Coarse-grained Online Monitoring of BTI Aging by Reusing Power Gating Infrastructure

Vasileios Tenentes, *Member*, *IEEE*, Daniele Rossi, *Member*, *IEEE*, Sheng Yang, Saqib Khursheed, and Bashir M. Al-Hashimi, *Fellow*, *IEEE* and Steve R. Gunn

***Abstract*—In this paper, we present a novel coarse-grained technique for monitoring online the Bias Temperature Instability (BTI) aging of circuits by exploiting their power gating infrastructure. The proposed technique relies on monitoring the discharge time of the virtual-power-network during stand-by operations, the value of which depends on the threshold voltage of the CMOS devices in a power-gated design (PGD). It does not require any distributed sensors, because the virtual-power-network is already distributed in a PGD. It consists of a hardware block for measuring the discharge time concurrently with normal stand-by operations and a processing block for estimating the BTI aging status of the PGD according to the collected measurements. Through SPICE simulation, we demonstrate that the BTI aging estimation error of the proposed technique is less than 1% and 6.2% for PGDs with static operating frequency and dynamic voltage and frequency scaling, respectively. Its area cost is also found negligible. The power gating Minimum Idle Time (MIT) cost induced by the energy consumed for monitoring the discharge time is evaluated on two scalar machine models using either x86 or ARM instruction sets. It is found less than 1.3X and 1.45X the original power gating MIT, respectively. We validate the proposed technique through accelerated aging experiments conducted with five actual chips that contain an ARM cortex M0 processor, manufactured with a 65nm CMOS technology.**

***Index Terms*—BTI, aging, sensor, power gating**

I. Introduction

Bias Temperature Instability (BTI), which is the major aging mechanism in very deep sub-micron CMOS technologies [1], induces detrimental effects to devices, such as performance degradation, which can lead to in-the-field failures. Many techniques for monitoring online the BTI provide a warning about imminent faults by focusing at its local detrimental effects. They monitor, in a *fine-grained* fashion, devices or paths in a design that are more vulnerable to aging [2]–[12].

The sensors utilized for such fine-grained BTI monitoring fall mainly into two categories: sensors monitoring path delay [2], [3] of logic circuits and sensors monitoring frequency drift in ring-oscillators [4]–[6]. The former require the sensitization of critical paths with patterns providing a warning indication, when the path-delay has violated a pre-defined delay threshold. The latter integrate ring oscillators at stressed areas and monitor the aging status of the sensors. Hybrid methods also exist [7], [8]. Other methods [9]–[11] reduce the area cost by selecting a subset of critical paths to monitor. However, many paths of modern circuits can become critical in-the-field due to temperature and workload variability [1], [12]. Therefore, for online fine-grained BTI monitoring, multiple devices or paths should be monitored at various pre-defined delay thresholds, impacting inevitably design complexity and area/power cost.

Many online applications require a global indication about the BTI status of a circuit without a warning indication about imminent faults. For such applications, a low cost indication about the BTI status of a design, in a coarse-grained fashion, can be practical, and the high cost of fine-grained monitoring could be avoided. One such application is the reliability management of multi-core systems that requires a BTI indication for balancing workload among identical cores under long-term reliability constraints. Such cores share similar workload and, therefore, similar fine-grained degradation characteristics. Another application is the Dynamic Thermal/Power Management (DTM) of System-on-Chips (SoCs), such as those of smart SoCs [13], [14], that tune online power reduction techniques [15], [16] according to measurements (voltage noise, temperature etc.) provided by on-chip sensors. Recent results [17], [18] show that the BTI induced threshold voltage V_{th} degradation of CMOS devices is not only accompanied by detrimental effects, but also by some benefits. Leakage power reduction techniques become more efficient [17], [19] and static power consumption decreases over time [18], [20]. Therefore, for DTM systems to harvest such *aging benefits*, a coarse-grained BTI indication would suffice. Finally, fine-grained BTI monitoring is not very practical for memories.

In this paper, we present a novel coarse-grained BTI aging monitoring technique, which is applicable on power-gated designs (PGDs). Power gating has already been proven as an effective solution to tackle static power consumption and has been widely adopted in many modern processors [21]. We show that the leakage current reduction of BTI aging in nanometer technologies [17], [18] impacts considerably the virtual-power-network discharge time during the stand-by of a PGD. The proposed technique consists of a hardware block for measuring online the virtual-power-network discharge time and a processing block for estimating the BTI aging status of the PGD according to the collected measurements. The proposed technique provides an indication about the average aging status of all the CMOS devices in the PGD and cannot

V. Tenentes, D. Rossi, B. M. Al-Hashimi and S. R. Gunn are with the Department of Electronics and Computer Science, University of Southampton, UK. Email: {V.Tenentes, D.Rossi, bmah, srg}@ecs.soton.ac.uk

S. Yang is with ARM Ltd, Cambridge, UK. Email: sheng.yang@arm.com

S.Khursheed is with the Department of Electrical Engineering & Electronics, University of Liverpool, UK. Email: S.Khursheed@liverpool.ac.uk

Manuscript received May XX, 2016.

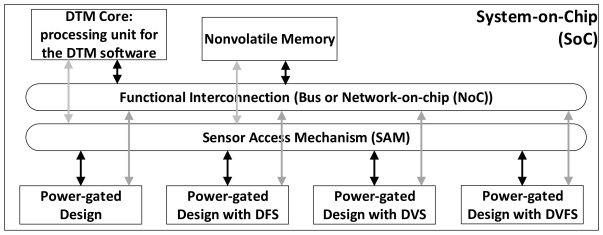

Fig. 1. SoC architecture with embedded Dynamic Thermal Management

be used for providing a warning about imminent faults. However, it features some advantages over path-based monitoring techniques. First, the discharge time is measured on the virtual-power-network, which is already distributed in the PGD, and thus distributed sensors are not required. Second, high aging estimation resolution is achieved, because the impact of aging on the discharge time is in the order of hundreds of nanoseconds, while on path-delay it is in the order of picoseconds. Third, it is also applicable to memories, because the discharge time is sensitive to the aging status of all the CMOS devices in the design and the workload is not required to be known during design. Finally, the proposed technique is performed concurrently with normal stand-by operations, enabling the harvesting of BTI static power reduction benefits by online applications, such as the DTM system of SoCs. To the best of our knowledge, this is the first coarse-grained technique for online BTI monitoring.

The remainder of this paper is organized as follows. The SoC architecture with DTM and the discharge time of the virtual-power-network, denoted as d_V hereafter, are introduced in Section II. Results of static power consumption reduction on designs due to BTI aging are also discussed. The proposed technique for monitoring the average threshold voltage degradation induced by BTI, which consists of an on-chip d_V sensor and a processing block is presented in Section III. The performance and the area cost of the proposed technique is evaluated by means of SPICE simulation of IWLS'05 [22] benchmarks in Section IV. Results on the energy consumed by the processing block using two scalar machine models with x86 and ARM instruction sets are also presented, and its impact on the power gating Minimum Idle Time (MIT) [23] is also evaluated. The discharge time d_V sensitivity to aging is validated through accelerated aging experiments conducted using five actual chips with an SoC that contains an ARM cortex M0 processor fabricated with a 65nm technology in Section V. Finally, conclusions are drawn in Section VI.

II. Background and Motivation

Figure 1 shows an SoC architecture with embedded DTM system [1], [14]. Designs with different power management capabilities, such as power gating and Dynamic Voltage and Frequency Scaling (DVFS), are integrated into the SoC. The DTM system consists of a DTM core and software. It collects measurements from on-chip sensors related to the status of the designs (power consumption, temperature, aging, etc), and optimizes their features (performance, power consumption, temperature, reliability) by controlling (accordingly) the

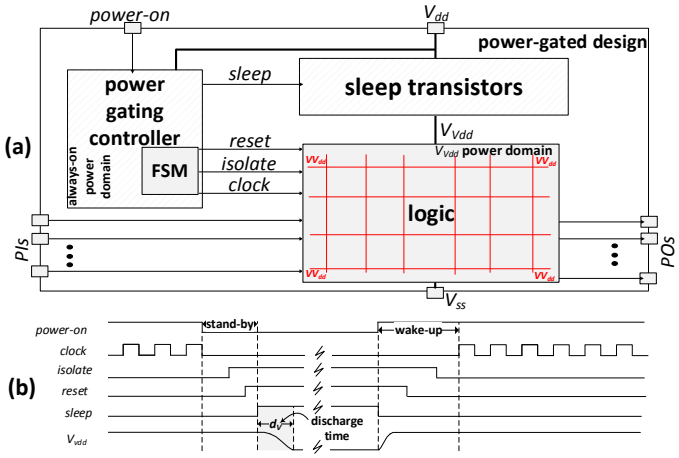

Fig. 2. (a) Power-gated design (PGD); (b) power gating control protocols

power-management capabilities of the designs [14]. The interconnection between the designs and the DTM core is achieved through the functional interconnection (Bus or Network-on-chip (NoC)) [1], shared Nonvolatile memory (NVM) [1] and Sensor Access Mechanisms (SAM) [24]. The DTM core will be used for processing data coming from an on-chip sensor.

