (a) 2-in NAND ($NF = 8.08$ ps); (b) 3-in NAND ($NF = 11.98$ ps); (c) 4-in NAND ($NF = 16.62$ ps); (d) 2-in NOR ($NF = 11.29$ ps); (e) 3-in NOR ($NF = 19.93$ ps); (f) 4-in NOR ($NF = 30.17$ ps).

Table 4. Input probability induced variability in the propagation delay and lifetime.

Logic Gate	Prop. Delay					Lifetime			
		0.25	0.5	0.75	$var_{1.S}$ (%)	$var_{2.S}$ (%)	0.25	0.5	0.75
NOT		1.179	1.177	1.165	1.40	1.19	4.3	4.4	6.5
NAND	2	1.156	1.149	1.142	1.38	1.20	8.3	>10	>10
	3	1.160	1.151	1.141	1.92	1.67	7.2	>10	>10
	4	1.165	1.154	1.142	2.27	1.97	6.2	8.9	>10
NOR	2	1.142	1.136	1.130	1.27	1.12	>10	>10	>10
	3	1.124	1.129	1.125	0.85	0.76	>10	>10	>10
	4	1.106	1.107	1.105	0.07	0.06	>10	>10	>10

As can be seen, the variability is very limited for all basic gates, with a NOR gate exhibiting a variability that is sensibly lower than for a NAND gate for all number of inputs. In the case of a NAND gate, $var_{1.S}$ ranges from 0.86% at 1 y to 1.38% at 10 y for a 2-in gate and from 1.39% at 1 y to 2.27 at 10 y for a 4-in gate; as for a NOR gate, $var_{1.S}$ is 0.82% at 1 y and 1.27% at 10 y for a 2-in gate and less than 0.07% for all lifetime for a 4-in gate. Similar considerations apply to $var_{2.S}$.

Although the delay variability with input probability is very small, the LT variation can be considerably larger. Values for LT and the corresponding variability are reported in the last four columns of Table 4. The variability metric for LT was evaluated as:

$$var_{3,S} = \frac{\min(LT_{0.75}, 10\text{years}) - \min(LT_{0.25}, 10\text{years})}{\min(LT_{0.5}, 10\text{years})}. \quad (7)$$

For NAND gates, $var_{3,S}$ ranges from 17% (two inputs) to 42.7% (four inputs), thus exhibiting a much larger variability than the propagation delay. For NOR gates, the variability evaluation turns out to be less meaningful. Indeed, a maximum value for LT equal to 10 y was considered, after which the design is no longer operational. As a result, LT values exceeding 10 y and the corresponding variability were not evaluated.

3.4. Temperature-Induced BTI Variability

So far, BTI and propagation delay variability as a function of workload-induced stress variability have been accounted for in aging simulation frameworks [2]. However, workload impacts BTI variability not only due to (electric) stress, but also due to temperature. Indeed, different workloads induce different switching activities, hence distinct power dissipation values in different blocks/paths in the considered design. Consequently, a considerable temperature difference may be exhibited by the blocks/paths in a design, which may exceed 20 °C [27]. In addition, the operating frequency and voltage, which do not affect the stress ratio [6], are the main players in determining the power consumption, thus the thermal profile of a design. Therefore, in a DVFS design, different operating modes are characterized by different thermal profiles and, consequently, by a different BTI aging. It should be noted that the operating modes of a DVFS design are usually controlled by a DTM system, which is responsible for monitoring online the circuit temperature through on-chip sensors. Moreover, it selects appropriately its operating mode (operating frequency and power supply) in order to honor pre-defined constraints, such as temperature, which is proportional to the power dissipation, and performance.

As an example, the Ethernet circuit from the IWLS 2005 benchmark suite was synthesized, which implements a DVFS technique with two operating modes, referred to as *low performance* (LP = 0.5 GHz@0.7 V) and *high performance* (HP = 2 GHz@1 V). Although the circuit is equipped with a DTM system selecting the proper DVFS operating mode at run-time, a fixed operating mode (either LP or HP) was assumed to better emphasize the impact of DVFS on the circuit thermal profile, and thus on BTI aging. The steady-state thermal analysis results for both LP and HP modes are shown in Figure 6a,b, respectively.

Figure 6. Thermal maps for (a) LP operating mode and (b) HP operating mode.

