BTI Aware Thermal Management for Reliable DVFS Designs

Hardeep Chahal, Vasileios Tenentes, Daniele Rossi, Bashir M. Al-Hashimi
*ECS, University of Southampton, UK. Email: {hsc1g13, V.Tenentes, D.Rossi, bmah}@ecs.soton.ac.uk

***Abstract*—In this paper, we show that dynamic voltage and frequency scaling (DVFS) designs, together with stress-induced BTI variability, exhibit high temperature-induced BTI variability, depending on their workload and operating modes. We show that the impact of temperature-induced variability on circuit lifetime can be higher than that due to stress and exceed 50% over the value estimated considering the circuit average temperature. In order to account for these variabilities in lifetime estimation at design time, we propose a simulation framework for the BTI degradation analysis of DVFS designs accounting for workload and actual temperature profiles. A profile is generated considering statistically probable workload and thermal management constraints by means of the HotSpot tool. Using the proposed framework we explore the expected lifetime of the ethernet circuit from the IWLS05 benchmark suite, synthesized with a 32nm CMOS technology library, for various thermal management constraints. We show that margin-based design can underestimate or overestimate lifetime of DVFS designs by up to 67.8% and 61.9%, respectively. Therefore, the proposed framework allows designers to select appropriately the dynamic thermal management constraints in order to tradeoff long-term reliability (lifetime) and performance with upto 35.8% and 26.3% higher accuracy, respectively, against a temperature-variability unaware BTI analysis.**

***Index Terms*—BTI, DVFS, DTM, temperature, lifetime**

I. Introduction

As the technology node shrinks, electronic systems become more prone to aging phenomena jeopardizing their reliability. Particularly, scaling to 32nm technology nodes and below leads to a change in the nature of reliability effects, which switch from abrupt functional problems to progressive degradation of the performance characteristics of devices and system components [1] induced by aging phenomena.

The dominating aging phenomenon for nanometer devices is bias temperature instability (BTI) [1], [2], whose main effect is to increase transistor threshold voltage, depending on technology parameters, operating voltage, temperature and stress ratio. If the induced performance degradation exceeds circuit time margins, it may lead to circuit failure and reduce lifetime of electronic systems. Great effort has been devoted to modeling of the effects of BTI and developing techniques to counteract them [2]–[6]. Design strategies adopted to tackle the negative effects of BTI aging include over-designing of the IC or monitoring the critical paths and opposing these effects during runtime [6], [7]. Besides, design margins are increased in order to design reliable integrated systems [1]. All these solutions may have a large impact on system performance and elevate the cost of reliability.

However, BTI aging degradation is workload and operating conditions (temperature, voltage, etc.) dependent [1], [7], [8]. Moreover, temperature and stress ratio values may vary from gate to gate, and even from transistor to transistor. In [1], a workload-dependent stress ratio computation framework is presented, which considers structural correlations of a circuit's logic. It is shown that different workloads can induce a variation up to 11% on circuit propagation delay. Therefore, paths that are not critical at time zero may become critical over time [1].

Nevertheless, the approach in [1] does not account for the effect on BTI aging of different temperatures at which different gates within a circuit may operate. In this regard, it is worth noticing that different workload leading to similar stress ratio distribution may lead to very different thermal profile. Indeed, whereas stress ratio depends on the time during which a transistor is ON, and is frequency independent, actual switching frequency may play an important role when the effect of temperature on aging is considered. 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. Moreover, in designs implementing dynamic voltage and frequency scaling (DVFS) techniques, the different operating modes impact considerably the circuit thermal profile and, consequently, its BTI aging. DVFS designs 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 and selecting appropriately its active operating mode in order to honour pre-defined constraints, such as performance and thermal design power (TDP). It is, therefore, expected that the pre-defined constraints used for DTM to affect the BTI aging of DVFS designs.