Power gating is a static power reduction technique that adds PMOS *Header* and/or NMOS *Footer* power switches, often referred to as *sleep transistors* (STs), that allow a circuit to operate in two modes: the *power-on* and the *power-off* mode. The general scheme using header STs is shown in Figure 2(a). During periods of inactivity, the circuit is set in the power-off mode in order to reduce static power consumption. STs are used for disconnecting the virtual power supply V_{Vdd} of the circuit from the power supply V_{dd}. The *wake-up* (power-off $\rightarrow$ power-on) and the *stand-by* (power-on $\rightarrow$ power-off) *operations* are implemented by a finite state machine (FSM) that resides in the always-on (operating with V_{dd}) power domain of the power gating controller. Each operation follows a protocol to coordinate the activation and deactivation of design features, such as clock gating, isolation and state retention [21]. A typical case, where the circuit is equipped with clock-gating and isolation features, is shown in Figure 2(b). With the de-assertion of the *power-on* signal, the protocol of the stand-by operation applied is: a) enable clock-gating; b) enable isolation by asserting *isolate* signal; c) reset the power-gated logic by asserting the *reset* signal; d) disconnect the V_{Vdd} from V_{dd} by asserting the *sleep* signal to open the STs. The protocol of the wake-up operation is the reverse sequence of actions. Note that the operations of a power gating design (PGD) can be self-controlled or externally controlled. For the first case, PGDs contain specialized idle-time monitoring circuitry for detecting idle periods during their operation, and for the second case, they are controlled by an external processing block (the DTM core at our case), which selects the best suited idle intervals according to system beneficial objectives (minimizing power, temperature and maximizing reliability etc.). The proposed coarse-grained BTI monitoring technique has been considered for the second case of PGDs. However, in principle, it is also applicable to the first case. Another

approach [25], provides the self-controlled ability without requiring any idle-time monitoring circuitry by deploying pre-defined idle intervals together with intervals that the circuit operates at higher than nominal voltage. A coarse-grained BTI indication could also be beneficial to this approach.

We point out that the virtual power supply V_{Vdd} is distributed by a virtual-power-network in the design, as shown in Figure 2(a). We consider to use the virtual-power-network discharge time d_V (shown in Figure 2(b)), which is the time required by the virtual-power-network to discharge after the assertion of the *sleep* signal during a stand-by protocol application, for monitoring the BTI of power-gated designs.

Recent research on the effect of V_{th} degradation of CMOS devices induced by BTI presented a significant leakage current reduction. It was shown in [18] that after only 1 month of operation, the power consumption due to leakage current drops to 50% compared to the initial power consumption at time t=0. It further reduces to less than 30% and 20% after 1 year and 10 years of operation, respectively. The results presented in [18] consider all possible leakage current components. However, since high-k technologies (thicker dielectrics) reduce considerably the gate leakage [21], and the junction leakage I_j is not affected by the V_{th} [26], this phenomenon has been attributed to a reduction of the sub-threshold leakage current I_{sub-th}. Particularly, when the STs are OFF the virtual-power-network V_{Vdd} discharges via the leakage current I_{leak} [21]:

$$I_{leak} \simeq I_j + I_{sub-th} \propto I_j + (W/L)e^{\frac{-|q|V_{th}}{\lambda kT}} \quad (1)$$

where W the width and L the length of device channel, q the electron charge, k the Boltzmann constant, T the temperature, and λ a fabrication characterization parameter. According to BTI aging models [27], [28], V_{th} increases over time, an effect that decreases circuit sub-threshold current I_{sub-th} exponentially over time, as derived by (1). Previous BTI monitoring techniques monitor either the path delay or the frequency drift of ring oscillators, which are effected by the active current. The active current varies almost linearly with the V_{th} [29]. However, the I_{sub-th} of a circuit, which affects the discharge time d_V, varies exponentially with V_{th}. Therefore, it is expected for I_{sub-th} to be more sensitive to the V_{th} than the active current, especially after the early lifetime of the circuit, when the variability of the V_{th} with time t is lower. These observations motivated the exploration of the virtual-power-network discharge time, which is affected by the I_{sub-th}, for monitoring BTI.

Note that other aging mechanism may also affect leakage. For example, the hot carrier injection (HCI) affects, similarly to BTI, the threshold voltage and the time dependent dielectric breakdown (TDDB) causes a sudden oxide collapse increasing the gate leakage. However, with high-k dielectrics the gate leakage is orders of magnitude lower than the sub-threshold.

III. Proposed BTI Monitoring Technique for PGDs

The proposed BTI aging monitoring technique consists of a virtual-power-network discharge time d_V sensor and an online processing block for estimating BTI aging according to the collected measurements, which are described next. The cost of the processing block is also analyzed.

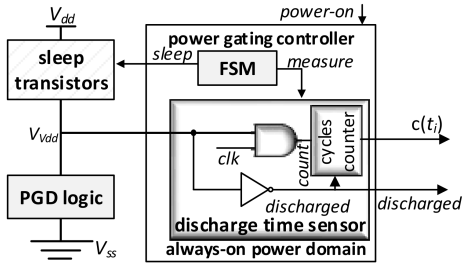

Fig. 3. Virtual-power-network discharge time d_V sensor architecture

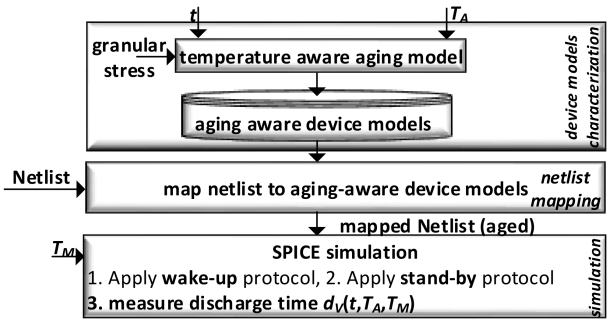

Fig. 4. Characterization process

A. The discharge time sensor

The d_V sensor, shown in Figure 3, is a very small circuit that resides in the power-gating controller and operates as a time-to-digital converter. This type of sensors is already used by power gating DFT infrastructure [30]. The power gating FSM controls the sensor by asserting the *measure* signal together with the *sleep* signal in order to collect the d_V measurement on every stand-by operation. Then, the sensor, which consists of only a logic AND gate, an inverter and a counter, counts the clock rising edges c until the virtual voltage V_{Vdd} drops to logic-'1'. This happens when the inverter input (V_{Vdd}) drops below $m \cdot V_{dd}$, where $m \cdot V_{dd}$ its logic threshold voltage. Then, its output, the *discharged* signal, switches to logic-'1', deasserting the enable signal of the counter. The $c(t_i)$ value of the counter is the d_V at time t_i expressed in clock cycles. Therefore, the measured d_V is $d_V(t_i) = c(t_i) \times T_{clk}$, where T_{clk} is the circuit clock period. Although the logic threshold voltage $m \cdot V_{dd}$ of the inverter affects the absolute $d_V(t_i)$ value, it does not affect the relative value, which is evaluated as $d_V(t_i)/d_V(t{=}0)$, where $d_V(t{=}0)$ is the discharge time at t=0. However, a logic threshold voltage $m \cdot V_{dd}$ lower than $0.15 \cdot V_{dd}$ should be avoided in order to limit the discharge time (and so the monitoring time) to hundreds of nanoseconds.

B. Collection and analysis of characterization data

The d_V BTI aware characterization process is shown in Figure 4. First, CMOS device models are characterized with the ΔV_{th} using [27], [28] for various values of the aging temperature T_A and operating time t. Statistical evaluation of the workload impact on devices stress was used [12] using structural correlations of the logic. We considered temperature $T_A{\in}S_T{=}[60, 80, 100, 120]^oC$ and time $t{\in}S_t{=}[0, 1/12, 2/12, \ldots, 1, 2, \ldots, 10]$ years. Next, given the

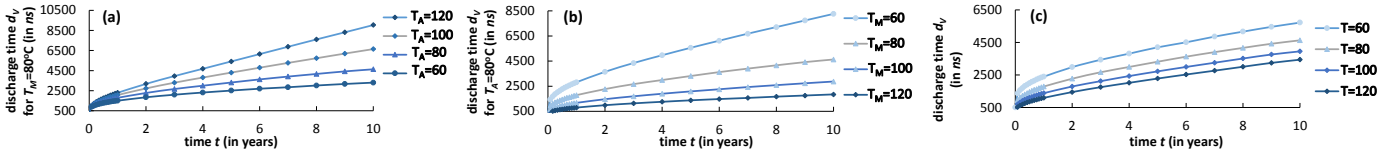

Fig. 5. Discharge time in time t when: (a) constant $T_M = 80^oC$; T_A varies (b) constant $T_A = 80^oC$; T_M varies; (c) $T_M = T_A = T$.