For LP mode, the maximum temperature T_{max} (hotspot) is approximately 72 °C, while T_{max} reaches 151 °C for HP mode. In particular, the hotspot is experienced in the upper area of the circuit in HP operating mode, due to the higher dynamic power dissipation of

the random logic implementing the Rx and Tx cores that are located in that area, compared to the memory in the lower part of the circuit. On the other hand, in LP mode, the hotspot is localized in the lower part, since in this case, the leakage power of the memory prevails over the dynamic power of the random logic.

If the DTM policies are ignored during the lifetime estimation of the circuit, then either the temperature during LP or that during HP will be used. In the first case, the lifetime estimation may be very optimistic, whereas in the second case, very pessimistic. However, a circuit is meant to operate under the influence of the DTM system, whose policies determine the temperature variability, which should be properly considered for lifetime estimation during circuit design.

To better assess the impact of temperature on aging and, as a consequence, on propagation delay and lifetime, the same basic gates as in Section 3.1 were considered. As an example, assume that the operating temperature is bounded in the interval $[50, 100]$ °C by the DTM system. In Figure 7, the temperature-induced BTI variability that reflects on the Pd and LT variability is pictured for a NOT gate (Figure 7a) and for two-input NAND and NOR gates (Figure 7b,c, respectively). As for the temperature, the upper and lower bounds were considered, as well as the average case (75 °C). The signal probability was set to 0.5 in all cases. Detailed results for the Ethernet benchmark are presented in Section 5. As can be seen, the effect of the temperature-induced BTI variability on propagation delay is higher than the stress-induced variability.

Figure 7. Normalized propagation delay for different aging temperatures for basic gates ($\alpha = 0.5$; NF : normalizing factor): (a) NOT ($NF = 5.00$ ps); (b) 2-input NAND ($NF = 8.08$ ps); (c) 2-input NOR ($NF = 11.29$ ps).

In Table 5, temperature-induced variability values for three- and four-input basic logic gates are reported. As for propagation delay, variabilities $var_{1,T}$ and $var_{2,T}$ were evaluated as follows:

$$var_{1,T} = \frac{\Delta delay|_{max,T}}{delay_{t0}}; \quad var_{2,T} = \frac{\Delta delay|_{max,T}}{delay(T_A = 75^\circ\text{C})}, \quad (8)$$

where $\Delta delay_{max,T}$ is the maximum propagation delay difference induced by different aging temperatures T_A and given by $\Delta delay|_{Tmax} = delay(T_A = 100^\circ\text{C}) - delay(T_A = 50^\circ\text{C})$. Variability $var_{1,T}$ is calculated over the delay of a fresh device, whereas variability $var_{2,T}$ is evaluated over the value of the propagation delay at $T_A = 75^\circ\text{C}$. Instead, lifetime variability is:

$$var_{3,T} = \frac{\min(LT_{50^\circ\text{C}}, 10\text{ y}) - \min(LT_{100^\circ\text{C}}, 10\text{ y})}{\min(LT_{75^\circ\text{C}}, 10\text{ y})}. \quad (9)$$

Table 5. Temperature-induced variability in propagation delay and LT.

Logic Gate		Prop. Delay					Lifetime			
		50°	Norm 75°	100°	$var_{1,T}$ (%)	$var_{2,T}$ (%)	50°	Abs. 75°	100°	$var_{3,T}$ (%)
NOT		1.138	1.182	1.225	8.66	7.33	>10	3.6	1.3	129.9
NAND	2	1.129	1.161	1.193	6.35	5.47	>10	6.7	2.4	113.4
	3	1.130	1.163	1.196	6.59	5.67	>10	6.4	2.2	121.9
	4	1.132	1.166	1.200	6.80	5.83	>10	5.7	1.9	142.1
NOR	2	1.119	1.147	1.176	5.71	4.98	>10	>10	3.9	61.0
	3	1.117	1.142	1.167	4.97	4.35	>10	>10	5.5	45.0
	4	1.108	1.129	1.151	4.28	3.79	>10	>10	8.8	12.0
Average var		-	-	-	6.19	5.35	-	-	-	89.3

Comparing the results in Figure 7 and Table 5 to those in Figures 4 and 5, it can be noticed that the effect of temperature-induced BTI variability exceeds the stress-induced one for both propagation delay and lifetime.