In this paper, by means of HSPICE simulation considering a 32nm High-k, CMOS technology from [9], we first show that the impact of temperature-induced BTI variability on circuit lifetime can be higher than that of stress-induced BTI variability. Particularly, by considering simple logic gates and three different input signal probabilities P_{IN} (0.25, 0.5, 0.75), with a constant temperature ($T = 80^oC$) we show that stress-induced variability on BTI can lead to a lifetime estimation variability reaching 43% for a 2-input NOR over the average value computed considering $P_{IN} = 0.5$ (3.59 years).

* This work is supported by EPSRC (UK) under grant no. EP/K000810/1 and EP/K034448/1.

Similarly, we assess the temperature induced variability on BTI aging. We consider an operating temperature varying from 70^oC to 90^oC, and we show that it can lead to a 58% lifetime estimation variability over the average value considered before.

In order to properly account for these variabilities in lifetime estimation at design time, we propose a simulation framework for the BTI degradation analysis of DVFS designs accounting for workload and actual thermal profiles, which are generated considering statistically probable workload and dynamic thermal management (DTM) constraints by means of the HotSpot tool [13]. The proposed approach allows us to obtain the fined-grain stress ratio for every transistor in a circuit as well as a fine-grained temperature profile. Particularly, for each and every transistor in the considered benchmark, we produce a unique model accounting for the unique stress ratio and operating temperature and we build a fine grained stress ratio and temperature aware aged library. We apply the developed simulation framework to the *Ethernet* circuit from the IWLS'05 suite, and we show that, depending on the considered DTM constraints, the margin-based design can underestimate or overestimate lifetime of DVFS designs by up to 67.8% and 61.9%, respectively. However, the proposed framework allows designers to explore the most appropriate DTM constraints according to a tradeoff between long-term reliability (lifetime) and performance with upto 35.8% and 26.3% higher accuracy, respectively, against a temperature-variability unaware BTI analysis.

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

II. Background

Bias temperature instability (BTI) has been extensively modelled using a number of methods, one of which is the reaction diffusion model [2]. 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 [2], [8]. In [8], [10], an analytical model has been proposed that allows designers to estimate long term, worst-case threshold voltage shift. It is:

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

The parameter C_{ox} is the oxide capacitance, t is the operating time, α is the fraction of the operating time during which a MOS transistor is under a stress condition (stress ratio), k is the Boltzmann constant, T the device temperature and E_a is a fitting parameter ($E_a \simeq 0.1eV$ [8]). The parameter K lumps technology specific and environmental parameters, and has been estimated to be $K \simeq 2.7V^{1/2}F^{-1/2}s^{-1/6}$ by fitting the model with the experimental results reported in [11]. The coefficient χ allows us to take into account the fact that NBTI ($\chi = 1$) prevails over PBTI ($\chi = 0.5$).

TABLE I
Stress ratio evaluation for 2-in NOR gate

IN1	IN2	MP1	MP2	MN1	MN2
0	0	s	s	r	r
0	1	s	r	r	s
1	0	r	r	s	r
1	1	r	r	s	s
Stress ratio		0.5	0.25	0.5	0.5

Note that the stress ratio allows designers to account for the actual stress applied to a transistor, which corresponds to the ratio of time the transistor is 'ON' over the total operating time. This value depends on input statistics, thus on workload. Also, note that the stress of individual transistors within a logic gate depends on the design logic structure.

III. Analysis

In this section, we assess the impact on propagation delay and lifetime of BTI variability induced by different stress ratios and operating temperatures. We show that these variability sources depend on workload (accounted for through signal probabilities, as in [1]), and on the operating modes of a DVFS design. In this regard, we should note that the DTM system strongly influence the power consumption of the DVFS design inducing temperature variabilities that should be considered for an accurate BTI aging estimation.

A. Stress-induced BTI Variability

To accurately consider aging during the timing analysis of a design, only the time each transistor is under stress should be considered. Since the workload may not be known, signal probabilities need to be considered, which are strongly influenced by the structural correlations of a logic design [1].