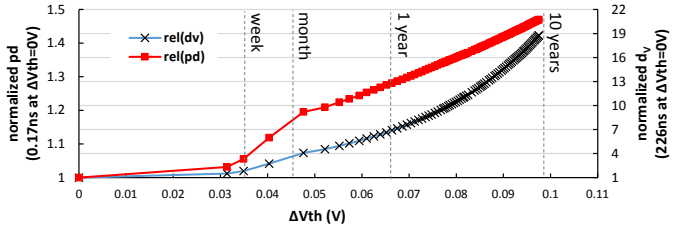

Fig. 6. Propagation delay and discharge time trend with ΔV_{th} at pMOS

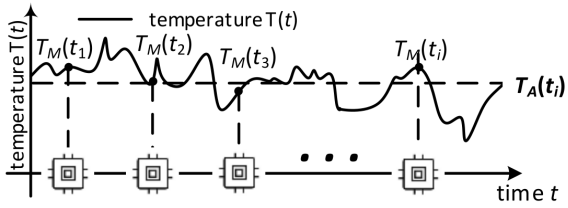

Fig. 7. $\sum_{j=1}^{j=i} T_M(t_j) \rightarrow T_A(t_i)$: the average temperature during stand-by converges to average aging temperature.

time t and aging temperature T_A, a PGD netlist is mapped with the device models, accounting for the proper BTI degradation. Finally, given the temperature T_M during the stand-by operation, we measure the d_V through SPICE simulation of the mapped netlist. The collected values will be referred to as d_V *characterization data*.

The characterization is applied to a PGD of 21 cascaded inverters (casc21) synthesized with a 32nm High-k metal gate CMOS technology [31]. We have considered a small circuit in order to explore the trade-offs using SPICE simulation. The operating frequency of this circuit, including 30% guardband, is lower than 4Ghz, which is usually the highest frequency of commercial applications. The number of sleep transistors is selected to fulfil the constraint of an IR-drop $\leq$10% in this analysis. The synthesis and SPICE simulations are conducted using commercial EDA tools. The d_V characterization data are presented in Figure 5 and are discussed below.

In Figure 5(a), we show the d_V characterization data when the temperature during stand-by operation is kept constant at T_M=80^oC and the average aging temperature T_A varies as follows: $T_A \in S_T$=$[60, 80, 100, 120]^oC$. As expected, the d_V increases as time t and aging temperature T_A increase. Indeed, from (1) we derive that the sub-threshold leakage current of the devices of the circuit decreases as their threshold voltage increases because of BTI [18], [21]. In Figure 5(b), we present the d_V characterization data when the average aging temperature is kept constant at T_A=80^oC and the temperature during stand-by T_M varies as follows: $T_M \in S_T$=$[60, 80, 100, 120]^oC$. In this case, the d_V decreases considerably with the temperature during stand-by, since the sub-threshold leakage current (1) of the devices of the circuit increases substantially with the temperature [21]. If we compare the d_V range of values in Figure 5(a) ($2507ns$ to $5411ns$) with that in Figure 5(b) ($1375ns$ to $5561ns$) for a specific time (t=5 years), we conclude that the effect of the temperature during stand-by T_M on d_V overwhelms the effect of average aging temperature T_A. In Figure 5(c), we present the d_V characterization data for average aging temperature T_A equal to the temperature during stand-by T_M, T_A=T_M, selected by set S_T=$[60, 80, 100, 120]^oC$. We note that for the same time t_i the d_V decreases with temperature, thus confirming the great sensitivity of the d_V to the temperature during stand-by T_M. In Section V, we collect measurements from actual chips that follow the d_V trends shown in Figure 5. Hence, d_V characterization data could also be fitted on actual measurements and points can be obtained using extrapolation.

In Figure 6, we present the impact of the BTI induced ΔV_{th} ('x'-axis) of the pMOS devices at the propagation delay *pd* (left 'y'-axis) of the casc21 and at its virtual-power-network discharge time d_V (right 'y'-axis) measured for $T_A = T_M = 100^oC$. The graphs depict the relative values compared to those at $t = 0$, when also $\Delta V_{th} = 0$. As expected, the trends validate that the propagation delay *pd* is affected almost linearly by ΔV_{th}, increasing upto 1.47x (times) after 10 years, while the discharge time d_V is affected exponentially increasing upto 18.7x (times) after 10 years.

C. Online processing block and cost analysis

The basic concept for monitoring BTI aging by processing the virtual-power-network discharge time d_V is described by means of the example shown in Figure 7. During time ($x-$axis), a circuit operates at various temperatures $T(t)$ ($y-$axis) and executes many times the stand-by operation at various time moments t_i. We note that the temperature T_M can be considered constant during the dischage time, which is in the order of nanoseconds and much shorter than the thermal transient cool-down from power-on to power-off mode, which is in the order of microseconds [32]. While time increases, the average aging temperature $T_A(t_i) = \sum_{t=t_1}^{t=t_i} T(t)/i$ is affecting the ΔV_{th} due to BTI [27], [28]. However, while $t_i \rightarrow t_\infty$ both the average aging temperature $T_A(t_i)$ and the average temperature during stand-by $T_{AM}(t_i) = \sum_{t=t_1}^{t=t_i} T_M(t)/i$ converge to a constant value. Therefore, we consider that the temperature during stand-by $T_M(t)$ is a random variable that follows the deviation of $T(t)$. This assumption is realistic, because each $T_M(t_i)$ is a sample of the $T(t)$ at the moment of stand-by operation $t = t_i$, as shown in Figure 7. Later, in Section IV-D, we present results when this assumption is removed.

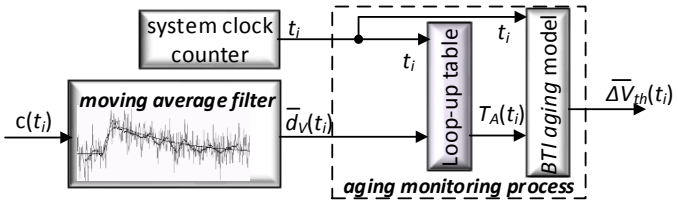

Fig. 8. Online processing for aging estimation

The online processing block is shown in Figure 8. A cumulative moving average filter is utilized to compute the *average* d_V from the history of stand-by operations. The filter is described by: $\overline{d_V}(t_i) = (\overline{d_V}(t_{i-1}) + s \cdot d_V(t_i))/(s+1)$, where s is the *convergence speed*, $d_V(t_i) = c(t_i) \times T_{clk}$ the measured d_V in seconds, $c(t_i)$ the discharge time d_V in circuit clock cycles, and T_{clk} the circuit clock period. This filter, which is applied whenever the *discharged* signal is asserted, requires time to converge to the average discharge time. A higher s value makes the filter to converge faster, but with a higher sensitivity to noise, as will be shown in Section IV. Note that the average discharge time $\overline{d_V}$ that is provided by the moving average filter depends on the average temperature T_{AM} during every previous stand-by. Therefore, as T_{AM} converges to the average aging temperature T_A, the computed $\overline{d_V}$ depends only on the aging status of the circuit. Based on the $|S_t| \times |S_T|$ collected d_V characterization data (Figure 5(c)), which are discrete d_V points in the space $t \times T$, the function $T_A(t, d_V)$ can be approximated using either interpolation coefficients [33] (cubic or linear) or a look-up table. A non-volatile memory, which is accessible for online processing, stores this data. The aging temperature $T_A(t_i, \overline{d_V})$ until time moment t_i is computed using the stored data and the average discharge time $\overline{d_V}(t_i)$ provided by the moving average filter. Then, a BTI model [27], [28] is used to compute the average ΔVth degradation of the CMOS devices in the PGD upon time t_i, as shown in Figure 8. The processing block is embedded in the DTM core (Section II) as a software.

The DTM core consumes power for the execution of the moving average filter affecting power gating efficiency. This cost is evaluated in terms of energy and minimum idle time (MIT) [23] impact, which represents the minimum time that a PGD must stay in power-OFF mode (denoted by MIT_{orig}) in order to save energy. The energy consumed by the PGD while it is idle is $E(idle) = P_{OFF} MIT_{orig}$, where P_{OFF} is the static power consumption in OFF state. The PGD also consumes energy $E(PGD)$ for recharging during wake-up. Thus, the energy consumed during idle and the recharging energy must be lower than the energy that would be consumed if the PGD were always ON:

$$\begin{aligned} E(PGD) + E(idle) \leq E(\text{if ON for } MIT_{orig}) \Rightarrow \\ E(PGD) + P_{OFF} MIT_{orig} \leq P_{ON} MIT_{orig} \end{aligned} \quad (2)$$

where P_{ON} is the circuit static power consumption in power-ON state. Considering that $P_{OFF} \simeq 0.05 P_{ON}$ due to power gating [23], (2) becomes:

$$MIT_{orig} \geq E(PGD)/(0.95 P_{ON}) \quad (3)$$

For the proposed MIT evaluation, we consider the dynamic energy $E(dyn)$ of the DTM core. Instead, we do not consider its static energy, since the DTM core is already present in the SoC, and is never power-gated. Thus, the proposed MIT, denoted by MIT_{prop}, is given by:

$$\begin{aligned} E(PGD) + E(idle) + E(dyn) \leq P_{ON} MIT_{orig} \\ \xRightarrow{\text{using (3)}} MIT_{prop} \geq MIT_{orig} \Big[1 + \frac{E(dyn)}{E(PGD)}\Big] \end{aligned} \quad (4)$$