As conclusive remarks for the performed analyses, it can be observed that during the BTI-aware timing analysis of a DVFS design, which is crucial for evaluating its performance degradation and the expected lifetime, the contribution of temperature variability can be considerably higher than the contribution of stress variability. Therefore, both the stress and temperature variability induced by the workload should be considered during BTI-aware timing analysis. Nevertheless, if the workload is not known and the average case for input probability is considered, the error in propagation delay degradation due to aging is negligible, even though the corresponding error in lifetime estimation can be larger than 10%. Moreover, different operating modes exhibit very different thermal profiles, which implies that thermal management constraints should also be considered for the evaluation of temperature-induced BTI variability and its impact on lifetime and performance.

4. Proposed BTI Simulation Framework for Run-Time Thermal Management

In this section, we discuss the proposed simulation framework (Figure 8) and the developed run-time system. The framework performs the BTI degradation analysis of DVFS designs accounting for the workload and actual thermal profiles considering statistically probable workload utilizing the HotSpot tool [27]. It was then used to explore the DTM constraints of the run-time aging-aware thermal management system.

Figure 8. Proposed framework for BTI degradation analysis of DVFS designs.

Given an RTL netlist of the DVFS design, the signal probabilities of the logic nets are computed accounting for possible signal dependencies. These dependencies may have different origins: they may occur because of topological correlations, signal split up and reconvergence, and data-dependent correlations due to signal correlations at the circuit inputs [2,31]. For the computation of BTI-induced threshold voltage degradation,

the aging model presented in Section 2 was utilized. Granular values of input signal probabilities, operating conditions (temperature T and operating voltage V_{dd}), and time t were considered. Then, static and dynamic power analysis was performed accounting for a probabilistic workload induced by the signal probabilities. The power analysis results were used for a DTM-aware thermal analysis in order to generate the temperature maps. Specifically, the HotSpot tool was executed using DTM constraints T_c and T_h , representing the lowest and the highest possible temperature at the hotspot of the circuit. A temperature sensor was considered to be monitoring the temperature at the hotspot of the circuit and feeding this information to the DTM controller. The operation of the circuit under such DTM constraints is described by means of the following example.

Example 1. Consider the graph shown in Figure 9, which refers to a hypothetical circuit with two operating modes, the high performance (HP) and the low performance (LP), with frequencies $f_{HP} > f_{LP}$. The operation of this circuit is controlled by a DTM system, which maximizes performance honoring pre-defined temperature constraints. While the circuit operates using f_{HP} , it heats up. When the temperature of the circuit reaches a pre-defined highest allowed temperature constraint T_h , the DTM system forces the circuit to switch to LP mode with frequency f_{LP} , which causes the circuit to cool down. When the temperature drops to a pre-defined lowest temperature constraint T_c , then the DTM system activates the HP mode again in order to maximize performance. This way, the circuit temperature is always in the admitted temperature window $w = T_h - T_c$. The actual duration of each heat-up/cool-down time frame in this loop depends on the combination of the workload-induced switching activity and temperature of the circuit.

Figure 9. DVFS design operating under DTM constraints.

From this process, a fine-grained temperature profile that considers the temperature constraints followed by the DTM system was generated and utilized for obtaining the temperature for each logic gate in the design. This information, together with the signal probabilities of the logic gates were utilized for mapping each logic gate in the design with aged models from the aged gate library (Figure 2). Finally, timing analysis was performed for the considered circuit in order to obtain a subset of the longest paths, thus reducing the size of the mapped SPICE netlist during subsequent simulations.

5. Simulations and Results

The developed simulation framework was applied to analyze the BTI degradation of the largest benchmark from the IWLS'05 suite [28], the Ethernet benchmark. The synthesis of the benchmark was conducted with a 32 nm high- k metal gate CMOS technology [26] with DVFS using two operating modes, referred to as *low performance* ($LP = 0.5 \text{ GHz}@0.7\text{V}$) and *high performance* ($HP = 2 \text{ GHz}@1\text{V}$), as introduced in Section 3. Finally, based on the results for various dynamic thermal management (DTM) constraints, the appropriate constraints were selected, which met either the lifetime or performance requirements. For the evaluation of the performance, the results of the thermal analysis regarding the utilization of each operating mode were used. Consider again the example presented in Section 4 (Figure 9). Once the thermal analysis had been conducted, the time that the circuit spent in either the HP or the LP operating mode, t_{HP} and t_{LP} (shown in Figure 9),