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 1 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. As an example, let us consider a 2-input NOR gate. While for the parallel nMOS transistors of the pull-down network straightforward considerations similar to those for the NOT gate apply, this is not the case for the series pMOS transistors of the pull-up network. Denote by MP1 the transistor connected to the power supply and MP2 that connected to the output. Table I is the *stress table* reporting either the stress (s) or recovery (r) condition for all transistors composing the NOR gate, for all input combinations. It is interesting to consider the input pattern IN1=1 and IN2=0. Since it is $V_{gs2} = V_{th}$ (absolute values), according to (1), transistor MP2 undergoes a recovery condition although its input is equal to a logic 0. The last row of the table shows the

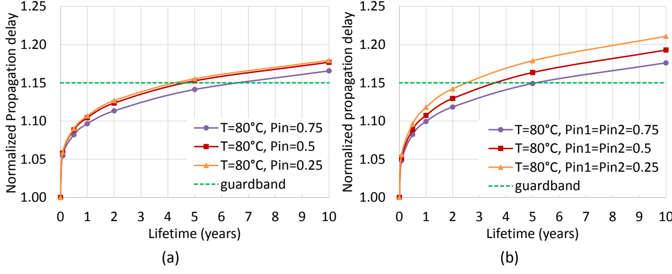

Fig. 1. Propagation delay variation with different input signal probabilities for a: (a) NOT gate; (b) NOR gate

stress ratio for all transistors, considering a signal probability $P_{IN1} = P_{IN2} = 0.5$ (input patterns equally likely).

The values of the stress ratio can be generalized as a function of the input probability as follows:

$$\alpha_{p1}^{nor} = \frac{t_{1s}}{LT} = 1 - P_{IN1}$$
$$\alpha_{p2}^{nor} = \frac{t_{2s}}{LT} = (1 - P_{IN2})(1 - P_{IN1}) \tag{2}$$

where t_{is} $(i = 1, 2)$ is the time during which a transistor is under stress and LT is the circuit lifetime. Hence, for the upper transistor MP1 to be stressed, its input has to be logic zero, whereas for the lower transistor MP1 both inputs have to be zero. Analogous considerations hold true for NOR gates with more than 2 inputs and for NAND gates.

In Fig. 1, we show that the stress-induced BTI variability from different signal probabilities is reflected on the propagation delay of logic paths and LT variability for the considered, implemented with a 32nm high-k metal-gate, CMOS technology from [9]. As for LT, it has been evaluated as the time interval the propagation delay takes to degrade by 15% over the value at *time zero* ($t0$) [12]. For the NOT gate, we considered an input signal probability $P_{IN} = 0.25, 0.5$ and 0.75. For the 2-input NOR gate, we show both the worst and best cases for propagation delay ($P_{IN1} = P_{IN2} = 0.25$ and $P_{IN1} = P_{IN2} = 0.75$, respectively), together with the average case ($P_{IN1} = P_{IN2} = 0.5$). As expected, the effect of stress-induced BTI variability on propagation delay is more evident in the NOR case, mainly due the cumulative aging effect on the series pMOS transistors. Although the propagation delay variability over the average value is limited to $\pm 2\%$, LT varies within a range of $[-30\%, +43\%]$ over the average value $LT = 3.59$ years.

Fig. 2 depicts the flow we have developed to evaluate the impact of logic gate input signal probabilities on the stress conditions of their transistors, considering also operating conditions (operating temperature and voltage). The obtained data for each gate are utilized to generate "aged" library. The library is used to map with the proper aging for each composing transistor and simulate complex circuits, as it will be discussed in more details in Section IV.

B. Temperature-induced BTI Variability

Workload impacts BTI variability not only due to stress ratio, but also due to temperature. Indeed, different workloads

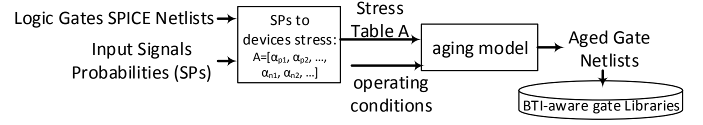

Fig. 2. Flow for stress & conditions aware gate models characterization

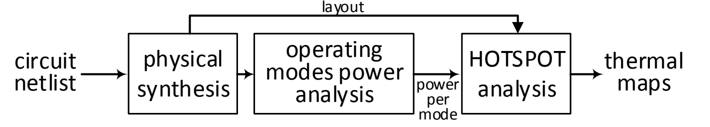

Fig. 3. Flow for obtaining thermal maps after physical synthesis

induce different switching activities, hence 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 reach 20^oC [13]. In addition, operating frequency and voltage, which do not affect stress ratio [2], 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. It is responsible for monitoring online the circuit temperature through on-chip sensors, and selecting appropriately its operating mode in order to honor pre-defined constraints, such as performance and power dissipation.