As in [34], we reasonably consider that half of the internal PGD nodes are in logic-'1' during wake-up. Thus, the energy E_{PGD} for recharging the PGD depends on the effective capacitance of the power network C_{PDN} and half of the capacitance of the logic: $E_{PGD} \simeq (C_{PDN} + 0.5 C_{PGD}) V_{dd}^2$. Also, the effective capacitance of the power network is almost half of the design [34], thus $C_{PDN} \simeq 0.5 C_{PGD}$. Therefore, $E(PGD) \simeq C_{PGD} V_{dd}^2$. As for $E(dyn)$, it is given by $E(dyn) = a C_{core} V_{dd}^2 s_{clk}$, where C_{core} is the capacitance of the DTM core, a the switching activity and the s_{clk} number of clock cycles to execute the software. Hence, the MIT cost $C_{MIT} = MIT_{prop}/MIT_{orig}$ becomes:

$$C_{MIT} = 1 + \frac{E(dyn)}{E(PGD)} = 1 + a \frac{C_{core}}{C_{PGD}} s_{clk} \quad (5)$$

For a relative evaluation, we consider the sizes of the PGD and DTM core similar ($C_{core} \simeq C_{PGD}$). Thus, (5) becomes:

$$C_{MIT} = 1 + a s_{clk} \quad (6)$$

The C_{MIT} of the proposed technique depends on the switching activity a of the DTM core and the elapsed clock cycles s_{clk}. As for the switching activity, we can consider a value $a = 0.15$, as in [35].

In addition, we evaluate the energy cost of the proposed technique. For this reason, we introduce a new metric, the ratio of the dynamic energy $E(dyn)$ consumed by the proposed technique on the DTM core against the energy that the power gating is saving when the circuit is idle for time t_{idle}. The *energy cost to energy savings ratio* will be simply referred to as *energy cost* E_{cost}, hereafter, and is given by:

$$E_{cost} = \frac{E(dyn)}{E_{sav\_orig}} = \frac{a C_{core} V_{dd}^2 s_{clk}}{0.95 P_{ON} t_{idle}} \quad (7)$$

When $E_{cost} > 100\%$ the consumed energy is greater than the saved energy. Since the energy stored in the circuit $E_{core} \simeq C_{core} V_{dd}^2$ is almost equal to the consumed energy during the discharge due to power gating $E_{core} \simeq P_{OFF} \cdot d_V$, (7) becomes:

$$\begin{aligned} E_{cost} = \frac{a(P_{OFF} \cdot d_V) s_{clk}}{0.95(P_{OFF}/0.05) t_{idle}} \Rightarrow \\ E_{cost}(t_{idle\_clk}) = \frac{0.15}{19} \frac{s_{clk}}{t_{idle\_clk}} d_{V\_clk} \end{aligned} \quad (8)$$

where s_{clk}, is the time to execute the software, whereas t_{idle_clk} the idle time and d_{V_clk} the discharge time d_V, expressed in clock cycles. As a worst case analysis using (8), we consider that $t_{idle\_clk} \simeq 10$ clock cycles, as in [34], whereas the $d_{V\_clk} \simeq 1000$ clock cycles, as evidenced by simulation results (Section IV) and experimental measurements (Section V). In Section IV-F, we present the energy and MIT cost of the processing block using metrics (6) and (8).

IV. Simulation Results

To evaluate the performance of the proposed technique, we apply it on a circuit of 21 cascaded inverters, referred to as *casc21*, on the *c432* and on the *s38584* and *s38417* benchmarks from the IWLS'05 suite [22]. All circuits have been synthesized with a 32nm high-k metal gate CMOS technology [31]. By means of SPICE simulations, we compare the aging estimation resolution achieved by the proposed technique against path-based approaches (Section IV-B). Also, we evaluate the performance of the proposed technique considering dynamic voltage and frequency scaling (DVFS) and we demonstrate its robustness against temperature variations. Finally, the cost of the proposed technique is evaluated in terms of area overhead, memory requirements, energy required by the processing block and its impact on the Minimum Idle Time (MIT). For any quantity Q at time t_i, we evaluate its relative error using $\varepsilon_Q(t_i) = |Est(Q(t_i)) - Act(Q(t_i))|/Act(Q(t_i))$, where $Est(Q(t_i))$ and $Act(Q(t_i))$ are the estimated and actual values of a quantity Q at time t_i. The average relative error at time t_i is computed as $\bar{\varepsilon}_Q(t_i) = \sum_{j=1}^{j=i} \varepsilon_Q(t_j)/i$.

A. Monte Carlo Simulation Setup

A circuit may operate using one or multiple DVFS operating modes that are controlled by DTM system policies, which affect its power consumption and its operating temperature. In order to simulate how the d_V is affected by the DTM policies, we examine pre-defined policies by generating 500 Monte Carlo permutations varying the probability that the circuit uses an operating mode. Particularly, each permutation is a Marcov Chain constructed by integrating the time range between t=0 to t=10 years with a time step of dt. For each step s_i, which corresponds to time from t_i to t_i+dt, we assume that the circuit executes a task with a task average temperature $T(t_i)$. Each $T(t_i)$ is considered to be a random value from a normal distribution with mean temperature T_p and standard deviation σ_p, the values of which are indicated by the policy. For each step s_i, the devices are characterized according to the models [27], [28] using the average temperature of all the tasks executed till task s_i: $T_A(t_i) = \sum_{j=1}^{j=i} T(t_j)/i$, and statistical stress values [12]. During the integration, unless it is stated differently, we assume that the circuit executes 8 tasks per day and each task is followed by a stand-by operation.

Example: Consider a scenario where the temperature $T(t_i)$ of a PGD during the execution of a task is a random variable with mean temperature $T_p = 80^oC$ and a standard deviation $\sigma_p = 3^oC$. A Monte Carlo permutation of this scenario, with $dt = 0.25$ days, is shown in Figure 9(a), where the temperature $T(t_i)$ of a task and the average temperature $T_A(t_i)$ of all tasks that have been executed till time t_i are shown. Next, Figure 9(b) presents the V_{th} degradation $\Delta V_{th}^i(t_i)$ at time t_i when the aging temperature is $T_A(t_i)$. The initial V_{th} for a pMOS is 0.49155V and T_A is 80^oC (Figure 9(a)). The ΔV_{th} is 16.88% after 4 years and reaches approximately 20% after 10 years. Finally, Figure 9(c) depicts the $d_V(t_i)$ after each task (shown as dots) and the average virtual-power-network discharge time $\overline{d_V}(t_i) = \sum_{j=1}^{j=i} d_V(t_j)/i$ (shown as a line) till time t_i, when we apply this scenario on benchmark *casc21*. ■

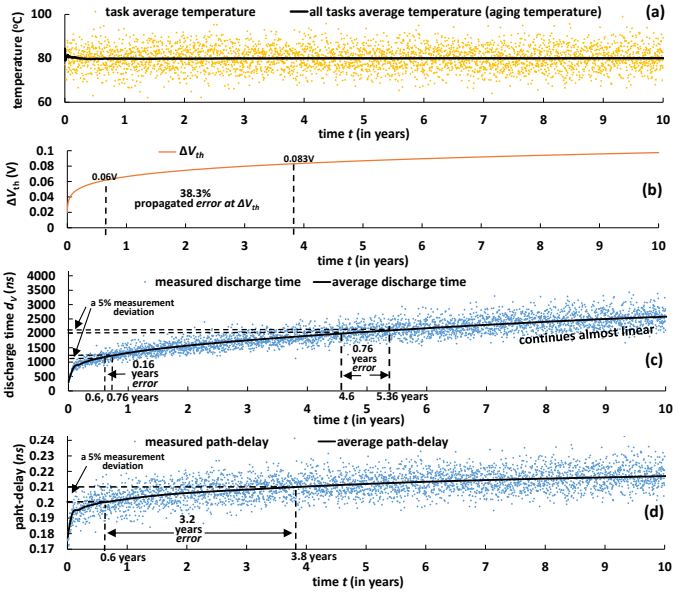

Fig. 9. For $T_A = 80^oC$: (a) Scenario of tasks temperature; (b) pMOS ΔV_{th} degradation; (c) discharge time d_V and (d) path-delay over time t.