respectively, was obtained. Then, the expected long-term performance of the circuit was evaluated with the *effective operating frequency* f_{eff} as:

$$f_{eff} = \frac{t_{LP}}{t_{LP} + t_{HP}} \cdot f_{LP} + \frac{t_{HP}}{t_{LP} + t_{HP}} \cdot f_{HP}. \quad (10)$$

In Figure 10, the average temperature of the Ethernet circuit is presented. The considered dynamic thermal management (DTM) constraints were $T_c = 80^\circ\text{C}$ and $T_h = 100^\circ\text{C}$. Note that for this case, the temperature window w (Figure 9) was $w = T_h - T_c = 20^\circ\text{C}$. The average temperature of the circuit at the hotspot was slightly higher than 90°C , while the average temperature of all the gates of the longest path was 88.1°C .

The propagation delay of the longest path for the considered temperatures is shown in Figure 11. In particular, in Figure 11a, propagation delay trend over time is depicted for the following operating temperatures: low DTM constraint $T_c = 80^\circ\text{C}$; high DTM constraint $T_h = 100^\circ\text{C}$; average temperature of the DTM constraints $TA = (T_c + T_h)/2 = 90^\circ\text{C}$; fine-grained temperature fgT , whose distribution throughout the circuit is shown in the thermal map in Figure 10. Figure 11b shows a zoom-in to highlight the region where the propagation delay curves cross the guard band considered to estimate the circuit lifetime LT , which was set equal to $1.2 \cdot Pd(t_0)$ to have a long enough lifetime for all considered temperatures. A margin-based temperature selection for aging evaluation, either using the DTM constraints T_c or T_h , resulted in a lifetime estimation of $LT_{T_c} = 4.41$ y and $LT_{T_h} = 2.01$ y, respectively. If the average temperature TA of the DTM constraints was instead considered, the lifetime estimation would be $LT_{TA} = 3.1$ y, which was expected to be more accurate than the estimation based on margin temperature selection. Finally, when all gates in the longest path were mapped with the fine-grained temperature fgT using the proposed framework, and thus the proper temperature was assigned to each gate, the lifetime estimation was $LT_{fgT} = 3.25$ y. As a result, an optimistic evaluation using the T_c temperature underestimated the detrimental effect of BTI aging on the lifetime of the circuit, which turned out to be overestimated by 35.8% compared with the lifetime value obtained by using fine-grained temperature mapping, LT_{fgT} . The estimation was instead pessimistic when using the T_h temperature, since it led to overestimating the actual effect of BTI on the circuit lifetime, thus underestimating the lifetime by 38.2% when compared with LT_{fgT} . Even when the average temperature $TA = 90^\circ\text{C}$ was considered, the BTI effect on the lifetime of the circuit was overestimated by 4.8%.

Figure 10. Temperature profile for DTM constraints $T_c = 80^\circ\text{C}$, $T_h = 100^\circ\text{C}$.

Figure 11. Ethernet longest path delay for various DTM constraints: (a) $T_c = 80\text{ }^{\circ}\text{C}$, $T_h = 100\text{ }^{\circ}\text{C}$; (b) $T_c = 80\text{ }^{\circ}\text{C}$, $T_h = 100\text{ }^{\circ}\text{C}$ zoom-in; (c) $T_c = 70\text{ }^{\circ}\text{C}$, $T_h = 110\text{ }^{\circ}\text{C}$ (d) $T_c = 70\text{ }^{\circ}\text{C}$, $T_h = 110\text{ }^{\circ}\text{C}$ zoom-in.