In order to account for the impact of different DVFS operating modes on a system thermal profile, we developed a flow that, starting from the physical synthesis of a circuit, generates the power analysis of each mode. This information, together with the layout, feeds the HotSpot tool [13] that conducts a thermal analysis of each operating mode. The block diagram of the developed flow is shown in Fig. 3.

As an example, we synthesize the *Ethernet* circuit (from the IWLS05 benchark suite) with DVFS with two operating modes, referred to as *low performance* (LP=0.5Ghz@0.7V) and *high performance* (HP=2GHz@1V). Although we consider that the circuit is equipped with a DTM system selecting the proper DVFS operating mode, we assume a fixed operating mode (either LP or HP). The steady state thermal analysis results for both LP and HP modes are shown in Fig. 4(a) and (b), respectively. As can be seen, the maximum temperature T_{max} (hotspot) for the LP mode is $\sim$72°C, while T_{max} reaches $\sim 121^oC$ for the HP mode. Particularly, in the HP operating mode, the hotspot is experienced in the upper area of the circuit, due to the higher dynamic power dissipation of the random logic located in that area compared to the memory in the lower part of the circuit. Instead, in the 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.

However, a circuit is meant to operate under the influence of the DTM system, the policies of which provide constraining of the temperature variability, which should be properly

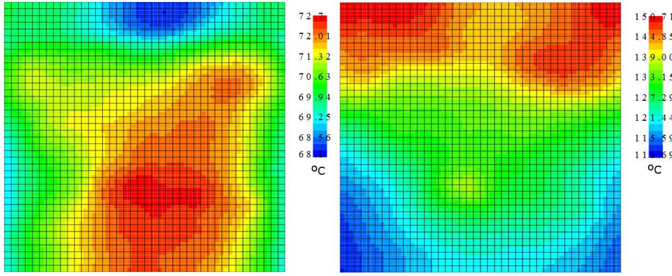

Fig. 4. Thermal maps for (a) LP operating mode and (b) HP operating mode

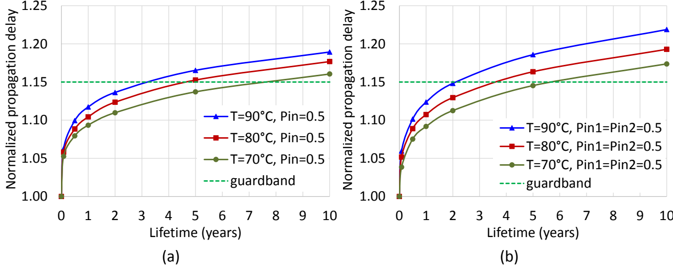

Fig. 5. Propagation delay variation with different temperature for (a) NOT gate; (b) NOR gate

considered for proper lifetime estimation during circuit design.

For comparison purposes, let us evaluate the effect of different temperatures considering the same basic gates as in Section III-A. Results for the *Ethernet* benchmark will be presented in Section V. As an example, assume that the operating temperature is bounded in the interval [70, 90]^oC by the DTM system. In Fig. 5 we show the temperature-induced BTI variability that reflects on propagation delay and lifetime (LT) variability, for a NOT gate (Fig. 5(a)) and a 2-input NOR gate (Fig. 5(b)). For the temperature, the upper and lower bounds are considered, and the average case (80^oC). The signal probability has been fixed at 0.5 in all cases.

Similarly to the stress-induced BTI variability analysis, the effect of temperature-induced BTI variability on propagation delay is more evident in the NOR case. The propagation delay variability over the average value is still rather limited $[-2\%, +3\%]$, whereas lifetime varies within a range of $[-41\%, +58\%]$ over the average value at $T = 80^oC$ ($LT = 3.59$ years). Comparing the results in Fig. 5 and Fig. 1, we can say that the effect of temperature-induced BTI variability on lifetime exceeds stress-induced one.