B. Robustness to noise: path-delay vs discharge time

During the simulations, we also collect path-delay data. Figure 9(d) presents the path-delay for each task (points) and the average path-delay (line), when the tasks shown in Figure 9(a) are applied on the cascaded inverters *casc21* circuit. Comparing the dicharge time (Figure 9(c)) with the path-delay (Figure 9(d)) values, we observe that the discharge time is in the order of hundreds of nanoseconds, while the path-delay is in the order of hundreds of picoseconds. If we assume a very small measured path-delay deviation of 5% at t=0.6 years (Figure 9(d)), where the average path-delay is $0.2ns$ and the pMOS devices ΔV_{th} is $60mV$ (Figure 9(b)), then the average path-delay increases from $0.2ns$ to $0.21ns$, which is the value ΔV_{th} degradation $83mV$ at time t=3.8 years. This corresponds to a time error of 3.2 years. The propagated error at the estimated ΔV_{th} using path-delay $\varepsilon^{pd}_{\Delta V_{th}} = 38\%$ (Figure 9(b)), which is also the aging estimation resolution that can be achieved by path-based techniques. If we now assume a small deviation of 5% at the measured discharge time, at t=0.6 years, then the average discharge time varies from $1176ns$ to $1235ns$, which corresponds to the discharge time due to $\Delta V_{th} = 60.5mV$ that occurs at time t=0.76 years (for the same operating conditions). The propagated time error is 0.16 years, and the error of estimation using the discharge time would be $\varepsilon^{d_V}_{\Delta V_{th}} < 1\%$, which is a 97% error reduction, and hence resolution increase, compared to the aging estimation resolution using path-delay. Finally, in Figure 9 we observe that path-delay increases by less than 23%, while discharge time more than 1100% after 10 years of lifetime. Note that the robustness evaluation of the ring oscillator frequency drift sensors is similar to that of the path-delay based sensors, because the path-delay of the ring oscillator is its oscillation period. Therefore, we conclude that the discharge time is more robust to random noise and offers higher aging estimation resolution than path-delay and ring oscillators frequency drift.

TABLE I
AVERAGE DISCHARGE TIME AND BTI ESTIMATION RESULTS FROM MONTE CARLO SIMULATIONS USING SINGLE AND MULTIPLE POLICIES

circuit	policies #	cp-every	discharge time d_V sensor			BTI monitoring	
			s	sb #	$\overline{\varepsilon}_{d_V}$	sb #	$\overline{\varepsilon}_{\overline{\Delta V}_{th}}$
casc21	1 (static)	never	0.01	376	0.66	268	0.87
			0.05	83	0.97	1	0.79
	3 (DVFS)	day	0.01	344	5.9	251	4.1
			0.05	67	6.4	3	4.5
		month	0.01	344	7.8	251	5.9
			0.05	67	8.6	3	6.2
c432	1 (static)	never	0.01	374	0.57	265	0.61
			0.05	82	0.88	29	0.59
	3 (DVFS)	day	0.01	343	4.9	248	3.2
			0.05	68	5.4	25	3.5
		month	0.01	343	6.7	248	4.7
			0.05	68	7.5	25	4.6
s38417	1 (static)	never	0.01	184	0.29	119	0.18
			0.05	99	0.28	55	0.18
	3 (DVFS)	day	0.01	150	4.1	5	0.5
			0.05	11	4.3	3	2.6
		month	0.01	42	4.4	5	1.2
			0.05	10	4.9	5	1.6
s38584	1 (static)	never	0.01	193	0.36	77	0.19
			0.05	85	0.37	52	0.2
	3 (DVFS)	day	0.01	148	5.6	19	0.5
			0.05	8	5.8	9	3.0
		month	0.01	37	5.8	5	1.1
			0.05	6	5.9	4	1.2

C. Results on circuits implementing various DTM policies

First, we consider that the benchmarks operate using a single policy (static operating frequency) that follows a thermal profile p=$[90^oC, 3^oC]$, with average aging temperature T_p=90^oC and deviation σ_p=3^oC. Second, we consider three policies with operating voltages $(V_{dd1}, V_{dd2}, V_{dd3})$=$(0.9, 1, 1.1)$V and thermal profiles p_L=$[75^oC, 2^oC]$, p_M=$[85^oC, 2^oC]$ and p_H=$[100^oC, 2^oC]$, respectively. Table I presents the results. Particularly, first column shows the circuit name and column "policies #" the number of available policies. We assume that 8 tasks/day are executed, therefore column "cp-every" reports the change-policy rule, which selects values from the set ['day', 'month', 'never']. When "cp-every" is set to value 'day' then the active policy of the circuit remains unchanged for 8 tasks and then it is randomly selected among the $[p_L, p_M, p_M]$ policies. Similarly, value 'month' indicates that the active policy remains unchanged for six months (30×8=240 tasks). Value 'never' applies only to the single policy case. The column labeled as "discharge time d_V sensor" contains information related to the d_V sensor (Section III-A): the parameter convergence speed 's' of the moving average filter, the number of stand-by operations required to converge '*sb #*', and the average relative error of the moving average filter $\overline{\varepsilon}_{\overline{d_V}}$ for all the Monte Carlo permutations. Note that, for s=0.01, the filter requires 265 stand-by operations to converge for the *c432* (single policy), while it requires only 29 operations for s=0.05. We also observe the earlier convergence of the sensor for higher 's' values, which, however, comes together with a higher error due to the filter's higher sensitivity to workload fluctuations. The error $\overline{\varepsilon}_{d_V}$ is small, in the range [0.36%-0.97%] and [4.1%-8.6%] for designs with single and multiple policies, respectively. The BTI estimation also requires a lower number of stand-by operations to converge, while s increases. For example, although the BTI monitoring of *casc21* requires 268 stand-by operations for s=0.01, it converges with the first stand-by operation for s=0.05. The error of the average threshold voltage degradation estimation $\overline{\varepsilon}_{\overline{\Delta V}_{th}}$ is very small, less than 1% for designs with a single policy and in the range [0.5%-6.2%] for designs with multiple policies. For all the Monte Carlo permutations conducted, the convergence occurs in the range 3 hours to 0.09 years. However, our results were obtained considering only *8 stand-by operations per day*, which is a small number. For circuits that are more frequenctly power-gated, the convergence could occur in minutes.

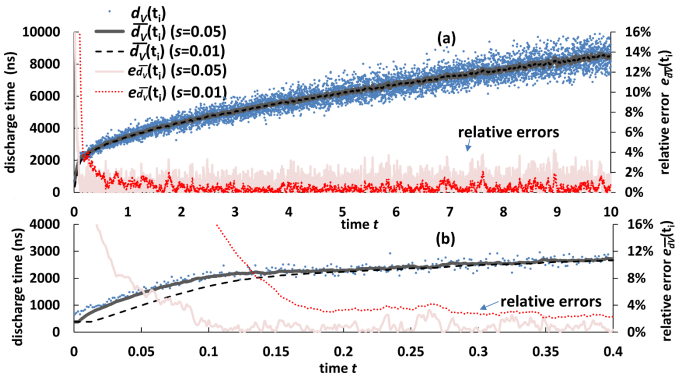

Fig. 10. Moving average $\overline{d_V}$ and error $\varepsilon_{\overline{d_V}}$ on single policy: (a) [0-10] years; (b) [0-0.4] years

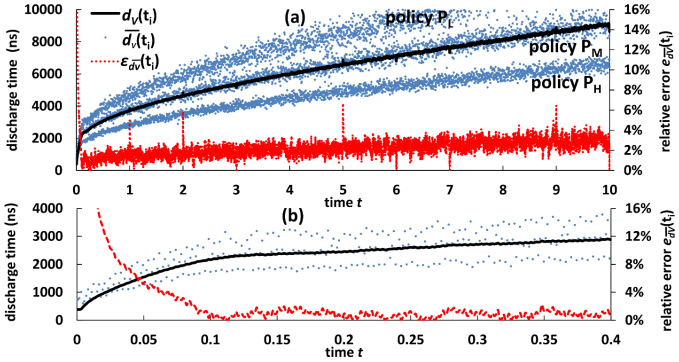

Fig. 11. Moving average $\overline{d_V}$ and error $\varepsilon_{\overline{d_V}}$ on three policies: (a) [0-10] years; (b) [0-0.4] years

Figures 10-12 focus on a single Monte Carlo permutation to present these trends in more detail. Figures 10 and 11 depict the discharge time $d_V(t_i)$ and the average $\overline{d_V}(t_i)$ (left y-axis) given by the moving average filter, as a function of time (x-axis), for circuit *c432* for both the single policy (Figure 10) and the three-policies (Figure 11, the three d_V regions represent one for each policy) cases, respectively. Figure 10 depicts results for the considered s values, s=0.01 and s=0.05. The relative error $\varepsilon_{\overline{d_V}}(t_i)$ (right y-axis) of the average discharge time estimation is also depicted. Figures 10(a) and 11(a) focus on the time range [0-10] years. The average relative error is 0.55% and 0.89% for s=0.01 and s=0.05, respectively, for the single policy case, and 3.2% for the three policies. Figures 10(b) and 11(b) focus on the time range [0-0.4] years.