Next, we show that the deviation between the lifetime LT_{T_c} , LT_{T_h} and LT_{TA} estimated considering DTM constraints T_c , T_h and average temperature TA , respectively, against the lifetime obtained with the proposed framework LT_{fgT} depends on the window size $w = T_h - T_c$. In Figure 11c,d, the trend over time of the longest path propagation delay is depicted for $T_c = 70\text{ }^{\circ}\text{C}$ and $T_h = 110\text{ }^{\circ}\text{C}$ ($w = 40\text{ }^{\circ}\text{C}$). From Figure 11d, the following values for the lifetime were derived: $LT_{T_c} = 6.8\text{ y}$, $LT_{T_h} = 1.35\text{ y}$, $LT_{TA} = 3.1\text{ y}$, and $LT_{fgT} = 4.2\text{ y}$. Therefore, an optimistic evaluation using the T_c temperature underestimated the detrimental effect of BTI on the lifetime of the circuit by 61.9% and a pessimistic one using the T_h temperature overestimated it by 67.8%. Additionally, when the average temperature $TA = 90\text{ }^{\circ}\text{C}$ between the two marginal constraints was considered, the BTI effect on the lifetime of the circuit was overestimated by 26.1%.

In Table 6, results on the performance and lifetime obtained from the proposed fine-grained approach for DTM constraints with $w \leq 20\text{ }^{\circ}\text{C}$ ($w = [2, 10, 16, 18, 20]\text{ }^{\circ}\text{C}$) are reported. The first set of columns present results obtained considering T_c and T_h such that $TA = 80\text{ }^{\circ}\text{C}$, while the results in the second set of columns were obtained with $TA = 130\text{ }^{\circ}\text{C}$. For $TA = 80\text{ }^{\circ}\text{C}$, it can be observed that the LT of the circuit increases with the increase of w . This was attributed to a performance reduction that was also observed while w increased. The reason for these trends was that the selected average temperature ($80\text{ }^{\circ}\text{C}$) caused a higher utilization of the LP operating mode. Since the sensor is located where the hotspot of the circuit is, it overestimated the average temperature of the circuit and forced an even higher utilization of LP operating mode than what would be necessary to meet the desired temperature constraints. For the higher average temperature ($130\text{ }^{\circ}\text{C}$), the exact opposite trend was observed. LT dropped as w increased and the performance increased, which was attributed to the already very high utilization of HP mode at those temperatures.

Table 6. Lifetime and performance trade-off of the constraints ($w < 20\text{ }^{\circ}\text{C}$).

w	DTM Constraints @ $80\text{ }^{\circ}\text{C}$			DTM Constraints @ $130\text{ }^{\circ}\text{C}$		
	$[T_c, T_h]$	LT	Perf	$[T_c, T_h]$	LT	Perf
2	79, 81	4.52	0.88	129, 131	0.61	1.57
4	78, 82	4.51	0.71	128, 132	0.61	1.57
6	77, 83	4.55	0.73	127, 133	0.61	1.61
8	76, 84	4.57	0.71	126, 134	0.61	1.60
10	75, 85	4.63	0.70	125, 135	0.61	1.59
12	74, 86	4.70	4.35	124, 136	0.60	1.64
14	73, 87	4.76	5.62	123, 137	0.60	1.63
16	72, 88	4.91	0.67	122, 138	0.60	1.62
18	71, 89	5.08	0.65	121, 139	0.59	1.63
20	70, 90	5.38	0.62	120, 140	0.59	1.63

Figure 12 shows the results enabling design exploration in terms of *performance* (Figure 12a) and *lifetime* (Figure 12b), as a function of temperature constraints $[T_c, T_h]$, for all possible temperature couples in the range $[70\text{ }^{\circ}\text{C}–150\text{ }^{\circ}\text{C}]$, with $T_h \geq T_c + 10\text{ }^{\circ}\text{C}$, that is the hot temperature T_h was set to be at least $10\text{ }^{\circ}\text{C}$ higher than the cold temperature T_c .

The lifetime (left “y”-axis) and performance (right “y”-axis) results evaluated by using the developed fine-grained framework are depicted in Figure 13a (solid lines) for a temperature window w (“x”-axis) that gradually increases in size in the range $[70\text{ }^{\circ}\text{C}–150\text{ }^{\circ}\text{C}]$. It also depicts the estimated lifetime and performance when the average temperature of the marginal DTM constraints TA is considered (dashed lines labeled as “LT@TA” and “Perf@TA”). Additionally, Figure 13b shows the estimation errors for the examined temperature constraints if the average temperature TA is considered and compared against the proposed approach based on a fine-grained temperature map Tfg . Lifetime was evaluated as discussed in Section 3.3, and the lifetime error is calculated as follows:

$$LT_{err\%} = \frac{LT_{TA} - LT_{Tfg}}{LT_{Tfg}} \times 100 \quad (11)$$

An analogous expression was used to calculate performance error. It is worth noting that at lower temperatures and while w increased, the underestimation of the lifetime using the average temperature TA also increased, reaching 35.8% for $w \geq 40\text{ }^{\circ}\text{C}$. At higher temperatures, the lifetime can be overestimated by more than 10% for $w \geq 40\text{ }^{\circ}\text{C}$. Similarly, Figure 13c depicts the error in the performance estimation. At lower temperatures and while w increased, the expected performance was overestimated by more than 20% for $w \geq 30\text{ }^{\circ}\text{C}$, reaching 26.3% for $w = 50\text{ }^{\circ}\text{C}$. At very high temperatures, the performance was slightly underestimated, with a difference lower than 1.2% over the fine-grained evaluation.

Figure 12. DTM constraints' exploration: (a) performance; (b) lifetime.

Figure 13. (a) Lifetime (left “y”-axis) against performance (right “y”-axis) results for a sliding window $w = T_h - T_c$ (“x”-axis) that gradually increases in size in the range [70 °C–150 °C]; (b) LT error when considering TA; (c) performance error when considering TA.

Next, possible trade-offs of various DTM constraints between lifetime and performance were examined. Figure 14a depicts the results for the constraints with a temperature window $w = T_h - T_c = 10$ °C sliding in the range [70 °C–150 °C]. As the window slid towards higher temperatures, the performance increased almost linearly, due to the HP operating mode being utilized more, as expected. At the same time, the lifetime dropped asymptotically towards zero. Figure 14b plots the relative error when comparing the lifetime prediction LT_{TA} against the lifetime LT_{fgT} computed by the proposed framework. The longest path was considered, and the lifetime estimation error was computed as in (11). The error was negative at lower temperatures, indicating that the lifetime estimation was optimistic, and became positive at higher temperatures, implying that it was pessimistic in this range. Figure 14c,d presents similar results for DTM constraints with $w = 20$ °C. Note that these results were consistent with those in Figure 13 and showed that, as w increased, the error of lifetime estimation using marginal, or even the average expected temperature, also increased and a fine-grained temperature consideration was required, which was only possible using the proposed framework.

Figure 14. Performance–lifetime trade-off exploration: (a) $T_c = 80\text{ }^{\circ}\text{C}$, $T_h = 100\text{ }^{\circ}\text{C}$; (b) $T_c = 80\text{ }^{\circ}\text{C}$, $T_h = 100\text{ }^{\circ}\text{C}$ zoom-in; (c) $T_c = 70\text{ }^{\circ}\text{C}$, $T_h = 110\text{ }^{\circ}\text{C}$ (d) $T_c = 70\text{ }^{\circ}\text{C}$, $T_h = 110\text{ }^{\circ}\text{C}$ zoom-in.

6. Conclusions

In this paper, it was shown that dynamic voltage and frequency scaling (DVFS) designs, together with stress-induced BTI variability exhibited high temperature-induced BTI variability due to their workload and different operating modes. Additionally, the impact of this variability on circuit lifetime was assessed, which can be higher than that due to stress. In order to account for the impact of this variability in lifetime estimation at design time, a simulation framework was proposed for the BTI degradation analysis of DVFS designs that considered their fine-grained thermal profiles as per the control of a dynamic thermal management (DTM) system. Using the proposed framework, we explored the expected lifetime and performance of the *Ethernet* circuit from the IWLS05 benchmark suite, synthesized with a 32 nm CMOS technology library, for various thermal constraints. The obtained results showed that a margin-based design can underestimate or overestimate the lifetime of DVFS designs by up to 67.8% and 61.9%, respectively. Finally, it was also demonstrated that, by using the proposed framework to appropriately select the dynamic thermal management constraints in order to trade-off long-term reliability (lifetime) and performance, a higher estimation accuracy up to 35.8% for lifetime and 26.3% for performance was obtained compared to a temperature-variability-unaware BTI analysis that considered only average temperature TA. The proposed framework can be suitably utilized for tuning run-time throttling policies of low-power designs, thus allowing designers to optimize lifetime–performance trade-offs, depending on the requirements mandated by specific applications and operating environments.