We can conclude 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 even higher than the contribution of stress variability. Therefore, both the stress and temperature variability induced by workload should be considered during BTI aware timing analysis. Moreover, we also showed that 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 estimation at design time.

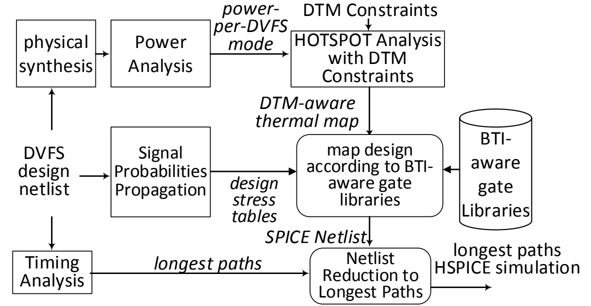

Fig. 6. Proposed framework for BTI degradation analysis of DVFS designs

IV. Proposed Simulation Framework

In this section, we discuss the simulation framework in Fig. 6 that we have developed for the BTI degradation analysis of DVFS designs accounting for workload and actual thermal profiles, which are generated considering statistically probable workload and DTM constraints utilizing the HotSpot tool [13].

Given an RTL netlist of the DVFS design, the signal probabilities of the logic nets are computed accounting for possible signal dependencies that may occur due to topological correlation, due to signals split up and reconvergence, and data-dependent correlation due to signal correlations at the circuit inputs [1], [14]. Then, static power and dynamic power analysis is performed accounting for a probabilistic workload induced by the signal probabilities. The power analysis data is used for a DTM aware thermal analysis in order to generate the temperature maps. Specifically, we execute the HotSpot tool using DTM constraints T_c and T_h at the lowest and the highest possible temperature at the hotspot of the circuit. We assume that a temperature sensor monitors the temperature at the hotspot of the circuit and feeds this information to the DTM controller. The operation of the circuit under such DTM constraints is described by means of an example:

Example 1. Consider the graph shown in Fig. 7, 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. 'x'-axis depicts time and 'y'-axis temperature of the circuit. 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 the LP mode with frequency f_{LP}, which causes the circuit to cool-down. When the temperature drops to a pre-defined cooling temperature constraint T_c, then the DTM system activates again the HP mode in order to maximize performance. The actual duration of each heat-up/cool-down time frames in this loop depends on the combination of workload-induced switching activity and temperature of the circuit. ■

From this process, a fine-grained temperature profile that considers the temperature constraints followed by the DTM system is generated and it is used for obtaining a fine-grained temperature for each logic gate in the design. This information,

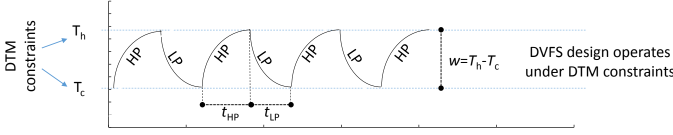

Fig. 7. DVFS design operating under DTM constraints

together with the signal probabilities of the logic gates is used for mapping each logic gate in the design with aged models from the aged gates library (Fig. 2). Finally, timing analysis is 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.

V. Simulation & Results

We applied the developed simulation framework for the BTI degradation analysis of the largest benchmark from the IWLS05 suite [15], the *Ethernet*. The synthesis of the benchmark has been conducted with a 32nm high-k metal gate CMOS technology [9] with DVFS using two operating modes, referred to as *low performance* (LP=0.5Ghz@0.7V) and *high performance* (HP=2GHz@1V), as introduced in Section III. Finally, based on the results for various dynamic thermal management (DTM) constraints, we select the appropriate constraints that meet either lifetime or performance requirements. For the evaluation of the performance, we use the results of the thermal analysis regarding the utilization of each operating mode. Consider again the example presented in Section IV (Fig. 7). Once the thermal analysis is conducted, the time that the circuit had spent in either the HP or the LP operating mode, t_{HP} and t_{LP} (shown in Fig. 7), respectively, is obtained. Then, the expected long-term performance of the circuit is evaluated using: $Perf = \frac{t_{LP}}{t_{LP}+t_{HP}} \cdot f_{LP} + \frac{t_{HP}}{t_{LP}+t_{HP}} \cdot f_{HP}$.