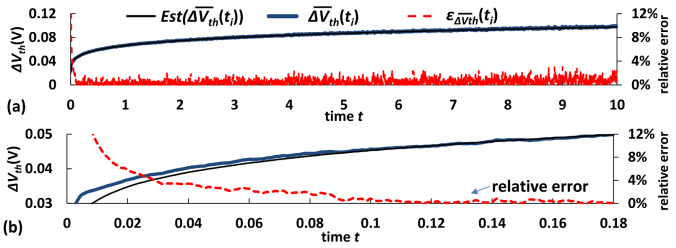

Fig. 12. Estimation error $\varepsilon_{\overline{\Delta V}_{th}}$: (a) [0-10] years; (b) [0-0.18] years

Figure 12 depicts the estimated ($Est(\overline{\Delta V}_{th}(t_i))$) and the actual ($\overline{\Delta V}_{th}(t_i)$) average V_{th} degradation (left y-axis) in time t ('x'-axis) for the single policy case (Figure 12). It also depicts their relative error $\varepsilon_{\overline{\Delta V}_{th}}(t_i)$ (right y-axis). The relative error between the estimated and the actual $\overline{\Delta V}_{th}$ values is higher at the beginning, but it reduces as the filter converges. The average value of the error $\overline{\varepsilon}_{\overline{\Delta V}_{th}}(t_i)$ is found 0.4% after the convergence. As convergence point, it is considered the moment when the relative error becomes <10% and occurs at 0.013 years (Figure 12(b)). For the case of three policies (Figure 11), the V_{th} degradation estimation error is following a similar trend. Its average value $\overline{\varepsilon}_{\overline{\Delta V}_{th}}(t_i)$ is found 3.2% after the convergence, which occurs at 0.024 years.

D. Temperature variations during stand-by operations

In previous paragraphs, both the temperature during stand-by operations $T_M(t_i)$ and the temperature of the executed task $T(i)$ were independent random numbers that were following the temperature variation of the active policy. Note that a reason to power-off a circuit could be the elevated temperature. Therefore, the average temperature during stand-by might be higher compared to the average temperature of the active policy. Therefore, we repeat all the simulations by considering that the average temperature during stand-by operations $\overline{T}_M$ is higher compared to the average temperature of the active policy by modelling T_M as $T_M(t_i) = T(t_i) + d_{T_M} + \sigma_{T_M}$, where d_{T_M} is a drift and σ_{T_M} a white noise deviation of temperature during stand-by at time t_i, compared to the task temperature $T(t_i)$. For a high deviation of $\sigma_{T_M} = 10^oC$ and without a drift ($d_{T_M} = 0^oC$), the proposed technique performs without any additional notable error, because the white noise is canceled by the moving average filter. The drift introduces an error in the average threshold voltage estimation, which for $d_{T_M} = 5^oC$ can reach 9.4%. However, we note that this error is systematic, thus it can be corrected by the processing block. Even in the case that this error is ignored, the drift is the same for identical designs, hence it does not affect the practicality of the proposed technique for comparing their aging status.

E. Area cost and system memory requirements

We evaluate the area cost of the hardware block as well as the memory requirements of the processing block. The discharge time sensor (Section III-A) consists of only a logic AND gate, an inverter and a clock cycles counter. Note that this type of delay sensors may already be part of the power gating DFT infrastructure [30], [36], [37]. The maximum number of bits $|CC|$ for the counter for all the examined cases was $|CC| = \lceil log_2(d_V(10, 120^oC, 60^oC)/T_{clk}) \rceil = 14$ bits, where $d_V(10, 120^oC, 60^oC)$ is the maximum d_V value, which is observed after time t=10 years, operating at average temperature of T_A=120^oC (higher temperature considered) with temperature during stand-by of the d_V set at T_M=60^oC (lower temperature considered) and operating clock period T_{clk}=$1ns$. The overall area overhead, when the DFT infrastructure [30] is not available, is $\leq$ 0.4% of s38417 and does not depend on the size of the design. In addition, we examined the non-volatile memory size $|M|$ required by the processing block software in order to approximate the $T_A(t, d_V)$ function. Using linear interpolation coefficients from 64 collected points for the processing block, $|M|$=$4 \times 4 \times$ (# of points) bytes, with 4 number of linear coefficients of 4 bytes each per point. Thus $|M|$=1 Kbyte, which is a very low memory cost. The discharge time sensor is accessible by the DTM core (Section II) through cross layer sensor access mechanisms that reuse DFT and interconnection infrastructure [1], [14], [24].

TABLE II
AVERAGE ENERGY-SAVING AND MIT COSTS FOR PROCESSING

processing	runs every	MIT_{cost}		$\overline{E}_{cost}(t_{idle})$ (%) $MIT \leq t_{idle} < d_V$		$d_V \leq t_{idle} \leq$1sec	
		x86	ARM	x86	ARM	x86	ARM
filter	$\forall$ stand-by	1.3X	1.45X	7.3	10.9	9.8E-05	1.7E-04
aging monitor.	2 months	negligible					

F. Energy and minimum idle time cost

We implemented the moving average filter in C programming language, which was compiled into 7 and 12 instructions from x86 and ARM instruction sets, respectively. We consider that each instruction is executed in one clock cycle, thus $s_{clk\_x86} = 7+2 = 9$ and $s_{clk\_ARM} = 12+3 = 15$, considering also the clock cycles for checking the *discharged* signal. Next, we use (6) and (8) to evaluate the processing block cost:
Moving average filter: Since MIT is less than the time of the circuit to discharge ($MIT<d_V$ [34]), we examine the energy cost, when the t_{idle} belongs to one of the two possible intervals: 1) $MIT \leq t_{idle} < d_V$, and 2) $d_V \leq t_{idle} \leq 1$ sec. The DTM core is aware if the PGD was fully discharged, through the *discharged* signal of the sensor. If the PGD wakes up before the circuit discharges ($MIT \leq t_{idle} < d_V$), the moving average filter execution is avoided and only 2 and 3 instructions are required from the x86 and ARM sets, respectively, to check the value of *discharged* signal, implying (using (6)) an C_{MIT} of 1.3X and 1.45X, respectively, as shown in Table II. Also, the average energy cost in this interval is:

$$\overline{E}_{cost}(t_{idle}) = \frac{\int_{t_{idle}=A}^{t_{idle}=B} E_{cost}(t_{idle}) dt_{idle}}{|B - A|} \quad (9)$$

This energy cost is evaluated for $MIT \leq t_{idle} < d_V$ by using A=MIT and B=d_V. For x86 and ARM architectures the $\overline{E}_{cost}$ results, which are shown in Table II, are 7.3% and 10.9%, respectively. When $d_V \leq t_{idle}$, the filter is executed and the energy cost is evaluated using (9) with A=d_V and B=1sec in clock cycles. It is found 9.8E-05% and 1.7E-04% for each

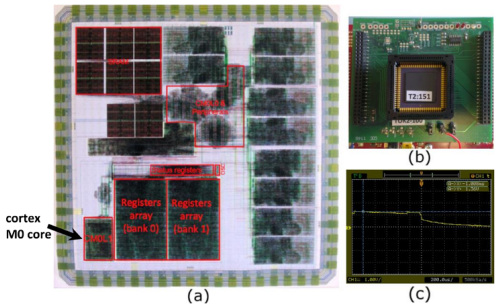

Fig. 13. (a) Chip floorplan; (b) exposed V_{Vdd} pin; (c) oscilloscope.

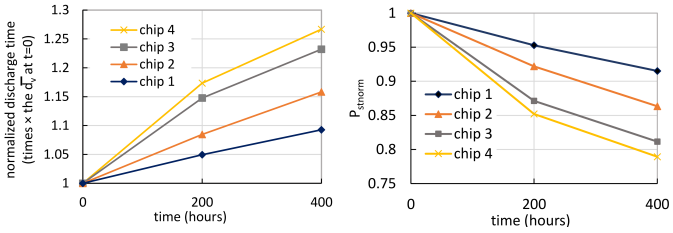

Fig. 14. (a) Average discharge time $\overline{d_V}$; and (b) P_{stnorm} trend of 4 chips

architecture, respectively (shown in Table II). The worst-case energy cost for this process is when t_{idle}= d_V and is evaluated using (8) at 7.1% and 11.8% for each architecture, respectively. *Aging monitoring process (Accessing of the Look-up Table):* We presented in Section IV-B that a 5% d_V variability propagates a V_{th} shift error <1% and that such V_{th} variability is exhibited between PGD with 0.16 years time difference. Due to this resolution bound, the aging monitoring process runs periodically with the very low period of 0.16 years (∼2 months) and, hence, its energy cost is negligible. Also, the larger the PGD is compared to the DTM core, the lower is the cost presented in Table II. Figure 13 shows the floorplan of an actual SoC, which has a DTM core that is an ARM cortex M0 processor and is located at the bottom-left corner of the SoC. Note that most blocks in the SoC are larger than the core.

V. Experimental Validation

To demonstrate the impact of aging on the discharge time, we conduct experiments with actual chips. The experimental setup is shown in Figure 13. The test-chips used in our experiment contain the SoC Tokashi [38] (Figure 13(a)) and are manufactured with a 65nm CMOS technology. The V_{dd} is connected to $1.2V$ power supply. The SoC has an ARM cortex M0 processor that is power-gated as a single block and has an exposed V_{Vdd} pin (Figure 13(b)) that can be directly accessed by an external oscilloscope (Figure 13(c)). Through the external oscilloscope, we collect virtual voltage V_{Vdd} waveforms during various stand-by operations of the processor in time. These measurements are post-processed for emulating the operation of the proposed processing block. The impact on the discharge time of oscilloscope's probe (∼$10M\Omega$ resistance) is negligible and the V_{Vdd} network discharges mainly through the chip (∼$50K\Omega$ resistance). The same instrument is used throughout the experiments and a relative evaluation of measurements compared to those obtained at t=0 is performed, thus any systematic variability induced by the instrument should not impact the observed trends.