Author Contributions: Conceptualization, D.R. and V.T.; methodology, D.R. and V.T.; validation, D.R. and V.T.; writing—original draft preparation, D.R. and V.T.; writing—review and editing, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is partially supported by the Italian Ministry of Education and Research in the framework of the CrossLab project (Departments of Excellence), Department of Information Engineering, University of Pisa, Italy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kang, K.; Park, S.P.; Roy, K.; Alam, M.A. Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performance. In Proceedings of the 2007 IEEE/ACM International Conference on Computer-Aided Design, San Jose, CA, USA, 4–8 November 2007; pp. 730–734.
2. Chandra, V. Monitoring reliability in embedded processors—A multi-layer view. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1–5 June 2014.
3. Henkel, J.; Bauer, L.; Zhang, H.; Rehman, S.; Shafique, M. Multi-layer dependability: From microarchitecture to application level. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1–5 June 2014.
4. Hong, H.; Lim, J.; Lim, H.; Kang, S. Lifetime reliability enhancement of microprocessors: Mitigating the impact of Negative Bias Temperature Instability. *ACM Comput. Surv.* **2015**, *48*, 1–25. [\[CrossRef\]](#)
5. Alam, M.A.; Mahapatra, S. A comprehensive model for PMOS NBTI degradation. *Microelectron. Reliab.* **2005**, *45*, 71–81. [\[CrossRef\]](#)
6. Alam, M.A.; Kufluoglu, H.; Varghese, D.; Mahapatra, S. A comprehensive model for PMOS NBTI degradation: Recent progress. *Microelectron. Reliab.* **2008**, *47*, 853–862. [\[CrossRef\]](#)
7. Joshi, K.; Mukhopadhyay, S.; Goel, N.; Mahapatra, S. A consistent physical framework for N and P BTI in HKMG MOS-FETs. In Proceedings of the IEEE International Reliability Physics Symposium (IRPS), Anaheim, CA, USA, 15–19 April 2012; pp. 5A.3.1–5A.3.10.
8. Liu, C.; Kochte, M.A.; Wunderlich, H.J. Efficient observation point selection for aging monitoring. In Proceedings of the 2015 IEEE 21st International On-Line Testing Symposium (IOLTS), Halkidiki, Greece, 6–8 July 2015; pp. 176–181.
9. Borkar, S. Electronics Beyond Nano-Scale CMOS. In Proceedings of the 2006 43rd IEEE/ACM Design Automation Conference (DAC), San Francisco, CA, USA, 24–28 July 2006; pp. 807–808.
10. Paul, B.C.; Kang, K.; Kufluoglu, H.; Alam, M.A.; Roy, K. Negative Bias Temperature Instability: Estimation and design for improved reliability of nanoscale circuits. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2007**, *26*, 743–751. [\[CrossRef\]](#)
11. Agarwal, M.; Balakrishnan, V.; Bhuyan, A.; Kim, K.; Paul, B.C.; Wang, W.; Yang, B.; Cao, Y.; Mitra, S. Optimized circuit failure prediction for aging: Practicality and promise. In Proceedings of the IEEE International Test Conference (ITC), Santa Clara, CA, USA, 28–30 October 2008; pp. 1–10.
12. Yi, H.; Yoneda, T.; Inoue, M.; Sato, Y.; Kajihara, S.; Fujiwara, H. A Failure Prediction Strategy for Transistor Aging. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2012**, *20*, 1951–1959. [\[CrossRef\]](#)
13. Omaña, M.; Rossi, D.; Bosio, N.; Metra, C. Low cost nbti degradation detection and masking approaches. *IEEE Trans. Comput.* **2013**, *62*, 496–509. [\[CrossRef\]](#)
14. Rossi, D.; Omaña, M.; Metra, C.; Paccagnella, A. Impact of Bias Temperature Instability on Soft Error Susceptibility. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2015**, *23*, 743–751. [\[CrossRef\]](#)
15. Vijayan A.; Kiamehr S.; Oboril F.; Chakrabarty K.; Tahoori, M.B. Workload-aware static aging monitoring and mitigation of timing-critical flip-flops. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2018**, *37*, 2098–2110. [\[CrossRef\]](#)
16. Vijayan A.; Koneru, A.; Kiamehr S.; Chakrabarty K.; Tahoori, M.B. Fine-grained aging-induced delay prediction based on the monitoring of run-time stress. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2018**, *37*, 1064–1075. [\[CrossRef\]](#)
17. Wang, W.; Wei, Z.; Yang, S.; Cao, Y. An efficient method to identify critical gates under circuit aging. In Proceedings of the IEEE/ACM ICCAD, San Jose, CA, USA, 4–8 November 2007; pp. 735–740.
18. Kihwan, C.; Soma, R.; Pedram, M. Dynamic voltage and frequency scaling based on workload decomposition. In Proceedings of the 2004 International Symposium on Low Power Electronics and Design, Newport Beach, CA, USA, 9–11 August 2004; pp. 174–179.
19. Flynn, D.; Aitken, R.; Gibbons, A.; Shi, K. *Low Power Methodology Manual: For System-on-Chip Design*; Springer: New York, NY, USA, 2007.
20. Ricketts, A.; Singh, J.; Ramakrishnan, K.; Vijaykrishnan, N.; Pradhan, D. K. Investigating the impact of NBTI on different power saving cache strategies. In Proceedings of the IEEE/ACM 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010), Dresden, Germany, 8–12 March 2010.
21. Basoglu, M.; Orshansky, M.; Erez, M. NBTI-aware DVFS: A new approach to saving energy and increasing processor lifetime. In Proceedings of the 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), Austin, TX, USA, 18–20 August 2010.
22. Rossi, D.; Tenentes, V.; Khursheed, S.; Al-Hashimi, B.M. BTI and leakage aware dynamic voltage scaling for reliable low power cache memories. In Proceedings of the 2015 IEEE 21st International On-Line Testing Symposium (IOLTS), Halkidiki, Greece, 6–8 July 2015; pp. 194–199.
23. Rossi, D.; Tenentes, V.; Reddy, S.M.; Al-Hashimi, B.M.; Brown, A. Exploiting aging benefits for the design of reliable drowsy cache memories. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2018**, *37*, 1345–1357. [\[CrossRef\]](#)
24. Chahal, H.; Tenentes, V.; Rossi, D.; Al-Hashimi, B.M. BTI aware thermal management for reliable DVFS designs. In Proceedings of the 2016 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Storrs, CT, USA, 19–20 September 2016; pp. 1–6.
25. Tenentes, V.; Rossi, D.; Yang, S.; Khursheed, S.; Al-Hashimi, B.M.; Gunn, S.R. Coarse-grained online monitoring of BTI aging by reusing power-gating infrastructure. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2017**, *25*, 1397–1407. [\[CrossRef\]](#)