In Fig. 8, we present the average temperature of the *Ethernet* circuit under the dynamic thermal management (DTM) constraints $T_c = 80^oC$ and $T_h = 100^oC$. Note that for this case the temperature window w (Fig. 7) of the marginal constraints T_c and T_h is w=$T_h - T_c$=20^oC. The average temperature of the circuit at the hotspot was found $\sim$$90^oC$, while the average temperature of all the gates of the longest path was found $\sim$$88.1^oC$. The variability of the longest path delay due to the BTI aging is shown in Fig.9(a) and (b) for the considered temperatures. A margin-based temperature selection for aging evaluation, either using the DTM constraints T_c or T_h, results in a lifetime estimation of LT_{T_c}=4.41 and LT_{T_h}=2.01 years, respectively. Moreover, if we consider the average temperature of the DTM constraints TA=$(T_c$+$T_h)/2 = 90^oC$, the lifetime estimation is LT_{TA}=3.1 years. However, when the fine-grained temperature fgT at the gates of the longest path are considered for mapping the design with the proposed framework, then the lifetime estimation becomes LT_{fgT}=3.25 years. Thus, an optimistic evaluation using the T_c temperature underestimates the effect of BTI on the lifetime of the circuit by 35% and a pessimistic one using the T_h temperature overestimates it by 38.2%. Even when the average TA =90^oC

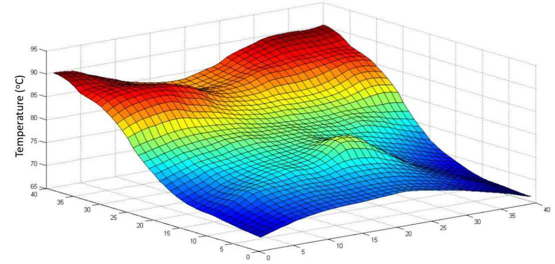

Fig. 8. Temperature profile for DTM constraints T_c=$80^oC, T_h$=100^oC

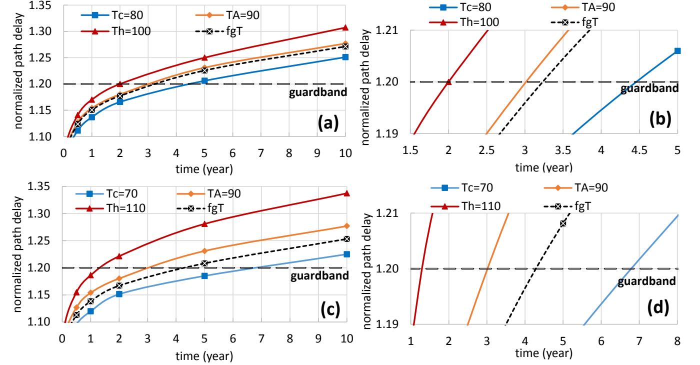

Fig. 9. Ethernet longest path delay for various DTM constraints (a) T_c=80^oC, T_h=100^oC; (b) T_c=80^oC, T_h=100^oC zoom-in; (c) T_c=70^oC, T_h=110^oC (d) T_c=70^oC, T_h=110^oC zoom-in

of the two marginal constraints is considered, the BTI effect on the lifetime of the circuit is overestimated by 4.8%.