To accelerate aging between measurements collection, we operate the chips at 70^oC, using a temperature chamber that has $\leq$ 5% accuracy error, while executing a computational intensive synthetic benchmark, the Dhrystone [39]. The discharge time is evaluated using oscilloscope measurements as the time interval from the assertion of the sleep signal to the moment where V_{Vdd} reaches a logic threshold of 25% of the V_{dd}. We collect K measurements at various time points t=0, 200 and 400 hours of operation. For each set of K measurements at a time point t, we compute the relative average discharge time compared to the average discharge time experienced by a chip at t=0. This normalized discharge time, which emulates the operation of the moving average filter, is simply referred to as *average discharge time*, hereafter, and is computed for each time point t by:

$$\overline{d_V}(t) = \sum_{i=1}^{i=K} d_{Vi}(t) / \sum_{i=1}^{i=K} d_{Vi}(t=0)$$

The $d_{Vi}(t{=}0)$ denotes one of the K measurements collected at the beginning of the experiment, when t=0. The measurements at each time point are considered to occur simultaneously, since the aging status of the chips is slightly affected during the few seconds of their manual collection. We expect that the average discharge time is sensitive to the aging status of the chips, if an increase trend over time is observed.

In Figure 14(a), we present the average discharge time of a set of $K = 10$ measurements for every time point t=0, 200 and 400 hours of operation for a set of chips. After 200 hours of operation there is a 5%-17.4% increase of the average discharge time, which increases to 9.3%-26.7% after 400 hours of operation compared to the average discharge time at t=0. As expected, a clear increase of the average discharge time for all the examined chips is observed confirming its sensitivity to the aging status of the chips. On the other hand, the absolute d_{Vi} measurements are highly sensitive to random noise and vary in the range [613 1240]ns. Next, we obtain a trend for the static power P_{stnorm} over time by considering that the charge which is stored in the circuit and the leakage current I_{leak} are constant during discharging: $P_{stnorm} = \frac{I_{leak}(t)}{I_{leak}(t=0)} \propto \frac{d_V(t=0)}{d_V(t)}$. Figure 14(b) depicts the computed static power trend for the examined chips. These results are consistent with the static power reduction with BTI aging reported by [18]. The aging of the chips at t=0 differs, since they were manufactured in 2012 and have also been used for other purposes.

In the next experiment, we focus on another chip, relatively "fresh" than those used for the previous experiments, and we repeat the experiment for 4000 hours≃5.5 months. Figure 15(a) depicts the collected data. We collect d_V measurements every 100 hours, while time t<600 hours (Figure 15(b)), and every 500 hours when time t>600 hours (Figure 15(c)). We also collect data at t=4000 hours. The same process as before is followed on each measurement. The reported d_V values are relative to time t=0. A clear incremental trend of the average discharge time $\overline{d_V}$ in time up to 2.79× (times ×) compared to the average d_V at time t=0 is shown after 4000 hours of

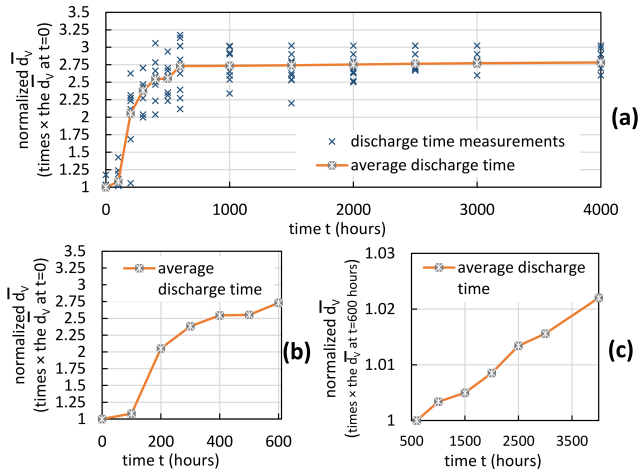

Fig. 15. Measurements from a 4000 hours (∼5.5 months) accelerated aging experiment: (a) d_V and $\overline{d_V}$ values, when t∈[0 4000] hours; (b) $\overline{d_V}$, when t∈ [0 600] hours; (c) $\overline{d_V}$, when t∈ [600 4000].

operation. Particularly, the $\overline{d_V}$ increases by 2.75× after almost a month (Figure 15(b)) and continues increasing, almost linearly, for ≃1% every 79 days (Figure 15(c)). The absolute $\overline{d_V}$ values are in the range [410 1650]ns. The observed trend of the average virtual-power-network discharge time $\overline{d_V}$ is in consistency with the expected trend, thus confirming its sensitivity to the BTI aging status of the design.

Note that the examined core (Figure 13) is power-gated as a single block. However, the proposed technique can also be adapted for cores with individually power-gated blocks by following coarse-grained rules, which depend on the objectives of the application that utilizes the coarse-grained BTI monitoring. For example, an application that targets to maximize reliability can consider the most aged block, as representative of the core, while an application that targets to maximize power consumption can consider the average aging of all blocks, instead. Nevertheless, the proposed technique remains unaffected in principle, while only additional software is required for following such coarse-grained rules. The analytical tools provided in Section III-C can be used for analyzing this additional cost, which is architectural and objective dependent.

VI. Conclusions

We presented a coarse-grained technique for monitoring online the impact of Bias Temperature Instability (BTI) aging on the CMOS devices of power-gated designs (PGDs) that consists of an on-chip virtual-power-network sensor embedded in the power-gating controller and a processing block for processing the collected measurements. The proposed technique features some advantages over fine-grained techniques: 1) It does not require the mission profile to be known during design, making it also applicable to memories; 2) upto 97% higher average aging estimation resolution is achieved than that of path-delay based techniques; 3) the virtual-power-network is already distributed in the PGD, and thus it does not require additional distributed sensors. By means of SPICE simulation, we evaluated the performance of the proposed technique on PGDs with static operating frequency and dynamic voltage and frequency scaling. The average threshold voltage estimation error induced by random temperature variations was found to be negligible. The minimum idle time increase caused by the energy consumed by the proposed software was evaluated on two scalar machine models that use x86 and ARM instruction sets and was found <30% and <45%, respectively. Through accelerated aging experiments using five actual chips with a System-on-Chip that contains an ARM Cortex processor, we validated the discharge time sensitivity to BTI.

Acknowledgments

The authors would like to thank Dr David Flynn (ARM Limited R&D Fellow) for providing feedback and the experimental chips. This work is supported by EPSRC (UK) under grant no. EP/K000810/1 and by the Department of Electrical Engineering and Electronics, University of Liverpool, UK. Experimental data used in this paper can be found at DOI:10.5258/SOTON/402489 (http://doi.org/10.5258/SOTON/402489).

References

[1] H. Yi, T. Yoneda, I. Inoue, Y. Sato, S. Kajihara, and H. Fujiwara, "A failure prediction strategy for transistor aging," *IEEE Trans. on VLSI*, vol. 20, no. 11, pp. 1951–1959, Nov 2012.

[2] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. Paul, W. Wang, B. Yang, Y. Cao, and S. Mitra, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. IEEE Int. Test Conf. (ITC)*, Oct 2008, pp. 1–10.

[3] A. Baba and S. Mitra, "Testing for transistor aging," in *Proc. IEEE VTS*, May 2009, pp. 215–220.

[4] T. Kim, P.-F. Lu, and C. H. Kim, "Design of ring oscillator structures for measuring isolated nbti and pbti," in *Proc. IEEE ISCAS*, May 2012, pp. 1580–1583.

[5] M. Chen, H. Kufluoglu, J. Carulli, and V. Reddy, "Aging sensors for workload centric guardbanding in dynamic voltage scaling applications," in *Proc. IEEE IRPS*, April 2013, pp. 4A.2.1–4A.2.5.

[6] P.-F. Lu and K. Jenkins, "A built-in bti monitor for long-term data collection in ibm microprocessors," in *Proc. IEEE IRPS*, April 2013, pp. 4A.1.1–4A.1.6.

[7] Y. Sato, S. Kajihara, Y. Miura, T. Yoneda, S. Ohtake, I. Inoue, and H. Fujiwara, "A circuit failure prediction mechanism (dart) for high field reliability," in *Proc. IEEE ASICON*, 2009, pp. 581–584.

[8] S. Wang, M. Tehranipoor, and L. Winemberg, "In-field aging measurement and calibration for power-performance optimization," in *Proc. ACM/EDAC/IEEE DAC*, June 2011, pp. 706–711.

[9] M. Noda, S. Kajihara, Y. Sato, K. Miyase, X. Wen, and Y. Miura, "On estimation of nbti-induced delay degradation," in *IEEE ETS*, May 2010, pp. 107–111.