26. Predictive Technology Model (PTM). Available online: <http://ptm.asu.edu> (accessed on 10 September 2021).
27. Huang, W.; Sankaranarayanan, K.; Skadron, K.; Ribando, R.J.; Stan, M.R. Accurate, pre-RTL temperature-aware design using a parameterized, geometric thermal model. *IEEE Trans. Comput.* **2008**, *57*, 1277–1288. [[CrossRef](#)]
28. IWLS 2005 Benchmarks. Available online: <http://iwls.org/iwls2005/benchmarks.html> (accessed on 30 September 2021).
29. Fukui, M.; Nakai, S.; Miki, H.; Tsukiyama, S. A dependable power grid optimization algorithm considering NBTI timing degradation. In Proceedings of the IEEE NEWCAS, Bordeaux, France, 26–29 June 2011; pp. 370–373.
30. Yang, H.-I.; Hwang, W.; Chuang, C.-T. Impacts of NBTI/PBTI and contact resistance on power-gated SRAM with high-metal-gate devices. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2011**, *19*, 1192–1204. [[CrossRef](#)]
31. Kleeberger, V.B.; Maier, P.R.; Schlichtmann, U. Workload- and instruction-aware timing analysis—The missing link between technology and system-level resilience. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conf. (DAC), San Francisco, CA, USA, 1–5 June 2014.
32. Rossi, D.; Tenentes, V.; Yang, S.; Khursheed, S.; Al-Hashimi, B.M. Reliable power gating with NBTI aging benefits. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2016**, *24*, 2735–2744. [[CrossRef](#)]