Next, we will show that the deviation between the estimated lifetime of marginal temperatures against the estimated lifetime with the proposed framework, which considers fine-grained temperatures for each logic gate, depends on the window size w between the DTM constraints. In Figs. 9(c) and (d), we present the longest path propagation delay with time for $T_c = 70^oC$ and $T_h = 110^oC$. For this case the window w is 40. From Fig. 9(d), we obtain that $LT_{T_c} =$ 6.8 years, $LT_{T_h} =$ 1.35 years, $LT_{TA} =$ 3.1 years and $LT_{fgT} =$ 4.2 years. As a result, an optimistic evaluation using the T_c temperature underestimates the effect of BTI on the lifetime of the circuit by 61.9% and a pessimistic one using the T_h temperature overestimates it by 67.8%. When the average TA =90^oC between the two marginal constraints is considered, the BTI effect on the lifetime of the circuit is overestimated by 26.1%.

In Table II, we present results on performance and lifetime obtained from the proposed technique for DTM constraints with $w \leq 20$ values ($w = 2, 10, 16, 18, 20$). The first column presents results for constraints with $TA = 80^oC$ and the second one with $TA = 130^oC$. For $TA = 80^oC$, we observe that LT of the circuit increases with the increase of w. This is attributed to a performance reduction that is also observed while w increases. The reason of these trends is that the selected average temperature (80^oC) causes a higher utilization of the LP operating mode. Since the sensor is located at the hotspot of the circuit, it overestimates the average temperature of the circuit and forces an even higher utilization of the

TABLE II
LIFETIME AND PERFORMANCE TRADEOFF OF CONSTRAINTS ($w < 20^oC$)

w	DTM Constraints @ 80°C $[T_c, T_h]$	LT	Perf	e_{LT} %	DTM Constraints @ 130°C $[T_c, T_h]$	LT	Perf	e_{LT} %
2	79, 81	4.52	0.88	0.56	129, 131	0.61	1.57	-0.50
10	75, 85	4.63	0.70	2.98	125, 135	0.61	1.59	-1.10
16	72, 88	4.91	0.67	8.40	122, 138	0.60	1.62	-2.81
18	71, 89	5.08	0.65	11.49	121, 139	0.59	1.63	-3.68
20	70, 90	5.38	0.62	**16.42**	120, 140	0.59	1.63	**-4.41**

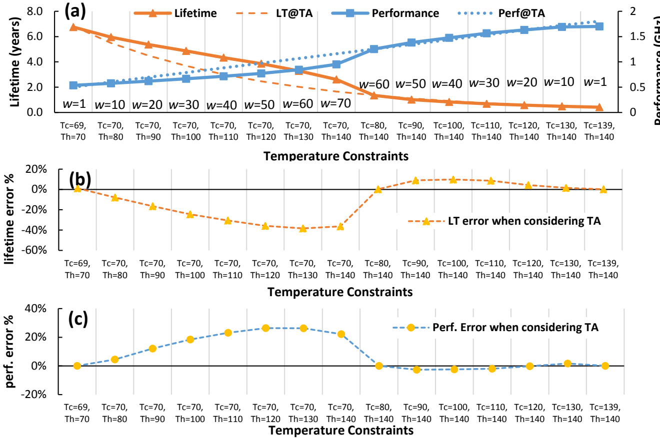

Fig. 10. (a) Lifetime (left 'y'-axis) against performance (right 'y'-axis) results for a slighting window $w = T_h - T_c$ ('x'-axis) that gradually increases in size in the range [70^oC-150^oC]; (b) LT error when considering TA; (c) performance error when considering TA

LP operating mode than what necessary to meet the desired constraints. For higher temperatures (second column, 130^oC), the exact opposite trend is observed. The LT drops as w increases and the performance increases, which is attributed to the already very high utilization of the HP mode at those temperatures. Column e_{LT} reports the lifetime estimation error when the average temperature TA is considered against the case of the fine-grained temperature of the longest path.

The lifetime (left 'y'-axis) against performance (right 'y'-axis) results are depicted in Fig. 10(a) for a temperature window w ('x'-axis) that gradually increases in size in the range [70^oC-150^oC]. It also depicts the expected lifetime and the performance when the average temperature of the margninal DTM constraints TA is only considered (dashed lines labeled as "LT@TA" and "Perf@TA"). Fig. 10(b) depicts the errors for the examined constraints between the expected and the actual lifetime at TA. Note that at lower temperatures and while the w increases, the underestimation of the lifetime using the average temperature of marginal constraints also increases, exceeding 35% for $w \geq 40^oC$. At higher temperatures, the lifetime can be overestimated by more than 10% for $w \geq 40^oC$. Fig. 10(c) depicts the error between the expected and actual performance at TA. Note that at lower temperatures and while the w increases the expected performance is overestimated by >20% for $w \geq 30^oC$ (reaching 26.3%), when the DTM constraints are not considered. At very high temperatures, the performance is slightly underestimated (<1.2%), which can be considered negligible.