[10] S. Wang, J. Chen, and M. Tehranipoor, "Representative critical reliability paths for low-cost and accurate on-chip aging evaluation," in *Proc. IEEE/ACM ICCAD*, Nov 2012, pp. 736–741.

[11] M. Omaña, D. Rossi, N. Bosio, and C. Metra, "Low cost nbti degradation detection and masking approaches," *IEEE Transactions on Computers*, vol. 62, no. 3, pp. 496–509, March 2013.

[12] E. Mintarno, V. Chandra, D. Pietromonaco, R. Aitken, and R. Dutton, "Workload dependent nbti and pbti analysis for a sub-45nm commercial microprocessor," in *IEEE IRPS*, April 2013, pp. 3A.1.1–3A.1.6.

[13] E. Mintarno, J. Skaf, R. Zheng, J. Velamala, Y. Cao, S. Boyd, R. Dutton, and S. Mitra, "Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging," *IEEE Trans. on CAD*, vol. 30, no. 5, pp. 760–773, May 2011.

[14] S. Sarma, N. Dutt, P. Gupta, N. Venkatasubramanian, and A. Nicolau, "Cyberphysical-system-on-chip (cpsoc): A self-aware mpsoc paradigm with cross-layer virtual sensing and actuation," in *IEEE DATE*, March 2015, pp. 625–628.

[15] T. Fischer, J. Desai, B. Doyle, S. Naffziger, and B. Patella, "A 90-nm variable frequency clock system for a power-managed itanium architecture processor," *IEEE Journal of SSC*, vol. 41, no. 1, pp. 218–228, Jan 2006.

[16] K. Nowka, G. Carpenter, E. MacDonald, H. Ngo, B. Brock, K. Ishii, T. Nguyen, and J. Burns, "A 32-bit powerpc system-on-a-chip with support for dynamic voltage scaling and dynamic frequency scaling," *IEEE Journal of SSC*, vol. 37, no. 11, pp. 1441–1447, Nov 2002.

[17] D. Rossi, V. Tenentes, S. Yang, S. Khursheed, and B. Al-Hashimi, "Reliable power gating with nbti aging benefits," *IEEE Trans. on VLSI*, 2016.

[18] ——, "Aging benefits in nanometer cmos designs," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. PP, no. 99, pp. 1–1, 2016.

[19] D. Rossi, V. Tenentes, S. Khursheed, and B. M. Al-Hashimi, "Bti and leakage aware dynamic voltage scaling for reliable low power cache memories," in *IEEE IOLTS*, 2015.

[20] D. Rossi, V. Tenentes, S. Khursheed, and B. Al-Hashimi, "Nbti and leakage aware sleep transistor design for reliable and energy efficient power gating," in *Proc. ETS*, May 2015, pp. 1–6.

[21] D. Flynn, R. Aitken, A. Gibbons, and K. Shi, *Low Power Methodology Manual: For System-on-Chip Design*. NY, Springer-Verlag, 2007.

[22] IWLS'05, online: http://iwls.org/iwls2005/benchmarks.html.

[23] Y.-F. Tsai, D. Duarte, N. Vijaykrishnan, and M. Irwin, "Characterization and modeling of run-time techniques for leakage power reduction," *IEEE Trans. on VLSI*, vol. 12, no. 11, pp. 1221–1233, Nov 2004.

[24] M. He and M. Tehranipoor, "Sam: A comprehensive mechanism for accessing embedded sensors in modern socs," in *Proc. DFT, IEEE*, Oct 2014, pp. 240–245.

[25] S. Gupta and S. S. Sapatnekar, "Employing circadian rhythms to enhance power and reliability," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 18, no. 3, pp. 38:1–38:23, Jul. 2013.

[26] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb 2003.

[27] M. Fukui, S. Nakai, H. Miki, and S. Tsukiyama, "A dependable power grid optimization algorithm considering nbti timing degradation," in *IEEE NEWCAS*, June 2011, pp. 370–373.

[28] K. Joshi, S. Mukhopadhyay, N. Goel, and S. Mahapatra, "A consistent physical framework for n and p bti in hkmg mosfets," in *IEEE Rel. Phys. Symp. (IRPS)*, April 2012, pp. 5A.3.1–5A.3.10.

[29] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits, 2nd ed.* NJ, USA: Prentice-Hall, 2003.

[30] V. Tenentes, S. Khursheed, D. Rossi, S. Yang, and B. Al-Hashimi, "Dft architecture with power-distribution-network consideration for delay-based power gating test," *IEEE Trans. on CAD.*, vol. PP, no. 99, pp. 1–1, 2015.

[31] "Predictive Technology Model (PTM)," http://ptm.asu.edu.

[32] M. Thoben, K. Mainka, A. Groove, and R. Herms, "Simulation vs. measurement of transient thermal resistance zth of power modules and its effect on lifetime prediction," in *Proc. of PCIM Europe*, 2013, pp. 1070–1076.

[33] L. Schumaker, *Spline Functions: Basic Theory*. Cambridge Mathematical. Library, Cambridge University Press, Cambridge, third edition, 2007.

[34] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *ISLPED*, Aug 2004, pp. 32–37.

[35] R. Andrew and M. Datta, "Pushing the performance boundaries of arm cortex-m processors for future embedded design," cadence, White Paper, 2014.

[36] V. Tenentes, S. Khursheed, B. Al-Hashimi, S. Zhong, and S. Yang, "High quality testing of grid style power gating," in *Proc. IEEE Asian Test Symposium (ATS)*, Nov 2014, pp. 186–191.

[37] V. Tenentes, D. Rossi, S. Khursheed, and B. Al-Hashimi, "Diagnosis of power switches with power-distribution-network consideration," in *Proc. ETS*, May 2015, pp. 1–6.

[38] S. Yang, S. Khursheed, B. Al-Hashimi, D. Flynn, and G. Merrett, "Improved state integrity of flip-flops for voltage scaled retention under pvt variation," *IEEE Transactions Circ. and Syst. I*, vol. 60, no. 11, pp. 2953–2961, Nov 2013.

[39] R. P. Weicker, "Dhrystone: A synthetic systems programming benchmark," *ACM Commun.*, vol. 27, no. 10, pp. 1013–1030, 1984.

Vasileios Tenentes (M'07) received the B.Sc. degree (2003) in Computer Science from the University of Piraeus, Greece, and the M.Sc. degree (2007) in Computer Science from the University of Ioannina, Greece. He was with Siemens Enterprise Networks and with the R&D of Helic S.A. In 2013, he obtained his Ph.D. from the Department of Computer Science and Engineering, at the University of Ioannina in Greece. He has been a Research Fellow at the University of Southampton, UK, since 2014. He is interested in electronic design automation, testing of electronic devices, and cross-layer techniques for energy efficient and reliable multi-core SoCs.

Daniele Rossi (M'02) received the Laurea degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Bologna, Italy, in 2001 and 2005, respectively. He was a Senior Research Fellow at the University of Southampton, UK, in period 2014-2015, and, currently, he has been a Senior Lecturer at the University of Westminster, UK, since 2016. His research interests include fault modeling and design for reliability and test, focusing on low power and reliable digital design, robust design for soft error and aging resiliency, and high quality test for low power systems.

Sheng Yang received the B.Eng. degree in Electronic Engineering from University of Southampton, UK, in 2008, and the Ph.D. degree in Electronic Engineering from University of Southampton, UK., in 2013. In the past, he had an internship with NXP investigating network hubs, and with ARM investigating data integrity of flip-flops. In 2013-2015, he was a Research Fellow at the University of Southampton. Currently, he is an ARM Research Engineer within the Applied Silicon Group, ARM R&D, Cambridge, UK. His research interests include low power embedded system design, signal processing and machine learning.

Saqib Khursheed received his Ph.D. degree in Electronics and Electrical Engineering from University of Southampton, U.K., in 2010. Currently he is working as a lecturer (Assistant Professor) in the Department of Electrical Engineering and Electronics, University of Liverpool, UK. He is interested in all issues related to design, test, reliability and yield improvement of low-power, high-performance designs and 3D ICs. He is the Program Co-Chair of DFT-2017 and is member of program committees of ETS, ATS, VLSI-SOC and iNIS. In the past, he served as the Program/General Chair of Friday workshop on 3D Integration, collocated with DATE conference (2012-2015).

Bashir M. Al-Hashimi is an ARM Professor of Computer Engineering and Dean of the Faculty of Physical Sciences and Engineering, University of Southampton. In 2009, he was elected fellow of the IEEE for significant contributions to the design and test of low-power circuits and systems. He holds a Royal Society Wolfson Research Merit Award (2014-2019). He has published over 300 technical papers, authored or co-authored 5 books and has graduated 31 PhD students.

Steve R. Gunn received his degree in Electronic Engineering from the University of Southampton in 1992 and the PhD degree from the University of Southampton in 1996. He is a Professor, leading the Electronic and Software Systems research group at the Department of Electronic and Computer Science, University of Southampton, UK. In the past, he was coordinating the EU Network of Excellence on Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL and PASCAL2) for machine learning, statistics and optimisation. He has published over 100 papers in the areas of image processing, machine learning and embedded systems.