VI. CONCLUSIONS

We showed that dynamic voltage and frequency scaling (DVFS) designs, together with stress-induced BTI variability, exhibit high temperature-induced BTI variability due to their workload and different operating modes. We, also, showed that the impact of this variability on circuit lifetime can be higher than that due to stress. In order to account for this variability during the lifetime estimation at the design time, we proposed a simulation framework for the BTI degradation analysis of DVFS designs that considers thermal profiles from DVFS designs under the influence 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 32nm CMOS technology library, for various thermal management constraints. We showed that margin-based design can underestimate or overestimate lifetime of DVFS designs by up to 67.8% and 61.9%, respectively. Finally, we used the proposed framework to select appropriately the dynamic thermal management constraints in order to tradeoff long-term reliability (lifetime) and performance with upto 35.8% and 26.3% higher accuracy, respectively, against a temperature-variability unaware BTI analysis.

REFERENCES

[1] V. Chandra, "Monitoring reliability in embedded processors - a multi-layer view," in *Design Automation Conf. (DAC)*, June 2014, pp. 1–6.
[2] M. Alam *et al.*, "A comprehensive model for PMOS NBTI degradation: Recent progress," *Microelectronics Reliability*, vol. 47, no. 6, pp. 853 – 862, 2007, modelling the Negative Bias Temperature Instability.
[3] S. Borkar, "Electronics beyond nano-scale cmos," in *in Proc. IEEE/ACM Design Automation Conference (DAC)*, 2006, pp. 807–808.
[4] M. Agarwal *et al.*, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. of IEEE International Test Conf. (ITC)*, 2008, pp. 1–10.
[5] H. Yi *et al.*, "Impact of bias temperature instability on soft error susceptibility," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 1951–1959, 2012.
[6] M. Omaña *et al.*, "Low cost nbti degradation detection and masking approaches," *IEEE Trans. on Computers*, vol. 62, no. 3, pp. 496–509, 2013.
[7] C. Liu, M. A. Kochte, and H. J. Wunderlich, "Efficient observation point selection for aging monitoring," in *On-Line Testing Symposium (IOLTS), 2015 IEEE 21st International*, July 2015, pp. 176–181.
[8] K. Joshi *et al.*, "A consistent physical framework for n and p bti in hkmg mosfets," in *in Proc. IEEE International Reliability Physics Symposium (IRPS)*, 2012, pp. 5A.3.1–5A.3.10.
[9] "Predictive Technology Model (PTM)," http://ptm.asu.edu.
[10] M. Fukui *et al.*, "A dependable power grid optimization algorithm considering nbti timing degradation," in *IEEE NEWCAS 2011*, June 2011, pp. 370–373.
[11] H.-I. Yang, W. Hwang, and C.-T. Chuang, "Impacts of nbti/pbti and contact resistance on power-gated sram with high-metal-gate devices," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 7, pp. 1192–1204, 2011.
[12] D. Rossi *et al.*, "Reliable power gating with nbti aging benefits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. PP, no. 99, pp. 1–10, 2016.
[13] W. Huang *et al.*, "Accurate, pre-rtl temperature-aware design using a parameterized, geometric thermal model," *IEEE Transactions on Computers*, vol. 57, no. 9, pp. 1277–1288, Sept 2008.
[14] V. Kleeberger, P. Maier, and U. Schlichtmann, "Workload- and instruction-aware timing analysis - the missing link between technology and system-level resilience," in *Design Automation Conf. (DAC)*, June 2014, pp. 1–6.
[15] IWLS'05, online: http://iwls.org/iwls2005/benchmarks.html.